Selçuk J. Appl. Math.
Special Issue. pp. 51-68, 2012

Selçuk Journal of Applied Mathematics

Correlation and Regression – Errors and Pitfalls

Manfred Borovcnik

Alpen-Adria-University Klagenfurt, Department of Statistics, Klagenfurt
e-mail: manfred.borovcnik@uni-klu.ac.at

⋆ *Presented in 3rd National Communication Days of Konya Eregli Kemal Akman Vocational School, 28-29 April 2011.*

Abstract. Correlation and regression are important tools for empirical research to establish, justify and describe relations between characteristics. The concepts stem from a specific historical background and were intended to imitate causal connections for empirical research in wider areas than physics. While in physics the rank of causality has been changed drastically by entrenching theoretical physics models with genuine – randomness, the search for causal relations in soft sciences is still an ideal. The methods are applied by scientists from other disciplines than mathematics and statistics who have to interpret them from a "wider perspective". However, misunderstandings and misinterpretations about the concepts involved are also wide-spread even among statisticians. Common errors and pitfalls with regression and correlation are dealt with to enhance the concepts.

Key words: Empirical research, Misconceptions, correlation, regression.
2000 Mathematics Subject Classification: 62P99, 97K70.

1. Issues on Correlation

In empirical research there is a strong need for concepts to investigate interrelations between variables. The mathematics of such concepts might be very sophisticated, especially in multidimensional problems. For those who apply statistical methods, inner-mathematical derivations become meaningless. Yet it is important that such people also get some comprehension of the involved concepts and that they can decide their use and evaluate the results accordingly. In this section we will illustrate that a correlation coefficient is hard to interpret. There are subsidiary ways to give at least an idea that the kind of coefficient is helpful to judge the strength of co-relation of two variables.

1.1. Explorations about the Value of the Correlation Coefficient

We will start from scatter plots to demonstrate that the usual correlation coefficient is hard to tell from the graphs. ANSCOMBE (1973) has given several data-sets for illustrative purpose. His graphs are useful and have been cited quite often. Nowadays, with the help of software, insights about the correlation coefficient might be gained interactively, in changing some data, trace the effect on the scatter plot and see immediately also the impact on the correlation coefficient and the best fitting line for the relation.
In applications, such a correlation coefficient is statistically tested for significance and if it is significant different from 0, a relation between the variables involved is "established". Such a test is important to find relations in empirical investigations. Beyond such a test, it is important to interpret the size of a correlation to judge whether it is relevant for further investigations. That means, the scientist has to get a feeling what different sizes of correlation do mean for him and how he can evaluate the relations between the variables from the context.
For an experiment to enhance such judgements, we can start without technical details such as a definition of r. We just tell the students, that such a coefficient, called (Pearson Bravais) correlation coefficient r is used by statisticians as a measure of linear co-relation of two variables. Furthermore, such a coefficients takes values only in the interval [–1, 1] with a monotonic scale, ie. the greater the absolute value of r, the higher the correlation; if data are situated along a perfectly increasing line, this would lead to r=1, while data along perfectly decreasing line would lead to r=–1. However, the coefficient has no metric scale so that r=0.4 does not mean the double correlation of r=0.2.
We use a spreadsheet for our calculations. The data set is simple, its purpose is simply to demonstrate that some scatter plots may be handled with correlation and some might need further concepts. One of the points in Fig. 1 (to the right) is moved (by rulers) and the impact on the correlation and the best fitting line is shown in the scatter plot. In Fig. 1a we see a strong influence of points far out of the bulk of points. It influences not only the size of the correlation but also the fitting line for the relation. In Fig. 1b, firstly there is no relation between the main part of the data; however, the point outside "creates" sometimes a high correlation.

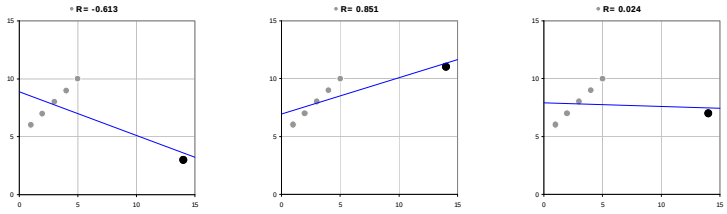

Fig.1a. A single point may strongly influences r and the regression line

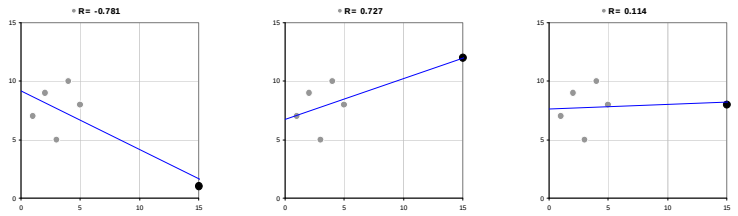

Fig.1b. A single point may "produce" a non-existent "relation"

We end this phase of exploration with the insight: The correlation coefficient r might measure the strength of correlation only if the data points form a scatter without gaps, with the same vertical scatter throughout. Otherwise, r might lose completely its meaning.

1.2. Standardized Scatter Plots

From the experiment in Fig. 1, we see that it is important, always to inspect the scatter gram of data to have a control on the values of r and their relevance. We need to calibrate our feeling about how to estimate $r$from scatter plots.

