Few Crystal Balls are Crystal Clear. Eyeballing Regression.

R.T. Wittebrood, Petrophysical Consultant
Wittebrood Petrophysical Services
Hycal Energy Research Laboratories Ltd.

Abstract

Regression lines and Eyeballed lines are often equally accurate. Theoretical arguments are illustrated using petrophysical examples. Apart from prerequisite requirements, such as proper sampling, the data has to satisfy 17 conditions for the regression line to be valid. Even then, one extra auxiliary assumption has to be made before the regression problem can be solved. Thus, regression does not necessarily yield the best fit line. An eyeballed line might be better.

1. Introduction

Modern reservoir analysis increasingly is done using packaged computer programs. These often have many routines and subroutines available, used like crystal balls, to calculate various parameters. However, most users often find these routines far from clear, when they start to think about them. Many routines appear opaque and some even may seem to be black to the casual user (black boxes). Many of us, of course, realize that crystal balls are only to be used by the initiated few, to gaze into the unknown.

Everyone knows, that statistical analysis can easily be abused, often even <u>un</u>intentionally. However, most users do not realize that, like in the field of petrophysics, the majority of our statistical results are in fact no better than an educated guess. The reason for this is a general lack of information about our data. This is also true for many other areas in the oil industry. For instance, most often no one is able to prove that a regression line is better or worse than a plain eyeballed line for the data. This in spite of all the crystal balls that were used in the analysis.

In this paper, we present many reasons why this is so. Many are revealed neither in standard statistical texts nor in school. <u>Before</u> regression may start enough knowledge should exist about:

1. The shape of the joint probability distribution
2. The presence of lurking (hidden) variables
3. The generation process of the data
4. The sampling procedures
5. The presence of outliers
6. The potential use
7. and more....

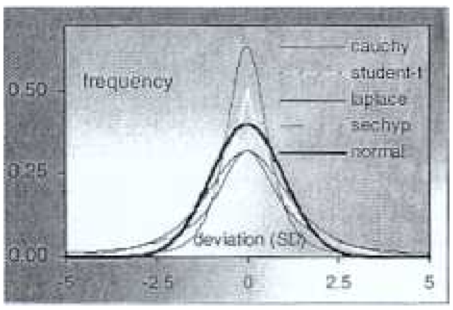
Next, regression <u>itself</u> sets a large number of conditions on the data, which will be discussed. Even when the data is fully cooperative and satisfies all regression conditions, the regression problem can not be solved! <u>*We always lack enough knowledge to solve the regression problem.*</u> This fact is hardly known. In order to solve the problem, we have to include at least one other assumption, which we ourselves have to make. It is usually based on the potential use of the regression results. The freedom in this assumption leads to the large variety of different regression techniques that are available: normal, weighed, inverse, reduced, bivariate, robust and special regressions. All these different techniques lead to different regression lines. The user has to select an assumption that he feels is most appropriate. The accuracy and the validity of this additional assumption are generally uncertain, making the regression line somewhat of a guess rather than a presumed truth. This is true even when the conditions for regression are satisfied. We will discuss many of the prerequisite conditions and the various common assumptions, using examples from the field of petrophysics. We show that in nearly all cases the regression lines are actually flawed to some extent. When we, for instance, want to regress the Formation Resistivity Factor against porosity (Archie), we need to use a weighed regression using the porosity (ϕ^2) as a weight factor and use inverse regression. We than will get the 'a' and 'm' factors. However, when we want to use the same data to find a relation between 'a' and 'm' the regression becomes much more complicated.

2. Prerequisite Conditions

To perform a proper regression, we first have to ensure that a number of conditions are satisfied. Some are more critical than others.

2.1 Distribution

To obtain the best-fit line using least squares regression the data should be normal distributed with respect to the equation fitted. However, many other different distributions are also possible. For different distributions the best-fit line will pass through different points:



1. Normal ($\exp -x^2$):	$\langle x \rangle$	average
2. Laplace ($e^{-\lvert x \rvert}$):	x_{med}	median
3. Uniform(1):	$(x_{lo}+x_{hi})/2$	low+high/2
4. Cauchy ($1/1+x^2$):	$\sum \varepsilon / 1+\varepsilon^2 = 0$	iterative

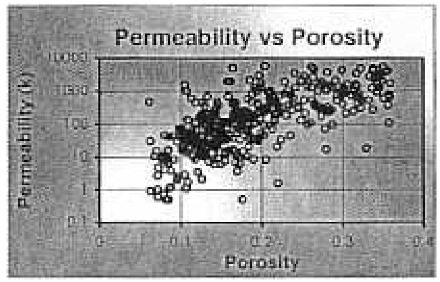
For many distributions, this point has to be calculated in an iterative manner to obtain the minimum for the sum of a function of the error, as for the Cauchy distribution. The Cauchy distribution has neither a defined average nor a standard deviation. This distribution is often used when outliers are frequent in the data set.

