Facts are Stubborn, Statistics are more Pliable

R. E. (Gene) Ballay, PhD
WWW.GeoNeurale.Com

Mark Twain said it, and while we all realize it, the fact remains that in a busy environment the implications can slip past us. And ***the risk is compounded when one recognizes that the default algorithms / display formats for some oilfield data, may lend itself to an improper numerical evaluation.***

Linear Regression

The common linear regression algorithm is based upon minimizing the sum of the squares of the deviation between the 'best fitting line' versus the 'observations'; ***the sum of the square of the residuals.***

Once the 'residual' is quantified, the mathematics are straight-forward and the computation is readily available in many software packages. These User Friendly packages, however, may very well assume the User is aware of the implications of how the residual is calculated, and the resulting consequences, and that the User has therefore set up the parameter determination appropriately. This is not always the case.

The basic slope - intercept form of the linear relation is

$$y - y(avg) = m\,[x - x(avg)]$$

where y(avg) & x(avg) are the mean values, and "m" is the slope.

There are, in fact, an infinity of solutions to this formulation, differing one to another by how the 'residual' is quantified.

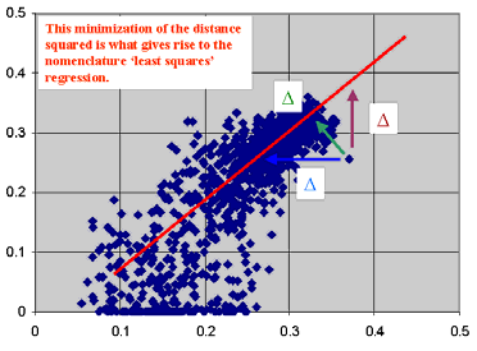
If 'residual' is calculated in the vertical direction, one is said to have done a 'y on x' regression. If 'residual' is referenced in the horizontal direction, an 'x on y' regression results.

Should one seek a 'middle of the road' residual, corresponding to a residual that is perpendicular to the best-fitting line, we have a Major Axis regression. Unfortunately, while this 'middle of the road' has an intuitive attraction, the results are dependent upon the units of the attributes involved: Figure 1.

If, for example, one is correlating core porosity against bulk density, and the units of either attribute are changed (decimal to fraction, oilfield to

Figure 1

- ***Four linear relations are most commonly referenced***
- ***y on x*** which minimizes the square of the ***vertical*** distance from the best-fitting line
- ***x on y*** which minimizes the square of the ***horizontal*** distance from the best-fitting line
- ***Major axis*** which minimizes the square of the ***perpendicular*** distance from the best-fitting line, but for which ***the slope is dependent upon the system of units*** (an undesirable feature)
- ***The fourth common regression protocol is Reduced Major Axis***



metric, etc), the numerical results will change; not an acceptable behavior.

Unit independence is achieved by working with 'normalized' or 'reduced' units, referencing all calculations to the respective standard deviations (x ➔ x/Sigma_x, y ➔ y/Sigma_y); Reduced Major Axis Regression.

Each of these 'infinity of possible linear regressions' corresponds to exactly the same correlation coefficient (which is an attribute of the calibration dataset), but can differ significantly one to the next in estimated results.

The slope of the three most common (simple) linear regressions are then as follows (with 'r' the correlation coefficient and σ the standard deviation).

Y on x minimizes the square of the vertical distance; ***m = r * σ(y) / σ(x).***

X on y minimizes the square of the horizontal distance; ***m = (1 / r) *[σ(y) / σ(x)].***

Reduced Major Axis minimizes the square of the perpendicular distance; ***m = σ(y) / σ(x).***