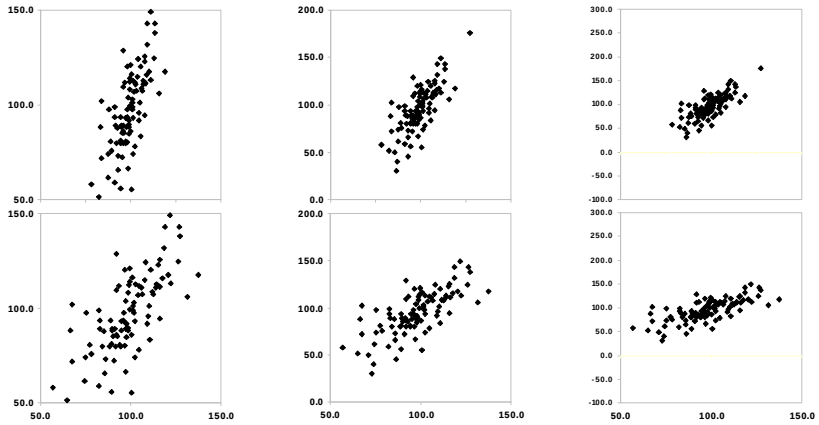

Fig.2. How big are the correlation coefficients in the various scatter plots?

The author used this set of scatter plots at various instances with statistical experts as well as with students or teachers of secondary schools and asked for an estimate. The estimates varied for the different plots but were remarkably similar for the different groups: 0.2–0.3 for the left and 0.9–0.95 for the right in the second row marked the extremes of the estimates given.
However, all the scatter plots do – in fact – represent the *same* data set with a correlation of r=0.75. It was a big surprise, also to statisticians to learn to know how influential the graphical appearance is. In their visual estimates, experts were no better than laypersons or students from several studies (including

mathematics and statistics). One needs to standardize scatter plots if r is to be estimated from the graph, or if the plots are communicated to applied scientists whom the statistician counsels to avoid confusion. Hereby the scales have to be chosen in the right proportions, ie., 1 unit on the x axis amounts to $1s_x$; likewise 1 unit on the $y$axis equals $1s_y$. By this choice, the dots for the points fill roughly a rectangle in the scatter plot. To calibrate one's visual estimation, it is helpful to have some standardized scatter plots in mind.

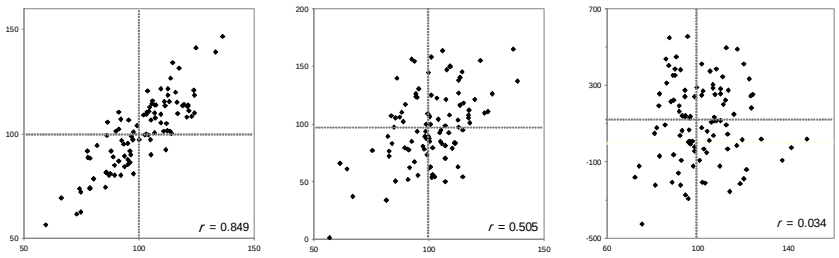

Fig.3. Thre efictional data sets with differing correlation to calibrate the visual estimation of r.

1.3. Co-Relation of Standard Scores

The idea of co-relation between to metric variables might be enhanced by the use of standardized scores (z values). For data x_i with mean $\bar{x}$ and standard deviation s_x, standard scores are defined usually as

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

Indirectly, ranks may be read off from z values (ideally in normal distributions). In terms of z values, $x_i = x + 2s_x$ eg., is 2 standard deviations above the mean and (in normal distributions) this has a small probability of less than 2.5%. Whether a data of an object is "extreme" may be read off from the pertinent standard score. In comparing whether the two variables under scrutiny co-relate, one may ask: Is the object as extreme on the y variable as it is extreme on the x variable? Or, are the standard scores of the two variables for this object of similar value, ie. does it hold

$$\frac{x_i - \bar{x}}{s_x} \quad \approx \quad \frac{y_i - \bar{y}}{s_y}?$$

In fact, the co-relation between the standardized values is exactly the same as it is for the original variables. This is a substantial property of a suitable measure of co-relation. Standardization should not change its values.

1.4. Two Interpretations of Correlation

As a question for co-relation has the same answer if put forward for original data or if put for standardized data, standardized values might be used further to understand how a concept to measure co-relation is built up.

Mean area of rectangles of standardized points

The coordinate system is transformed to standard units, for x the scale $z_x = \frac{x-\bar{x}}{s_x}$ is used.

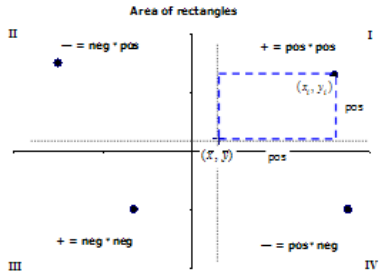

Fig. 4a. Signed area for the data points in a standardized co-ordinate system.

The signed area of one single rectangle related to the data point (x_i, y_i) amounts to:

$$\frac{x_i-\bar{x}}{s_x} \cdot \frac{y_i-\bar{y}}{s_y} .$$

The "mean" area of rectangles to all points equals the correlation coefficient:

$$r = \frac{1}{n-1}\sum_{i=1}^{n} \frac{x_i-\bar{x}}{s_x} \cdot \frac{y_i-\bar{y}}{s_y}$$

Balanced contribution of concordant points

First the standardization coordinates of x and y are introduced (as to the left), and then a further transformation of axes is applied.