In spite of our common assumption, the normal distribution is not that normal. This can easily be deduced from the occurrence of outliers. A data point should have a chance of 0.3% to fall beyond the 3σ range. Or, we should have an outlier less than once in 300 times. We all know, that outliers are significantly more frequent than this. If outliers are frequent, a robust form of regression should be used.

2.2 Lurking Variables

A lurking or hidden variable may have more effect on a relationship between two quantities than anticipated. To illustrate this, we will use the example of the permeability-porosity crossplot.

The problem is that the permeability is mainly dependent on the pore throat size. The porosity has a smaller effect. However, we usually have no good measurement for the pore throat size, so the permeability porosity cross plot is often used.



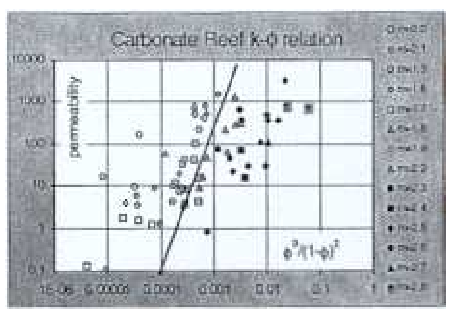
There exists a relation between resistivity and permeability, which is based on pore models:

$$\frac{kF}{P_d^2} = \text{const}$$

with F: the Formation Factor, k: the permeability and P_d the capillary displacement pressure. Because normally we may write $F = \phi^{-m}$ we get

$$\ln k = m.\ln\phi + 2.\ln P_d$$

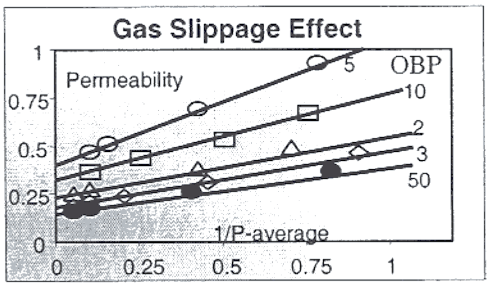
The effect of P_d will be larger than the effect of the porosity. The first graph applies to data from a carbonate reef, which shows a bad case of correlation. In the graph below, some of the data points for which FRF measurements were available are shown grouped according to the value of the m-exponent (m=-



lnF/lnϕ). The x-axis is the "Kozeny Porosity transform". The line shown is the line through the group of data for m=2 and 2.1 The other groups fall along lines parallel to this line. We observe that when we include the resistivity as a parameter in the regression, the remaining effect of the porosity on the permeability is very small. In this case, the addition of m-exponent as a parameter makes the relation between porosity and permeability more understandable. However, how to use this under normal evaluation conditions is another matter. (The above also holds in sandstones).

2.3 Origin of Data

How the data was obtained is of quite some importance on the method of regression. A good example of this can be found in the data for the gas slippage effect on the permeability known as the Klinkenberg effect.



Under laboratory conditions the permeability behaves as follows:

$$k = k_o(1+b/P_a)$$

with

- k_o : slip free permeability
- b : slip factor
- P_a : average pressure (Pin-Pout)/2

By plotting k versus 1/Pa at a given overburden pressure we can determine the intercept k_o and the slope $k_o.b$ from the regression line. In the plot, an example is given for one sample at different overburden pressures. We would like to know the slip free permeability. To calculate this value we need to know the b-factor. The factor 'b' follows from the line in the plot as:

$$b = \text{slope}/\text{intercept}$$

To measure b is expensive. In order to avoid measuring b we want to predict this value. Most often b is plotted versus the permeability. In the literature many examples of this can be found, which are compiled into the plot below. It is not always clear which value of k is used in the literature,but one may assume that in most cases it will be k_o. In normal regression, the regression line will go through a fixed point (the center of gravity). If the slope is somewhat too low, the intercept has to be higher. Therefore, an error in the intercept (k_o) will induce a strongly correlated error in the b-factor. In fact b is skewed in the same direction twice.

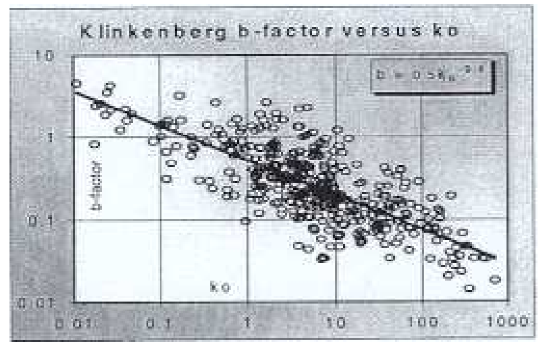
Errors should not be correlated in regression. Standard regression results in:

$$b = 0.5k_o^{-0.4}$$

This is flawed because of the correlated errors. When an estimate is made of the correlation of the errors and accounted for in the regression, the result becomes

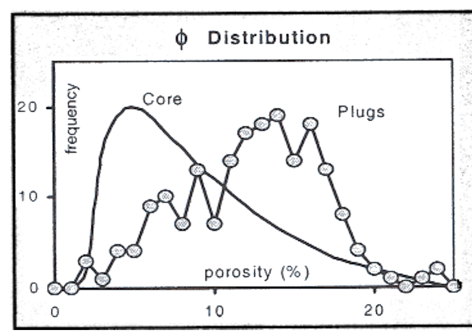
$$b = 0.6/\sqrt{k_o}$$

This result corresponds to the theory.