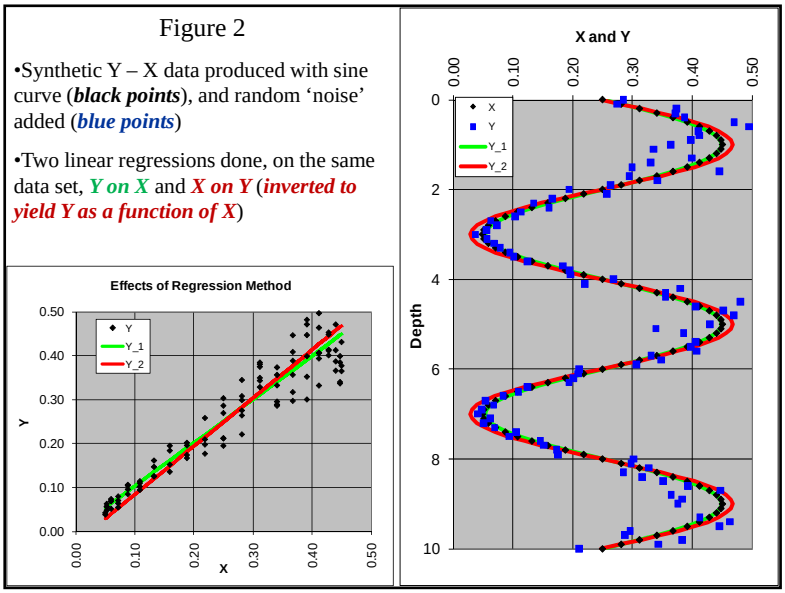
Since 'r' <= 1, the various slopes are related as; m(y on x) < = m(RMA) < = m(x on y). The better the correlation ('r' ➔ 1), the more similar the results; however, with the scatter that is sometimes present in oilfield data, there can be significant differences.

It is important to realize that inversion of a Y on X relation, does not lead to an inverted regression algorithm. That is, regressing Y on X, and then solving for X as a function of Y, is not the same as regression X on Y to start with: Figure 2.

Figure 2

•Synthetic Y – X data produced with sine curve (***black points***), and random 'noise' added (***blue points***)

•Two linear regressions done, on the same data set, ***Y on X*** and ***X on Y*** (***inverted to yield Y as a function of X***)



When setting the regression determination up, one must recognize which (if either) attribute is better known than the other; in effect which 'sum of squared residuals' is to be minimized, and what it is that we seek to estimate.

As Woodhouse points out, some default oilfield display formats can lead us astray. Pressure profiles, as one of several possible examples, are usually displayed with depth on the vertical axis, and measured fluid pressure on the horizontal. Since the default regression algorithm is often Y on X, residuals are then minimized in the vertical (depth) direction, even though depth is relatively well known while it may be that pressure (horizontal axis) is subject to more uncertainly.

An additional concern arises with logarithms, and the tendency to use linear regression to correlate logarithm values; porosity and log(perm) for example. The average of a series of permeability values is not the same as 10 raised to the average of the logarithmic values; Figure 3.

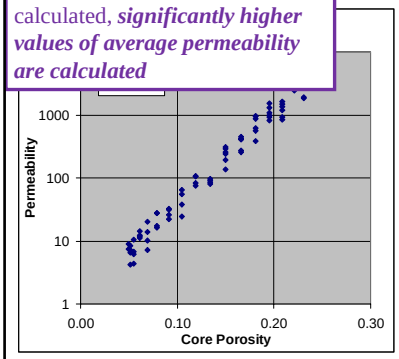
Figure 3

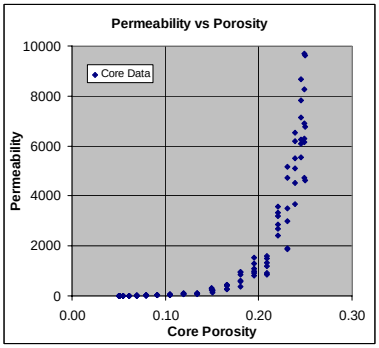
•A ***comparison of Avg(Permeability) and 10^Avg(Log(Perm)) illustrates the discrepancy*** that the Author is referring to

•Woodhouse: this process ***suppresses the high-permeability values***.

•If ***arithmetic means*** for various sections of the *x*-axis are calculated, ***significantly higher values of average permeability are calculated***

All Data	
Avg(Prm)	1957.5
Avg(Log(Prm)	2.53
10^Avg(Log(Prm)	335.2





Saturation - Height

Modern spreadsheets offer analytical power that is often not even generally known, let along used. Statistical summaries are simple, residuals may be explicitly defined in various orientations and minimized in mathematical formulations that are non-linear, Monte Carlo simulations can be performed, and much more.

One common oilfield requirement is a saturation – height relation, which can serve as a vehicle to illustrate the issues, and how to address them. Be aware that the screen shots included herein are of Excel2007 in the backwards compatible mode, and may thus be different than what would appear in a different version.