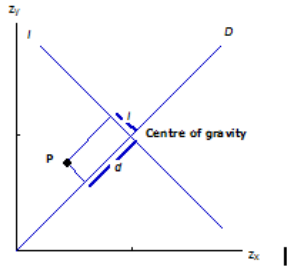

Fig. 4b. Co-ordinates of direct (D) and indirect (I) relation between standardized values.

The new coordinates of the points in this system are (d, i). The weights 'influencing' one point show either in the direction of D (direct relation of the variables) or I (indirect relation).

The correlation coefficient is an adjusted sum of weights for both directions:

$$r = \frac{\sum d_i^2 - \sum i_i^2}{\sum d_i^2 + \sum i_i^2}$$

For the new coordinates D and I in Fig. 4b, one may mention that these are identical to the two orthogonal components of a principal component analysis – the first is identical with D in case of r positive, and identical to I in case of r negative (see eg. HUCK et al, 2007).

2. Issues on Regression

Besides the standard regression line, a further line to represent data in a scatter plot will be re-introduced in this section. This concept, which was used historically before the method of least squares was established, will help to clarify an often misunderstood phenomenon of regression towards the mean. Also, it will help to understand the purpose of a regression line and the "shape" of the predictions. The evaluation of the fit of the best line may be misleading and the impact of the relation may be over- or underrated if the graphical displays use arbitrary scales. From normalized plots such an evaluation should be more reliable.

2.1. Two Different Lines to Describe Co-Relation: R- and S-Line

Regression is the name for the method of fitting lines (or non-linear functions) to data in order to describe the general model (trend) for the relation between the variables involved, or to estimate the values of the dependent variable y from the known value of the independent variable x. The best model is derived by means of the method of least squares. Again, the mathematical details are beyond many researchers who use the method of regression in empirical research. To enhance understanding of the concepts for such researchers, arguments beyond mathematics may be used. We will take up the historical perspective of developing the concepts of regression and correlation by Galton and Pearson (see MACKENZIE, 1981), which has been used also by FREEDMAN et al (2007).

The following data should be analyzed whether there is a linear relation between the variables.

- To get an idea how good the relation between the variables is described by the correlation coefficient, it is good to standardize the scales of the scatter plot – Fig. 5a.

not all data displayed	x	y
	100.07	61.58
	112.94	96.99
	102.42	54.25
R	0.50	0.26
mean	101.24	98.31
s.d.	15.22	32.73
proportion of scales of axes		2.15

- To get an idea how big the influence of the independent variable is, it is wise to turn back to original scales. The following plots use original scales – two different lines are used to describe the relation between the variables – Fig. 5b.

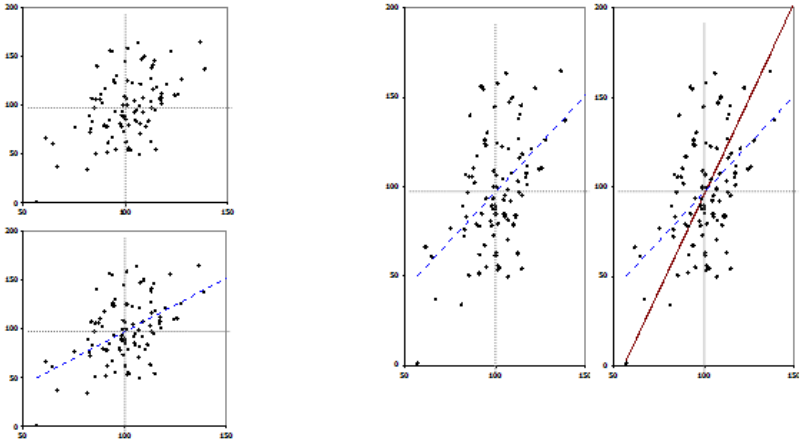

Fig.5. Scatter plots in standardized and original scale of variables

The broken line in Fig. 5 is the usual regression line derived by least squares; we will name it r line and compare it to the so called s line. This comparison will shed a light on the nature of the regression line and enhance the formulae involved.

s line: $\frac{y-\bar{y}}{s_y} = \frac{x-\bar{x}}{s_x}$

I.e., standard scores are equated. An object is set to be "extreme" to the same degree in both distributions. If one knows the value of the independent variable x, the dependent y may be predicted from this equation.

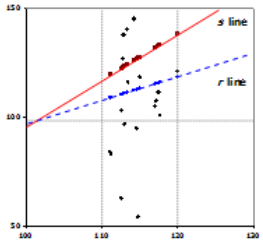

Fig. 6a. Two different lines for predicting the dependent variable.

r line: $y = \hat{a} + \hat{b} \cdot x$

The regression line asks for the mean of the dependent variable if the independent variable is known. This comes close to the mean of the data in single vertical stripes. The parameters a and b are estimated from the data.