2.4 Sampling Procedure

No prerequisite regression condition is violated more often than the condition of measuring the results on a representative sample. On almost every occasion, we tend to take more samples or readings from the good portion of our reservoir than of the less developed section. This tends to bias our results in a significant way.

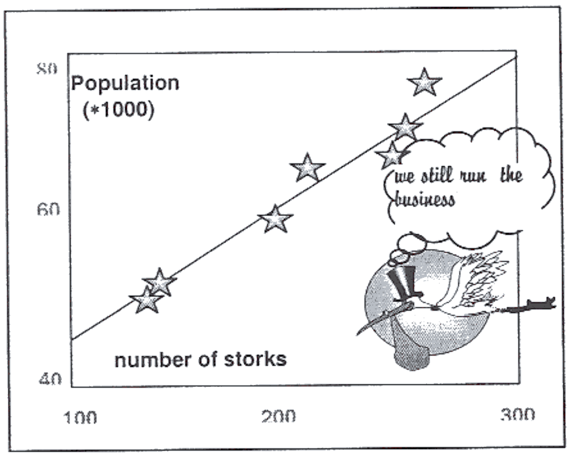


In the graph above, the estimated frequency distribution of the porosity in a number of cores is given with the actual measured porosity distribution over the samples for this core. It is evident that the high porosity sections are excessively sampled, by a factor of two or more, and the low porosity sections are not sampled frequently enough. The good porosity sections are weighed about five times heavier than the poor porosity sections. This will invalidate any regression line calculated and even eyeballed lines might be in error.

2.5 Causality

When we regress two variables against each other, it is necessary that the variables are correlated in causal manner. When we only know the effect but not the cause, we will be limited in the use of the regression results. We should have an expectation that some physical mechanism will change one variable when the other is varied. In the permeability porosity cross plot it can be expected that the porosity has only a minor effect on the permeability, because the permeability is determined by the pore throat size and not by the porosity. For the same porosity, a change in pore throat size of a factor of ten, will change the permeability by a factor of hundred, while the porosity has no effect.

There are a number of correlations to be found in the literature where a humorous result is presented. One is shown in the graph below.



The population of the town Oldenburg in Germany in the years 1930 to 1936 is plotted against the number of storks nesting within the city limits. The correlation coefficient is high, around 0.9. In Oldenburg the population apparently is controled by the stork polulation in a statistically significant manner. This type of humorous (?) regression is also to be found within petrophysics.

2.6 Use of Regression Result

The manner in which the regression results are going to be used will also have an effect on the method of regression. As an obvious example, we may use the example of the Klinkenberg relation for the gas slippage effect in the measurement of routine permeabilities. When we just want to know the slippage corrected permeability, the intercept, we want to use a different technique than when we want to use the results to predict b from ko. For just ko we could use bivariate regression, while for the prediction of b from k, we have to use a regression where we correct for the covariance between these two. As this covariance is unknown, we have to estimate it in some way, which is cumbersome.

The regression relation should mainly be used for interpolation. If extrapolation is envisioned the causal model should be good. In this case, the regression is most often used to determine the parameters of the model.

2.7 Other Factors

- We have to decide which variable will be used as the independent and which as the dependent variable. This is sometimes not self-evident.
- We have to make some decision about our outliers. Rejection of more than one outlier will change the reliability of the regression. The decision when to treat a data point as an outlier is dependent on the shape of the joint probability distribution and is not always easy.
- We also have to decide in what order we will do data manipulation and regression. Do we e.g. filter before or after the regression.

2.8 General Remarks

It is often stated that regression is rather insensitive to the normality of the joint probability distribution. This is not quite true. When we do regression, we have an objective in mind, e.g. we want to find the best-fit line. The best-fit line goes through a point determined by the shape of the distribution. Most regression techniques will draw the line through the center of gravity of the data (sometimes weighing is applied). This will not give the best-fit line if the distribution is not close to normal. An eyeballed line might also give good results. Is regression worth the trouble if the result is questionable? The answer is, of course,: it all depends.

3. Regression Theory

We start with a brief review of the standard theory of linear regression for a straight line.