Following Ding and Pham, we recall that Leverett's J-Function is defined as

$$J(Sw) = [Pc / (\text{Sigma} * \text{Cos}(\text{Theta}))] * \text{Sqrt}[\text{Perm}/\text{Porosity}]$$

Pc and Height Above Free Water Level (HFWL) are related by

$$Pc = HFWL * 0.433 * [\rho(\text{brine}) - \rho(\text{hydrocarbon})] = HFWL * 0.433 * [\Delta(\text{Specific Gravity})]$$

Two (of several) possible formulations of the mathematical relation between Sw and J are

$$J = a * (Sw)\ ^\wedge\ b \text{ and/or } J = a * \exp(b*Sw)$$

"a" and "b" are determined by some "best fit" methodology (and as you have already guessed, we will do it with a spreadsheet).

At the simplest level, the relation between Saturation and Height, can then be written as

$$HFWL = A * (Sw)\ ^\wedge\ B \text{ and/or } HFWL = A * \exp(B*Sw)$$

"A" and "B" reflect the consolidation of the various contributing factors in the Pc ➔ Height estimation ("a" ➔ "A", "b" ➔ "B").

Inserting the details, the two possible expressions above become.

J (1) =

= a * (Sw) ^ b

= [Pc / (Sigma * Cos(Theta))] * Sqrt[Perm/Porosity]

= {HFWL*0.433*[Δ(Specific Gravity)]}/ (Sigma*Cos(Theta))]*Sqrt[Perm/Porosity]

and/ or

J (2) =

= a * exp(b*Sw)

= {HFWL*0.433*[Δ(Specific Gravity)]}/ (Sigma*Cos(Theta))]*Sqrt[Perm/Porosity]

At application (reserves calculation, simulator initialization, etc), the relation will often be inverted to allow calculation of Sw, given HFWL.

$$Ln[J(1)] = Ln(a) + b * Ln(Sw) \text{ and / or } Ln[J(2)] = Ln(a) + b * Sw$$

As a specific, complete example (continuing to follow Ding), the second formulation would be written as below.

Sw =

= [Ln(J(2) – Ln(a)] / b

= {Ln{{HFWL*0.433*[Δ(SG)]}/ (a*Sigma*Cos(Theta))]*Sqrt[Perm/Phi]}}/b

In practice, there will be "uncertainties" in the various parameters, and so a live-linked spreadsheet becomes a useful platform within which to perform the parameter determination (curve fit).

Our oilfield requirement is to determine "a" and "b" in the physically appropriate manner.

- Work with the actual mathematical algorithm, rather than some simplification thereof (ie minimize the actual residuals, not for example the residuals of the logarithms).
- Orient the residuals properly. In general, at some specific capillary pressure, there will be a range of Sw observations, even after rock typing has been imposed. The preferred residual minimization is then between observed and predicted Sw, and not along the vertical HFWL axis.

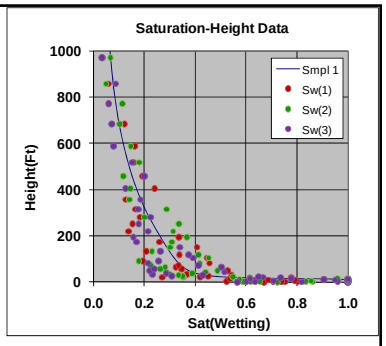
The procedure is illustrated by taking an actual capillary pressure curve, converting to reservoir conditions (Pc ➔ HFWL, per live-linked attributes and relations) ***and then adding random noise about that actual curve to simulate the consolidation of several samples***; Figure 4.

Figure 4

Sw(Noise) = Random*(m * X + b))	
0.75 LE Random LE 1.25	
RAND()*(b-a)+a	
a	1.5
b	0.5

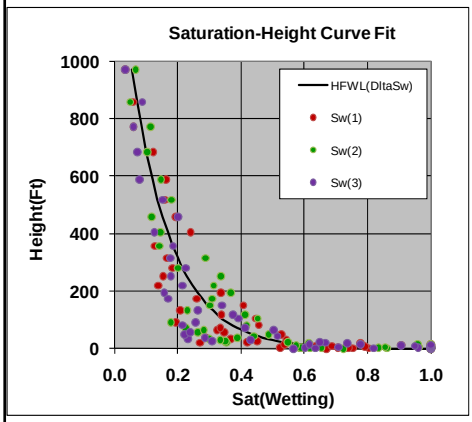
•The procedure is illustrated by taking an actual capillary pressure curve and adding random noise about that curve to simulate the consolidation of several samples.

•Three samples (1, 2 & 3) are created, each with a different Random Seed, yielding a realistic scatter.