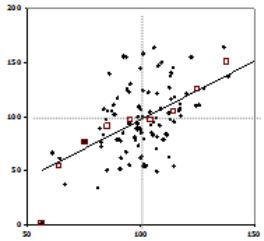

Fig. 6b. The mean of the dependent variable y in small vertical stripes.

s line

$$y = \bar{y} + \frac{s_y}{s_x}(x - \bar{x})$$

Slope $= \frac{s_y}{s_x}$

r line

$$y = \hat{a} + \hat{b} \cdot x$$

with intercept $\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$

and slope $\hat{b} = \frac{s_{xy}}{s_x^2}$

2.2. Regression to the Mean

A "natural law" behind naming new concepts

We will write the equations for both prediction models in terms of standardized values. The equation for the regression line simplifies by $r = \frac{s_{xy}}{s_x \cdot s_y}$.

s line in standardized variables	r line in standardized variables
$\frac{y-\bar{y}}{s_y} = \frac{x-\bar{x}}{s_x}$	$\frac{y-\bar{y}}{s_y} = r\frac{x-\bar{x}}{s_x}$

This makes clear the basic distinction between the two models for prediction: If the s line is used, then a standard value of 2 in x will lead to a prediction of y to be of 2 standard deviations above its mean. Using the least squares line, this prediction will be systematically changed by a factor r, which is the regression coefficient – this coefficient was therefore termed as *reversion coefficient* by Francis Galton (see eg., MACKENZIE, 1981). In the historic example of body length of fathers (the independent variable) and sons (the dependent variable), this has become known as the law of reversion, or the law of regression: If fathers are extreme off the mean value – eg., two standard deviations above – then their sons tend to be taller than average; but their length will be less extreme than their fathers': they regress towards the mean.

This was coined to be a major law of nature and a sign of heredity. By the way, the concepts of correlation and regression were developed to prove that –

besides body length, intelligence is hereditary. Such a phenomenon was advocated as an argument to foster intelligent persons to bring up children. While historically, this erroneous perception of regression played an important role for its promotion, the phenomenon of regression towards the mean should be now clear as an artefact: It holds for all variables under scrutiny. The slope of the regression line between standardized values for both variables is always less than 1!

Special case of regression to the mean – the case of equal variances for both variables

In case the variation of both variables is identical, ie., if it holds

$$s_y = s_x,$$

the phenomenon of "regression" towards the mean may already be read off the data in the original scale as it holds then:

$$y - \bar{y} = r \cdot (x - \bar{x}).$$

The variances of both variables may be equal (but need not be) in the case of a test-retest situation. The slope of the regression line equals r and as $|r| < 1$, the y values tend to be less extreme than the x values on the original scale in this case.

2.3. Standardized Plots Also With Regression Analysis

The impact of the strength of a relation is twofold: Firstly, the regression coefficient r gives a confirmation that such a relation is worthwhile as a model to describe the co-relation; also, the higher r, the more precise predictions will be. Secondly, the slope of the regression line influences the "grade" of dependence: for a steep slope the predictions will be highly influenced by changes in the independent variable. Again, we will illustrate that the visualization may mislead if scatter plots are arbitrarily scaled.

A study should analyze the effect of TV consumption on the length of deep sleep phase.

i	1	2	3	4	5	6	7	8	9
x_i	0.3	2.2	0.5	0.7	1	1.8	3	0.2	2.3
y_i	5.8	4.4	6.5	5.8	5.6	5	4.8	6	6.1
$\hat{y}_i$	6.02	5.17	5.93	5.84	5.71	5.35	4.81	6.07	5.12
$\varepsilon_i = y_i - \hat{y}_i$	-0.22	-0.77	0.57	-0.04	-0.11	-0.35	-0.01	-0.07	0.98

Two scatter plots with different scales for the given data induce differing awareness of correlation *and* regression coefficients. The effect of chosen scales is enlarged if the regression lines are drawn.

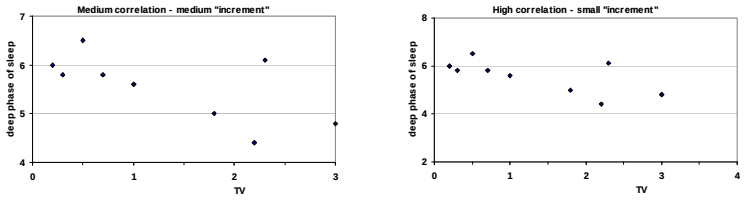

Fig. 7a. The scales used influence the impression of the influence of the independent variable

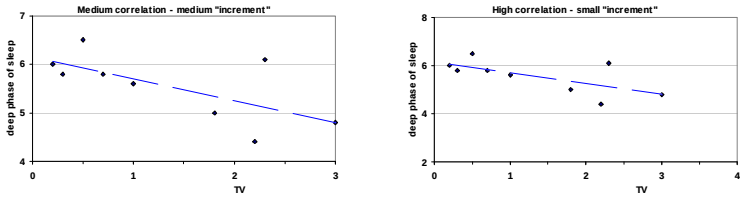

Fig. 7b. The effect of scales is increased if the regression line is also displayed

If coefficients are interpreted from scatter plots, the plots should be standardized again.
A graph – approximately standardized – shows the effect (Fig. 7c): High correlation - high increment. The coefficients confirm the visual impression as nearly 45% of the variation of the dependent variable may be explained by the independent variable:

intercept	6.16
regression coefficient	-0.45
regression equation $\hat{y} = \hat{\alpha} + \hat{\beta}x = 6.16 - 0.45x$	
correlation coefficient	-0.669
determination coefficient	0.448