3.1 Standard Regression Approach

1. Assume linear relation between y and x:

$$y = mx + b \qquad \textbf{(1)}$$

2. Define an error in y as the difference between the measured y-value: y_i and the predicted value according to equation (1): $\hat{y}$ as:

$$e_i = y_i - \hat{y}_i \qquad (2)$$

3. Calculate the sum of all squared errors SSE:

$$SSE = \sum e_i^2 \qquad \textbf{(3)}$$

4. Minimize SSE by varying the values of m and b (SSE will always be positive). At the minimum the slope of the curve of SSE versus m (the derivative) will be zero. Determine:

$$\partial SSE/\partial m = 0 \qquad (4a)$$

and

$$\partial SSE/\partial b = 0 \qquad (4b)$$

5. After some slight math we find:

$$m = SXY/SXX \qquad (5a)$$

and

$$b = \overline{y} - m\overline{x} \qquad (5b)$$

with

$$SXX = \sum(x_i - \overline{x})^2 \qquad (6a)$$

$$SXY = \sum(x_i - \overline{x})\cdot(y_i - \overline{y}) \qquad (6b)$$

$$SYY = \sum(y_i - \overline{y})^2 \qquad (6c)$$

Also

$$\overline{x} = \sum x_i / n \qquad (7a)$$

$$\overline{y} = \sum y_i / n \qquad (7b)$$

Equations (5), (6) and (7) form the standard regression results.

It is often stated that the errors should be normal distributed and not correlated to each other or to the variables etc. Where were these requirements needed in the above approach? At no point in the above reasoning (standard approach) are these conditions used. They do not enter the picture at all. These conditions only become important when we want the regression to yield the best-fit line. So, you can always calculate the standard regression line, but how well it fits can not be estimated from the above approach. To do this we need another approach.

3.2 Structural Model for Regression

We have selected the structural model to find an expression for the regression problem. In the structural model, we assume that there is only one line and that the data set fits to the line within some statistical limits. Another model assumes that there exist an infinite number of lines and we have to find the one that best fits the data in some sense. The two models both could be applied to our problem. Their difference is mainly philosophical.

In the structural model, we assume x_i and y_i differ from their intrinsic (χ_i resp. ψ_i unknown) values by some deviation (error):

$$x_i = \chi_i + d_i \qquad (8a)$$

$$y_i = \psi_i + e_i \qquad (8b)$$

and that we have a linear relation between Ψ,χ:

$$\Psi = m\chi + b \qquad (9)$$

We assume that the joint probability distribution is normal and are going to maximize the likelihood for the values of m and b, so that they best match the data set. Before we can solve the problem, we must satisfy 17 conditions. The first four conditions are trivial:

$$\mathrm{Exp}(x_i) = \chi_i \qquad (10a)$$

$$\mathrm{Exp}(d_i) = 0 \qquad (10b)$$

$$\mathrm{Exp}(y_i) = \Psi_i \qquad (10c)$$

$$\mathrm{Exp}(e_i) = 0 \qquad (10d)$$

The symbol 'Exp' means: the expected value of. If these are not satisfied, the data set is biased and must be removed. The next 13conditions are:

$$\mathrm{Var}(d_i) = \sigma_x^2 \qquad (11a)$$

$$\mathrm{Var}(e_i) = \sigma_y^2 \qquad (11b)$$

$$\mathrm{Var}(\chi_i) = \sigma^2 \qquad (11c)$$

$$\mathrm{Var}(\Psi_i) = m^2\mathrm{Var}(\chi_i) \qquad \text{due to eqn (9)}$$

$$\mathrm{Cov}(d_i,d_j) = 0 \qquad (12a)$$

$$\mathrm{Cov}(e_i,e_j) = 0 \qquad (12b)$$

$$\mathrm{Cov}(d_i,e_j) = 0 \qquad (12c)$$

$$\mathrm{Cov}(\chi_i,d_j) = 0 \qquad (13a)$$

$$\mathrm{Cov}(\Psi_i,e_j) = 0 \qquad (13b)$$

$$\mathrm{Cov}(\Psi_i,e_j) = 0 \qquad (14a)$$

$$\mathrm{Cov}(\Psi_i,d_j) = 0 \qquad (14b)$$

$$\mathrm{Cov}(\chi_i,\chi_j) = 0 \qquad (15a)$$

$$\mathrm{Cov}(\Psi_i,\Psi_j) = 0 \qquad (15b)$$

$$\mathrm{Cov}((\chi_i,\Psi_j) = 0 \qquad (15c)$$

Condition (11) requires that the deviations have variances, which do exist and can be calculated. They are a consequence of the use of the normal distribution. Conditions (12), (13), (14) and (15) require that all the errors and variables are independent of everything else. If these conditions are satisfied we can execute the math and obtain a set of three equations:

$$SXX = \sigma_x^2 + \sigma^2 \qquad (16a)$$

$$SXY = m\sigma^2 \qquad (16b)$$

$$SYY = \sigma_y^2 + m^2.\sigma^2 \qquad (16c)$$

Equation (6) still defines SXX, SXY and SYY.