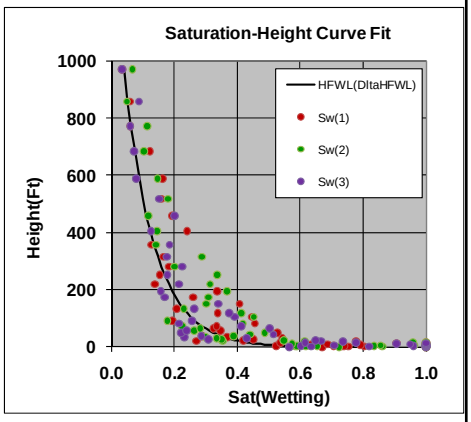


Three samples are created, each with a different Random Seed, yielding a realistic scatter.

Figure 5



Residuals calculated as
Sw(Model) – Sw(Actual)
Sw(Mod) calculated from Solver Results



Residuals calculated as
HFWL(Model) – HFWL(Actual)
Sw(Mod) calculated from Solver Results

•At left, the best-fitting line visually passes through the horizontal distribution of values, while at right the best-fitting line passes through the vertical distribution of values

The Excel Solver function is then used to minimize the sum of the square of the residuals with respect to the best fitting line, in two orientations; Figure 5.

Residuals have been calculated vertically (this might very well be the default orientation in a routine linear regression package) and horizontally.

Visually, we observe that if the residuals are minimized along the horizontal Sw axis, the best-fitting line honors the 'middle of the road' in an Sw sense. If, however, the residuals are minimized vertically, the 'middle of the road' is HFWL and a different set of parameters result. Physically, we will typically regard HFWL as better known than Sw (which is a consolidation of multiple measurements, even after rock typing has been done, with the capillary pressure well-defined).

By personally coding our local evaluations we reinforce our understanding of the basic physical phenomena and constraints, and furthermore build a spreadsheet skill set. Sharing observations and concerns with Colleagues, along with the spreadsheets, will compound the benefit.

Additional Excel Applications

While there are a number of commercially available spreadsheet applications to address common oilfield requirements, there is merit to at least considering local development.

- The spreadsheet development process will mentally impress the various, key assumptions.
- Our spreadsheet skills are honed, with benefit in other areas.
- By sharing these spreadsheets with one another, it is likely that our understanding of both the spreadsheet, and the physical process it seeks to model, will be enhanced.

Spreadsheet applications developed to date include.

- Differential analysis of Sw(Archie) to identify which attribute has the greatest uncertainty impact. Ballay (March 2009).
- Monte Carlo simulation of Sw(Archie) to identify which attribute has the greatest uncertainty impact. Ballay (July 2009).
- Thomeer Model of Dual Porosity Capillary Pressure Curves. Ballay (December 2009).

References

Adams, S. J. Quantifying Petrophysical Uncertainties. Asia Pacific Oil & Gas Conference and Exhibition, Jakarta. April 2005.

Ballay, Gene. Risky Business. March 2009. www.GeoNeurale.com.

Ballay, Gene. Rolling The Dice. July 2009. www.GeoNeurale.com.

Ballay, Gene. Split Personality. December 2009. www.GeoNeurale.com.

Bowers, M. C. & D. E. Fitz. A Probabilistic Approach to Determine Uncertaint in Calculated Water Saturation. Dialog; 8 April 2003. SPWLA 41st Annual Logging Symposium; June 2000.

Bryant, Ian and Alberto Malinverno, Michael Prange, Mauro Gonfalini, James Moffat, Dennis Swager, Philippe Theys, Francesca Verga. Understanding Uncertainty. Oilfield Review. Autumn 2002.

Burnie, Steve. Error / Uncertainty and The Archie Equation. Insight : Canadian Well Logging Society. January 2004.

Case Western Reserve University. Appendix V of the Mechanics Lab Manual, Uncertainty and Error Propagation (available on-line).

Chen, C and J. H. Fang. Sensitivity Analysis of the Parameters in Archie's Water Saturation Equation. The Log Analyst. Sept – Oct 1986.

Cheng, C. L. and J. W. Van Ness. Statistical Regression With Measurement Error. Arnold. Oxford, UK (1999) 16.

Clerke, Edward and Harry W Mueller, Eugene C Phillips, Ramsin Y Eyvazzadeh, David H Jones, Raghu Ramamoorthy & Ashok Srisvastava. Application of Thomeer Hyperbolas to decode the pore systems, facies and reservoir properties of the Upper Jurassic Arab D Limestone, Ghawar Field, Saudi Arabia: A "Rosetta Stone" approach. GeoArabia, Vol 13 No 4 2008.