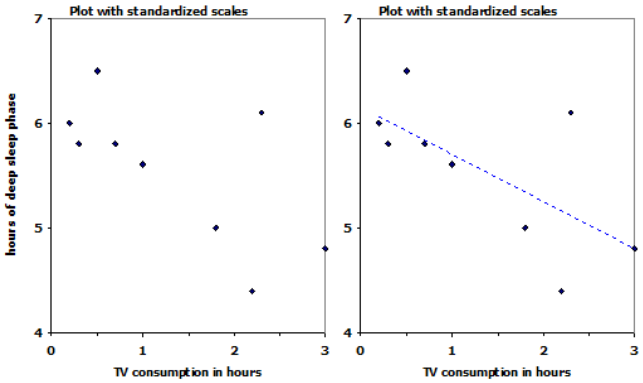

Fig.7c. Standardized plots reveal the relation of quite high correlation and high increment.

3. Back to Correlation

The split of sums of squares total into a sum related to the model and a sum related to residuals (errors) is basic for several statistical procedures. It paves the way for deriving statistical tests for the model and testing for significance for its parameters. It discloses also the term of variance (variation) explained by the model, which is substantial for empirical researchers in their judgement of the results for their contextual questions. However, usually this is proven generally, in mathematical terms, which hinders more than it gives an insight for those who are mathematically less inclined. This variance split will here be motivated by a special data set and simple techniques of data analysis. All the sums will get a concrete meaning thereby. A corroboration of the result might be gained by changing the data set purposefully. The remainder of this section will encounter some pitfalls, which are very frequent in empirical research. The discussion of interpreting correlation will also include items about causality, which is so different from correlation but is always in a researcher's mind

3.1. The Variance Split for Interpreting Correlations

The split of variances into an explained sums of squares (explained by the used model) and a residual sums of squares is highly important, not only for theoretical purpose but also for the interpretation of concepts used for the researcher who applies them. Again, for many researchers using regression and correlation in various fields, a mathematical argument is hardly convincing. We will establish the variance split using a simple data set. To corroborate the finding, the data may be systematically changed (by a slide in the spreadsheet used) with the result that the additive split remains invariant. In this experiment, we will use two different models to predict the value of the dependent variable:

i) naïve model: $\hat{y}_i = \bar{y}$; independently of x_i the mean value of y will be used as a prediction;

ii) linear model (least square): $\hat{y}_i = \hat{a} + \hat{b} \cdot x_i$; the point on the regression line is predicted.

data		model for y: mean		lin. model for y		resid
independent	dependent	prediction	residuals	prediction = fit		
	D	F_i	R_i=D-F_i	F	F - F_i	R=D-F
x	y	$\bar{y}$	$y - \bar{y}$	$\hat{y}$	$\hat{y} - \bar{y}$	$y - \hat{y}$
0		4.17				
1	3	4.17	-1.17	2.66	-1.51	0.34
2	2	4.17	-2.17	3.07	-1.10	-1.07
3	5	4.17	0.83	3.48	-0.69	1.52
6	2	4.17	-2.17	4.72	0.55	-2.72
7	7	4.17	2.83	5.13	0.96	1.87
9	6	4.17	1.83	5.95	1.79	0.05
10		4.17				
Sum	25.00	25.00	0.00	25.00	0.00	0.00
Mean	4.17	4.17	0.00	4.17	0.00	0.00
Variance	4.57	0.00	4.57	1.68	1.68	2.89
R^2	0.37					

Fig.8a. The additive split of variances: Total variance equals variance explained+unexplained

The mean value of the y data in Fig. 8a is 4.17; rule i) yields always a prediction of 4.17. The residuals (errors) of such a prediction have a mean of 0, which is good (no systematic bias of prediction). However, the variance of this prediction i) equals the variance of y originally (4.57), which is bad as – compared to the original data – no improvement is reached at. Using the least squares prediction ii) gives also a mean value of predicted values of 4.17 (which equals $\bar{y}$). For the error of procedure ii) it holds: mean error equals 0.
With regard to the variance of these residuals, we notice a value of 2.89, which is much smaller than the variance of 4.57 in the original data – a clear sign of improvement. The variance (interpreted as uncertainty of predicting values) has improved by 4.57 – 2.89. Inspecting the variance of the fitted points (their y value, of course), one notes the value of 1.68, which coincides exactly by the difference for the improvement. It holds

$$\underbrace{\text{var(data)}}_{\text{Variation total}} = \underbrace{\text{var(predictions)}}_{\text{Variation explained by model}} + \underbrace{\text{var(residuals of linear prediction)}}_{\text{Variation not explained by model}}$$

which is the famous additive split of the variances. Usually, statisticians note this law in the form of "sums of squares" (which form independent chi-square statistics etc). By simple techniques of data analysis, one may see important

relations of statistics.