Although equation (11) defines σ_x^2, σ_y^2 and σ^2, these values are unknown, as is the value for the slope m. Equation (16) gives us three relations with four unknowns. An infinite number of slopes is possible, which all satisfy equation (16). To arrive at a definite answer <u>we</u> have to introduce at least one extra relation. The choice of this relation defines the method of regression. Examples:

Standard Regression:	$\sigma_x^2 = 0$	$m = SXY/SXX$
Inverse Regression:	$\sigma_y^2 = 0$	$m = SYY/SXY$
Reduced Regression:	$\sigma_y^2/\sigma_x^2 = m^2 = SYY/SXX$	
Bivariate Regression	$\sigma_y^2/\sigma_x^2 = \lambda^2$	(sectn 4.5)

The structural model shows exactly why the conditions are needed to arrive at the 'best fit' line. In this model we minimize the sum of the cross product over all i and j of:

$$(\Psi_i+e_i-m(\chi_i+di)).(\Psi_j+e_j-m(\chi_j+d_j))$$

and this leads directly to Equations (11) to (15)

The best fit line will equal the standard regression line when the x-variable is error free. When we use a different probability distribution, the calculations change. For instance for a Laplace distribution ($e^{-|x|}$) we find that we have to minimize the Sum of the Absolute Errors (SAE) and not the Sum of the Squared Errors (SSE), which makes the calculations rather complicated. For a Cauchy distribution $[1/(1+x^2)]$ the average and the variance do not even exist, but their expected values needed in equation (10) do exist. Now we minimize the Sum of Declining Errors:

$$SDE = \sum \varepsilon_i/1+\varepsilon_i^2 \qquad (17)$$

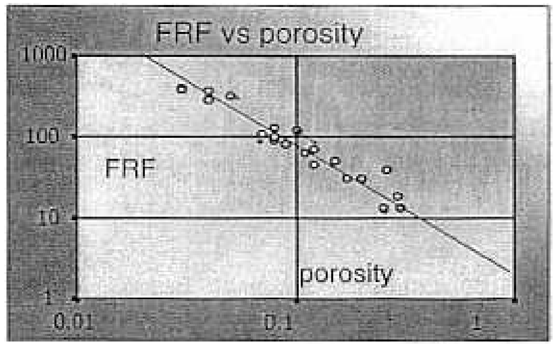
The value of ε_i is the deviation from the expected value. The expected value is not yet known before the minimum is found for the SDE. So the SDE can only be minimized in an iterative manner. In equation (17), the very small errors and the very large errors are not important.

4. Violations of the Conditions

We will not show an example of a reliable regression in this paper. First, they are hard to come by, however, more importantly, our aim is to illustrate violations of the regression conditions as we introduced them in the previous section. If one condition is violated the resulting equation (16) will be different (assuming it can be calculated). In cases where certain conditions are not satisfied, we still can calculate a best-fit line.

4.1 Weighted Regression

We will use an example where we want to calculate an Archie Relation: $FRF=a.\phi^{-m}$. In this case, we have a plot of FRF on a logarithmic scale versus the porosity also on a logarithmic scale as shown below.



The data are measured on core plugs. The resistivity can be measured accurately, while the porosity has some constant error. The linear regression equation reads:

$$Ln(FRF) = -m.\ln(\phi) + \ln(a) \qquad (18)$$

The variable with the error is the porosity. The error in the porosity (ε) is constant. This means that the error in $\ln(\phi)$ is not and depends on ϕ as follows:

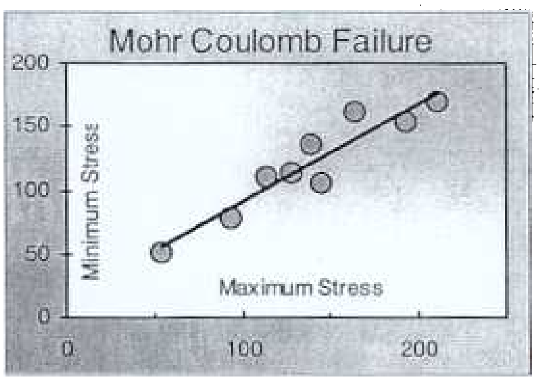
$$\begin{aligned}\ln(\phi+\varepsilon) &= \ln(\phi(1+\varepsilon/\phi)) \\ &= \ln(\phi) + \ln(1+\varepsilon/\phi) \\ &= \ln(\phi) + \varepsilon/\phi\end{aligned}$$

So, the error in $x = \ln(\phi)$ increases as the porosity decreases. A standard regression would be flawed. In regression we need σ^2 to be constant for all data points. So, by applying ϕ^2 as a weight factor we can still do a regression. Because the porosity has the largest error by far, we should treat the porosity as the dependent variable, or do an inverse regression.

4.2 Dependent Variables

When the two variables are dependent, regression is difficult. However, the solution is straightforward. The practical method is to eliminate the effect. The variables should be transformed in such a manner that they become independent. We will show this on the regression of the minimum stress to the maximum stress on the Mohr circle for stress failure for the rock. In the graph below the minimum stress is plotted versus the maximum stress and a standard regression line is shown.

It is interesting to know that this type of regression used to be standard in this application.