Clerke, Ed. Permeability, Relative Permeability, Microscopic Displacement Efficiency and Pore Geometry of M_1 Bimodal Pore Systems in Arab D Limestone. SPE Middle East Oil and Gas Show. Bahrain. March 2007.

Clerke, Ed. Beyond Porosity-Permeability Relationships-Determining Pore Network Parameters for the Ghawar Arab-D Using the Thomeer Method. 6th Middle East Geoscience Conference. Bahrain. March, 2004.

Denney, D. Quantifying Petrophysical Uncertainties. SPE JPT. September, 2005.

Ding, Sheng and Tai Pham. An Integrated Approach For Reducing Uncertainty In The Estimation Of Formation Water Saturation And Free Water Level In tight Gas Reservoirs – Case Studies. Society of Core Analysts 2002-41.

Ding, Sheng and Tai Pham & An Ping Yang. The Use Of An Integrated Approach In Estimation Of Water Saturation And Free Water Level In Tight Gas Reservoirs: Case Studies. SPE Annual Technical Conference And Exhibition. October, 2003. Denver, Colorado.

Excel Tips. http://people.stfx.ca/bliengme/exceltips.htm.

Freedman, R. And B. Ausburn. The Waxman-Smits Equation of Shaly Sands: I, Simple Methods of Solution, II Error Analysis. The Log Analyst. 1985.

Hill, T. & P. Lewicki (2007)
Statistics, Methods and Applications. StatSoft, Tulsa, OK
http://www.statsoft.com/textbook/stathome.html

Hook, J. R. The Precision of Core Analysis Data and Some Implications for Reservoir Evaluation. SPWLA 24th Annual Symposium, June 27-30, 1983.

Limpert, L. and W. Stahel & M. Abbt. Log-normal Distributions across the Sciences: Keys and Clues. BioScience, Vol 51 No 5, May 2001.

LSU. Probabilistic Approach to Oil and Gas Prospect Evaluation Using the Excel Spreadsheet. Found with Google, Author n/a. http://www.enrg.lsu.edu/pttc/.

Sweeney, S. A. and H Y Jennings Jr: The Electrical Resistivity of Preferentially Water-Wet and Preferentially Oil-Wet Carbonate Rock, Producers Monthly 24, No 7 (May 1960): 29-32

Jensen, J. L. et al. Statistics For Petroleum Engineers and GeoScientists. Elservier. Amsterday (2002).

Leverett, M. C. Capillary Behavior In Porous Solids; Petroleum Transactions of AIME (1941); 142; 152-169.

Linear Regression. http://en.wikipedia.org/wiki/Linear_regression.

Mr Excel. http://www.mrexcel.com/.

Voss, David, 1998, Quantitative Risk Analysis: John Wiley and Sons, New York.

Woodhouse, Richard. Statistical Regression Line-Fitting In The Oil & Gas Industry. PennWell. Tulsa (2003) 8, 26.

Woodhouse, Richard. Developments in Regression Line Fitting; Improved Evaluation Equations by Proper Choices Between Satististical Models. SPE Distinguished Author Series. December 2005.

Biography

R. E. (Gene) Ballay's ***34 years in petrophysics*** include ***research and operations*** assignments in Houston (Shell Research), Texas; Anchorage (ARCO), Alaska; Dallas (Arco Research), Texas; Jakarta (Huffco), Indonesia; Bakersfield (ARCO), California; and Dhahran, Saudi Arabia. His carbonate experience ranges from individual Niagaran reefs in Michigan to the Lisburne in Alaska to Ghawar, Saudi Arabia (the largest oilfield in the world).

He holds a ***PhD*** in ***Theoretical Physics*** with ***double minors in Electrical Engineering & Mathematics***, has ***taught physics in two universities***, ***mentored Nationals*** in Indonesia and Saudi Arabia, published ***numerous technical articles*** and been designated ***co-inventor on both American and European patents***.

At retirement from the Saudi Arabian Oil Company he was the senior technical petrophysicist in the Reservoir Description Division and had represented petrophysics in three multi-discipline teams bringing on-line three (one clastic, two carbonate) multi-billion barrel increments. Subsequent to retirement from Saudi Aramco he established Robert E Ballay LLC, which ***provides physics - petrophysics consulting services.***

He served in the US Army as a Microwave Repairman and in the US Navy as an Electronics Technician, and he is a USPA Parachutist and a PADI Dive Master.