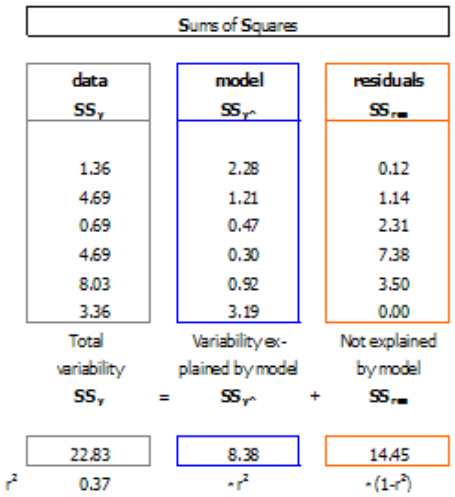

Fig.8b. The additive split of sums of squares

These relations remain stable when some of the data is changed in the spreadsheet by a slider. This is some kind of confirmation for an applied researcher who can now appreciate what is meant by variance explained and variance not explained. Furthermore, a new insight into the correlation coefficient is possible: To multiply the total variance byr^2 yields the explained variance:

$$SS_{\hat{y}} = SS_{total} \cdot r^2, or r^2 = \frac{SS_{\hat{y}}}{SS_{total}},$$

ie., the square of the correlation coefficient equals the percentage of explained variation of the total variation. Such a (determination) coefficient may then also be defined for non-linear models for prediction by way of this interpretation of percentage of variation explained by the model used for prediction.

3.2. Some Pitfalls

Finally, we will deal with some examples to illustrate various phenomena related to the concepts of correlation and regression: Firstly, additive effects of variables are not shown by correlations. Secondly, if aggregate data are used, correlations are increased by much and the fit of a relation is strongly suggested. And thirdly, third variables may enhance or even reverse a relation between two variables under scrutiny. Always is the question of causality to be analyzed separately from the statistical analysis.

Effects that correlation misses to capture

Correlation measures co-relation between variables. If the scatter plots are shifted eg., upwards, the mean value of the dependent variable would indicate this; however, the correlation coefficient does not reflect such a shift. It is

important to note this as there seems to be some confusion about it. An example from FREEDMAN et al (2007) deals with intelligence studies and whether intelligence is hereditary.
One study of SKODAK and SKEELS (1949) investigate the development of a co-relation of intelligence between children that have been adopted and their natural mother as well as to their foster mother in a longitudinal study. At a young age, correlations of child's intelligence to mother as well as to foster mother were small ($r < 0.1$). With ages 4 to 13 this correlation developed completely differently: while the correlation of child to natural mother became larger than 0.3, the correlation to fostering mother remained small; in fact it increased to 0.1 and 0.15 at age 13. This suggests that the child develops an intelligence that is much more closely related to its natural mother and much less to its fostering mother. A further study by HONZIK (1957) confirms this connection indirectly. It deals with the development of the correlation between intelligence of a child with its natural mother when the child was reared by its natural mother; see the dashed curve in Fig.9.

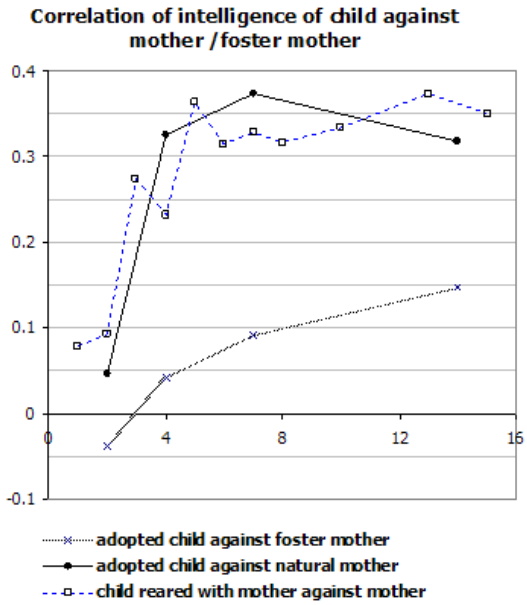

Fig.9. Development of intelligence between children and natural and fostering mothers

And, surprisingly or not, the correlation curve over the ages of the child develops quite similarly to the SKODAK and SKEELS study (see the two upper curves in Fig. 9). From this one may conclude that the intelligence of a child develops in some connection to its natural mother whether is reared by its own mother or it is fostered by an adoptive mother. Is intelligence mainly hereditary? Are

influences from the educational environment not strong enough? Is an influence of this educational environment completely missing?
It is interesting to note that despite the small correlation of children to fostering mothers, the influence of the adoption is relevant and high – however, it is not captured by the correlation coefficient: the natural mothers of those children who were adopted had an IQ of 86 while their children had an IQ of 106 so that they gained 20 points on average. Such an additive influence is not shown by correlations – and the summary of the investigation is based on the change of correlation coefficients (between intelligence of child and mother respectively fostering mother) with age of the child. The additive impact of enviroment to intelligence development is not reflected by correlation. There still remains the question of how to explain this increase of 20 points. Is it due to the supportive environment of the adopted child? Or, is it simply due to a selective process as only the brighter babies have been adopted?
Ecological correlations
If the correlation between variables is based on rates or averages, the term ecological correlation is in use. Such an aggregate analysis always increases the correlation coefficient and let the best line fit much better as compared to an analysis based on original data. FREEDMAN et al (2007) gives an example based on data about income and education level (using 1970 census data). If income is related to educational level for original data, the correlation is 0.4. However, if the data are aggregated over nine geographic regions, the correlation increases to 0.7.