We should look at what the minimum and maximum stresses actually are:

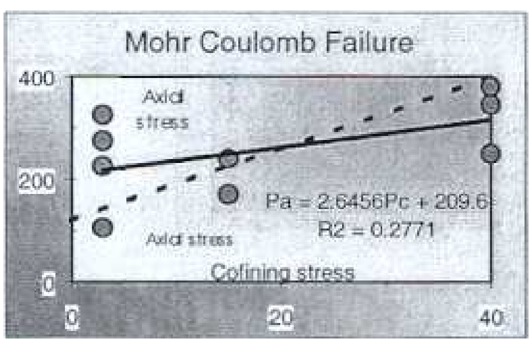
$\sigma_{min} = (Pa - Pc)/2$ (20a)

$\sigma_{max} = (Pa + Pc)/2$ (20b)

with Pa : axial fracture stress
Pc : confining stress

In the case that Pc = 0 we could expect a perfect fit. Usually Pc is much smaller than Pa. This result in some scatter around the regression line, but still the fit is good, even if Pa and Pc are totally random. This is caused by the dependency of the variables and is not real.

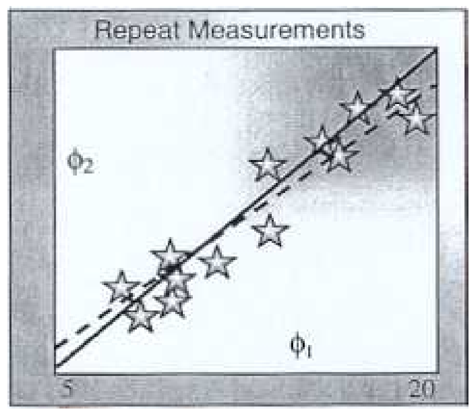
The situation changes when we plot Pc versus Pa as in the graph below.



The scatter in the points increases dramatically and the regression coefficient is significantly lower. We have also shown the regression line obtained in the previous plot. This line was actually close to thee eyeballed line drawn by the expert, who measured and used this data on a regular basis. He did not trust the three high points at the low value of Pc, and the two low points at the high value of Pc. In fact, he rejected five out of nine data points. We are not in a position to debate his decision to ignore more than half of his data. Just looking at the way this data is treated from a statistical point of view, the situation is more than hair raising. One might even question why these measurements were made in the first place. If the rejection of data is common practice, the analysis could well be very biased and this bias may again direct the grounds for rejection. The onus is on the expert, to circumvent this vicious bias.

4.3 Example with Dependent Errors

Errors in the dependent and independent variable can sometimes be correlated to each other. This often happens in so called repeat measurements, when part of the measurement sequence is shared by the repeated measurements. A case in point is porosity estimates from drill cuttings using Archimedes Principle. This is shown in the plot below (reproduced by hand).



The regression line is shown as the dotted line and has a correlation coefficient of r^2=0.82. The covariance between the errors in y and x is estimated as $Cov(d_i,e_i)/\sigma_x\sigma_y = 0.5$. Correction for a non-zero covariance is possible, as shown in the next section, and leads to a slope of 0.97. In fact this example was used to show that the covariance was about 0.5. The slope was found 0.97, but the slope for the inverse regression was found 1.18.

4.4 Solution for Dependent Errors

The structural model allows calculating the best-fit line in the case where we know the following conditions:

$$\lambda^2 = \sigma_x^2/\sigma_y^2$$

$$\rho = Cov(d,e)/(\sigma_x \cdot \sigma_y)$$

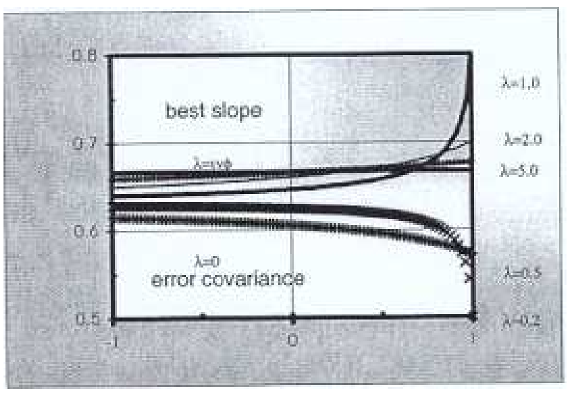
The equation for m is quadratic and somewhat complicated:

$$m^2(m_1-\rho\lambda)-m(m_1m_2-\lambda^2)-(m_1\lambda^2-\rho\lambda m_1m_2)=0 \qquad (21)$$

with

m_1 = SXY/SXX standard regression slope
(assumes that y has all the error or $\sigma_x^2=0$)
m_2 = SYY/SXY inverse regression slope
(assumes that x has all the error or $\sigma_y^2=0$)

We have plotted the calculated slope in the graph below as a function of the covariance (ρ) for various error ratios (λ). We assumed SXX=5, SXY=3 and SYY=2. Or: m_1 = 3/5 , m_2 = 2/3 and r^2 = 0.9.



The available data does not change. What does change is our additional knowledge about the data. With the extra information, the slope of the best-fit line changes. When we lack this extra information, we are also unsure of the slope of the best-fit line. The importance of this graph is that it clearly shows that the slope of the best-fit line is uncertain when we lack the necessary information.

This conclusion has nothing to do with the so-called confidence interval for the slope, as can be calculated from the available data. The confidence interval is related to the scatter around the line and becomes smaller with the number of data points, or the precision improves with the amount of data. The confidence interval follows straightforward from the standard regression approach as:

$$\text{m-st.dev} = +/- \; 2m\sqrt{(1/r^2-1)/n} \qquad (22)$$

r^2 : correlation coefficient (SXY^2/SXXSYY)

Our present discussion centers on the premise that the standard regression approach is often invalid, because the data does not obey the necessary conditions. The uncertainty in the slope due to the lack of information is a bias error, and does not change when more data points are measured. There are not many people who have a reasonable understanding of the shape of the error distribution, whether errors are correlated, whether variables are dependent, what the proper size of the errors are, whether the sampling was done properly etc. This knowledge is hard to come by and usually expensive.

The graph shows that even with a relatively minor change in the error ratio or an error covariance that is non zero the slope changes. Therefore, the slope of the best-fit line is uncertain.
We do not discuss the confidence limits of the regression line (precision) but the reliability of the slope itself (accuracy, bias error).

4.5 Recommended Slope Calculations

When we have some approximate knowledge of the errors associated with the x and y variables and all conditions are reasonably satisfied we calculate the slope from somewhat adjusted equations.

Again we use m_1 = SXY/SXX , m_2 = SYY/SXY

1. if we know $q_x = \sigma_x^2$/SXX while
   $q_x < 1$ and $m_2 > m_1/(1-q_x)$
   use $m = m_1/(1-q_x)$
   else $m = m_1$

2. if we know $q_y = \sigma_y^2$/SYY while
   $q_y < 1$ and $m_1 > m_2(1-q_y)$
   use $m = m_2.(1-q_y)$
   else $m = m_2$

3. if we know σ_x^2 and σ_y^2 while
   $q_x > 1$ or $q_y > 1$ or $m_1/(1-q_x) > m_2.(1-q_y)$
   use $m = \sqrt{SYY/SXX}$
   else $m = m_{bi}$

4. if we know $\lambda^2 = \sigma_x^2/\sigma_y^2$
   use always $m = m_{bi}$

5. nothing eyeballed line

m_{bi} is the slope for bivariate regression, which is calculated from the following equation:

$$m_{bi} = t \pm \sqrt{t^2 + \lambda^2}$$

Where

$$t = (m_2 - \lambda/m_1)/2$$

In Table I, we show a number of different cases leading to different slopes. It is important to realize that for these calculated slopes all the regression conditions are assumed satisfied. The only difference between the slopes is that we have some extra information that is ignored in standard regression.

$Q_x = \sigma_x^2/SXX$	$q_y = \sigma_y^2/SYY$	$\lambda^2 = \sigma_x^2/\sigma_y^2$	m
0.25	-	-	2.133
0.50	-	-	2.500
1.50	-	-	2.500
-	0.25	-	1.875
-	0.50	-	1.600
-	1.50	-	1.600
1.50	1.125	(1/3)	2.073
2.20	1.65	(1/3)	2.000
0.50	1.125	(1/9)	1.820
-	-	1/3	2.073
-	-	3	2.429
-	-	1	2.308
SXX=100	**SYY = 400**	**SXY=160**	**r^2 = 0.6**

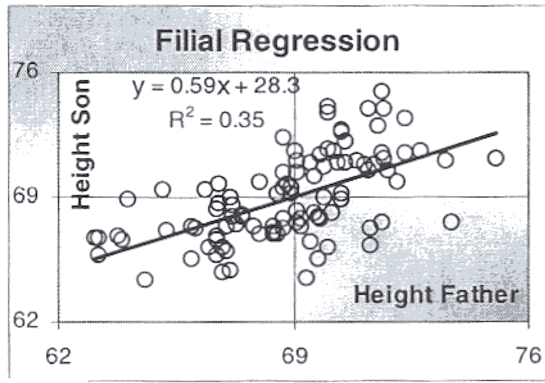
Table I : Different slopes for different conditions.

It is very evident from the table that the slope varies with the availability of extra information, while all the conditions for normal regression are satisfied. Here, we are not dealing with deviations from the regression conditions as discussed in section 4.4. Now, everything is all right and still the slope can be different.

The recommended slopes, as given in this section, are directly derived from the solution for the structural model. The recommended slopes always fall in between m_1 and m_2.