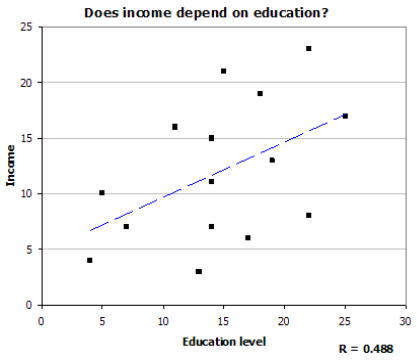

Fig.10a. Education and income -original data

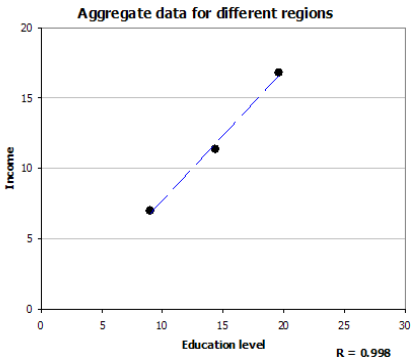

Fig.10b. Aggregate data for 3 regions

A fictional example (going back to FREEDMAN et al 2007) should illustrate matters: For original data, the correlation is 0.488 and the fit is quite reasonable, Fig. 10a). If the data are hypothesized to be from three regions and averages are taken, the scatter plot with the averages shows a perfect fit, Fig. 10b. The fit to single regions shows no uniform pattern for the relation between income and educational level, the correlation has gone down to values 0.109, 0.311, and

-0.128 (from left to right in Fig. 10c).

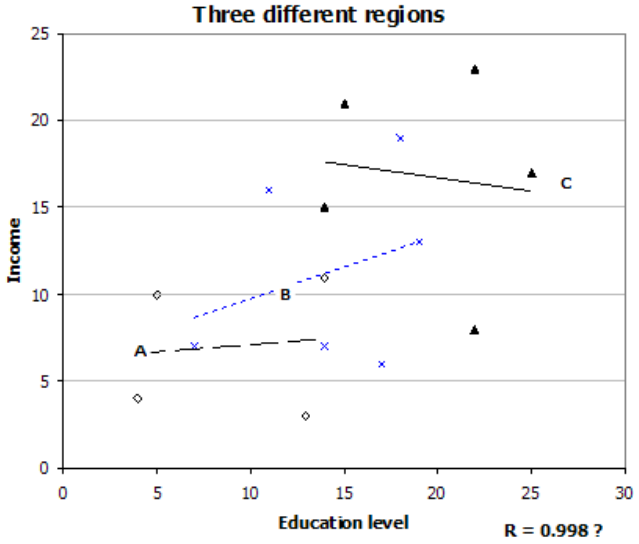

Fig.10c. Different relation in the 3r egions

The investigation of aggregate data is very popular in political sciences or in sociology as original data are missing in many cases (an early example about smoking causing cancer is DOLL 1957). The effect of such an approach should be clear and relations found have to be clarified by further investigations. The results come more close to preliminary hypotheses than they resemble findings from the study.

Third variables

An investigation of the initial income of academic persons (see KRÄMER 2007) has shown a strong relation of the income to the duration of the study, see Fig. 11. Should one give the advice to students to study longer (perhaps more carefully) to earn more? At closer scrutiny, the relation falls apart when one brings into the analysis a third variable, namely the study type. Clearly, business studies are shorter, physics studies last longer, and it is well-known that chemistry takes a long time to finish the studies. Across these studies,

however, the initial income increase highly from business to chemistry.

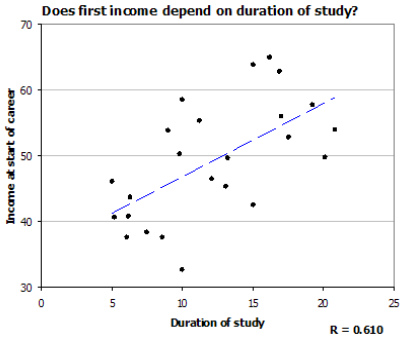

Fig.11a. First income and study length

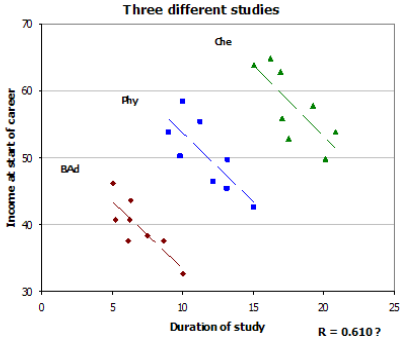

Fig.11b. Within studies the relation is reversed

Within the studies, the relation between study length and income is reversed. The longer the students take their time to finish the study, the lower the income at the start of the career. Third variables might enhance, blur, or even reverse a relation. The case of reversing the relation is also known as Simpson paradox.

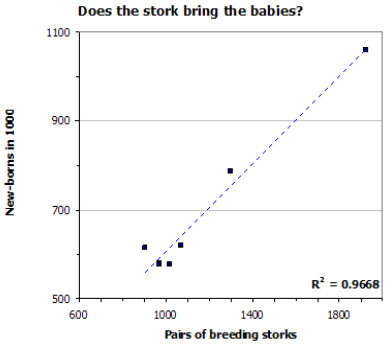

Fig.12a. A strong influence of storks to births?