4.6 Incorrect Usage

Regression can lead an unwary user easily astray as happened to Sir Francis Galton, who actually introduced the name regression. He studied the height of the son in relation to the height of his father. Similar data is shown for employees of a mine in Newcastle. He observed that the predicted height of the son was closer to the average than the height of the father. A tall father tends to have a son who is smaller than him and a small father tends to have a son who is larger than him. He called this incorrect observation "the law of filial regression". The name regression has stuck ever since. *His observation is an artifact of regression.* He should also have studied the inverse relation i.e. the height of the father in relation to the height of the son. Then he would have observed, that smaller sons would tend to have fathers who are taller than they are and tall sons would tend to have fathers who are smaller than they are. Both conclusions cannot be correct at the same time Comparing the distributions of the height of fathers with the height of sons, they appear to be identical, as if taken from the same overall distribution. The data is limited to persons of heights between 62 and 74 inches (156 and 187 cm) and ages of 22 and 32 as recorded by the mine, to avoid outliers.



The inverse regression of height father versus height son would also yield a slope of 0.59, just as in the graph above. ***Reduced-regression*** is appropriate in this case, where errors in both variables are the same. This yields a slope of 1.00, which means father and sons do not differ in height

Similar regression effects occur when we compare the effect of well treatment by comparing pre- and post test results. If we treat only the wells with low productivity, and compare productivity levels before and after the treatment, we always will observe a statistically significant improvement from standard statistical treatments, even when the effects are completely random. The reason is that just as with fathers and sons, the low scores tend to regress back to the mean and so yield higher scores on a post-test. We need to control this regression effect, whenever we are going beyond just predicting a y-value when knowing the x-value.

4.7 Extrapolation Using Polynomials

Polynomials are easy to use in regression, but apart from interpolation, they carry potential hazards. They should never be used to extrapolate beyond the range of the data, because they tend to explode. They are also not very useful in understanding the physical meaning of the regression relation. Again, as long as they are used to predict Y from X, they are all right.

5. Discussion

The purpose of this paper is not to dissuade any potential user to apply statistics and regression. Our aim was to present a number of arguments to show that regression lines are not the final truth at all. A substantial amount of additional information is required before we may accept a regression line as being better than a good eyeballed line. The justification for the wide application of standard regression is based mainly on two arguments:

1. Gauss-Markov theorem, which shows that the standard regression slope actually is the best linear and unbiased estimator for the slope.
2. The regression theory is very concise, elegant and straightforward.

The conditions for the Gauss-Markov theorem are weak, which means that it has a wide field of application. It is usually presented as a strong argument. However, it does not provide a strong argument at all. The weakness of the theorem lies in the linear and unbiased estimator property. All other linear unbiased estimators ignore some data, and so they can not be as good. In effect the Gauss-Markov theorem is of minor importance. All the improved slopes we have discussed are biased, or when the distribution is not normal, they are not linear. They still yield better estimates, sometimes much better estimates.

The standard least squares regression theory is indeed quite concise, elegant and well developed. The amount of theory available for standard regression is very much larger than for non-standard regression. However, this does not necessarily give us a better line. The calculations for not normal distributions involve quite often sorting and iterative operations. They are much more involved and the equations can become complicated. They require the use of computers. The algorithms are generally not available in software packages and have to be written and tested by the user, which is a significant hurdle for most of us.

The concise treatment of non-linear and biased estimators is beyond the scope of this paper. We are not trying to bring the reader on the path of methods like robust regression. Nearly all of the arguments presented in this paper for normal regression are equally applicable to other regression techniques. We only wanted to demonstrate how much one has to know about his data before one can be confident that the results are better than an educated guess. How the results are going to be applied is also important.

6. Summary

We have shown that there are a number of conditions to be fulfilled before a proper regression can be attempted. Next we introduced seventeen conditions to be satisfied in order to solve the regression problem. The solution leads to three equations with four unknowns. The user has to introduce at least one auxiliary relation to be able to calculate the "best fit" line. This auxiliary relation leads to a variety of regression techniques, for data that obeys all the conditions and prerequisite conditions.

Usually we are unsure about the degree in which our data conforms to these requirements, which means that we are also unsure about how well the calculated line conforms to the "best fit" line. The calculated line becomes a guess and is most often not better or worse than a good eyeballed line.

When we want to use regression to estimate physical parameters, we are in for a lot of work, because simple standard regression will most often yield biased results.

Literature

No textbook or paper discusses regression as we have here. Books that consider some aspects are:

1. Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P., Numerical Recipes, 2nd Ed, Cambridge University Press. (1995).

2. Strasser H., Mathematical Theory of Statistics (1985).

3. Balakrishan N., Roa Calyanpundi R., Handbook of Statistics Volumes 1 to 18 (1984-1998).

4. Meeker W., Escobar L., Statistical Methods for Reliability Data (1998).

5. Rao Calyampundi R., Toutenberg H., Linear Models, Least Squares and Alternatives (1995).

The first book is a must have for every engineer. Not only for the algorithms, but also for the no nonsense discussion preceding them. The third book "The Handbook of Statistics" has components of this paper strewn throughout various volumes. It should be in your library.