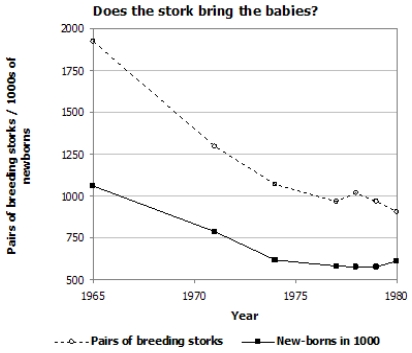

Fig.12b. Two time-related variables co-relate

Often the hidden variable is just time as shown in Fig. 12. Two time-series correlated to time are also correlated together. Clearly, we know that "the stork does not bring the babies" as may – at first sight – be concluded from the strong correlation between the numbers of storks and the new-borns in Western Germany for the years 1965–80 (see SIES, 1988, for this nice example). Sometimes, however, researchers are erroneously induced to think about causal relations between the variables under scrutiny if a correlation is higher (and with the storks and the babies it is higher than 0.983).

4. Summary – Correlation and Causality

The properties of correlation and regression discussed here are not new. The difficulties for understanding the underlying concepts and the restrictions on the interpretation of findings in applications are well-known. The twist here is to focus on a comprehension beyond mathematical justification, which is important for empirical researchers whose prime interest is their scientific discipline in which they work; they might also lack profound mathematical background. Yet they have to orientate themselves about the potential of the used statistical concepts. To enhance the meaning of the methods and to improve the capability to evaluate the findings, explorations of the effect of the used concepts might be more promising than further mathematical explanations:

- Many of the relations need not be played to the end mathematically.
- Exploration might help to orientate.
- Mathematical concepts always have two sides, they can clarify or blur.
- For applications, inner-mathematical relations get less important.
- Some of the explorations might be done interactively, we used EXCEL.

There is a final caveat: Correlation cannot "measure" causality of the phenomenon.

- Statistical laws orientate at the external phenomenon: data vary, or co-relate.
- Causal laws refer to the internal mechanism of connections: why data "influence" each other?

Once, a correlation is found to be significant and its size is judged to be relevant, further investigations have to be undertaken: replication studies should show that the phenomenon is stable; the search for contextual connections between the variables under scrutiny has to be pursued with detective care and persistence. If time related variables correlate with each other, an obvious co-factor is time. As was seen with the storks' example, such a correlation per se has no meaning and it is not always so obvious that a causal connection is absurd. All too easily a researcher is tempted to cross the line between statistical co-relation and causation as the latter always gives a much more attractive and acknowledged research finding.

Other co-factors may influence both variables. In the TV example, eg.,

- a factor of "un-quietness" might let these people look more at TV (to "relax")
- and these people might have shorter periods of deep sleep during night.

To clarify issues by further analyses, it is important to consider such co-factors early in the study otherwise pertinent data will be missing. The third variable or co-factor in the example with the income and study length was the discipline. After it is introduced in the analysis, the co-relation breaks down. The search for such co-factors is the key of empirical research and resembles more to an art and expertise knowledge about the context than it is connected to statistical expertise. If all potential co-factors are excluded, the task still remains to give a contextual explanation for a high correlation found.

GOOD and HARDIN (2003), or KHARSIKAR and KUNTE (2002) illustrate further pitfalls and how to deal with them. GOLDACRE (2008) analyses common errors in the approach within medical sciences. There is no statistical substitute for such arguments based on knowledge from the discipline. That makes the communication between the statistical advisor and the expert from the discipline so important.

References

1. F. J. Anscombe, Graphs in statistical analysis. The American Statistician 27, 17–21, 1973.
2. M. Borovcnik, Korrelation und Regression – Ein inhaltlicher Zugang zu den grundlegenden mathematischen Konzepten. Stochastik in der Schule 8(1), 5–32, 1988.
3. R. Doll, Etiology of Lung Cancer. Advances of Cancer Research 3(1955), 1–50.
4. D. Freedman, R. Pisani, R. Purves, Statistics. 4^{th} edition. W.W. Norton, 2007.
5. P. I. Good, J. W. Hardin, Common Errors in Statistics (and How to Avoid Them), Wiley, 2003.
6. M.P. Honzik, Developmental Studies of Parent-Child Resemblance in Intelligence. Child Development 28(1957), 215-228.
7. S. W. Huck, B. Ren, H. Yang, A New Way to Teach (or Compute) Pearson's r without Reliance on Cross-Products. Teaching Statistics 29(2007)1, 13–16.
8. B. Goldacre , Badscience. Fourth Estate, 2008.
9. A.V Kharsikar., S. Kunte, Understanding correlation. Teaching Statistics 24(2), 66–67, 2002.
10. W. Krämer, So lügt man mit Statistik. Piper 2007 (This Is the Way to Lie with Statistics).
11. D. A. Mackenzie, Statistics in Britain – 1865-1930 – The Social Constructioin of Scientific Knowledge. Edinburgh University Press 1981.
12. H. Sies, A New Parameter for Sex Education. Nature 332(1988), 495.
13. M. Skodak, H. M. Skeels, A final follow-up study of one hundred adopted children. Journal of Genetic Psychology 75(1949), 85–125.