The Jungle of Linear Regression Revisited

Lorentz JÄNTSCHI[1], Sorana-Daniela BOLBOACĂ[2]

[1] *Technical University of Cluj-Napoca, Romania*
[2] *"Iuliu Haţieganu" University of Medicine and Pharmacy, Cluj-Napoca, Romania*
http://lori.academicdirect.org, http://sorana.academicdirect.ro

Abstract

Simple linear regression is reviewed. Some well known facts are analyzed from different approaches. Some new formulas and equations are posted and discussed.

Keywords

Simple Linear Regression, Least Squares Method, Independent and Dependent Variables

Hypothesis

Let's assume that we have two series of experimental measurements X = $x_1 \ldots x_n$ and Y = $y_1 .. y_n$ on which we suppose that a linear dependence exists:

X, Y linear dependent (1)

Algebraic approach

The general formula for a linear dependence can be written as:

$$aX + bY + c = 0 \Leftrightarrow aX + bY = -c \quad (2)$$

In terms of linear algebra our system (2) has three unknowns (a, b, and c) for only two known (X and Y). In order to provide a finite solution we must reduce the number of unknowns. Let us analyze the values of coefficients. Eight cases are presented in Table 1, for a, b, and c in terms of (0, ≠0):

Table 1. Cases for a, b, and c

Cases	1	2	3	4	5	6	7	8
a	=0	≠0	=0	≠0	=0	≠0	=0	≠0
b	=0	=0	≠0	≠0	=0	=0	≠0	≠0
c	=0	=0	=0	=0	≠0	≠0	≠0	≠0

Case 1 is the trivial case (0 = 0, and it fit for any (X,Y) pair of data). Case 5 goes to an impossibility (c = 0 for c≠0). Cases 2 (aX = 0), 3 (bY = 0), 6 (aX + c =0), and 7 (bY + c = 0) are in disagreement with hypothesis (1).

For further discussions it remains only cases 4 (aX + bY = 0) and 8 (aX + bY + c = 0). Note that in these cases neither of coefficients is null. The following table contains variants of the above described two cases:

Table 2. Linear dependences for (X, Y)

aX+bY=0 (Case 4)			aX+bY+c=0 (Case 8)			
0	1	2	0	1	2	3
aX+bY=0	X+nY=0	mX+Y=0	aX+bY+c=0	X+nY+p=0	mX+Y+p=0	mX+nY+1=0

In the table 3, elementary transformations to the equations from table 2 were applied. In addition, remarks are made.

Table 3. Cases analysis (variants of cases from table 2)

Variant	Equation	Remarks - statistical imposed assumptions
4.0	aX+bY=0	particular case of 8.0
4.1	X=-nY	X: dependent variable; Y: independent variable
4.2	Y=-mX	Y: dependent variable; X: independent variable
8.0	aX+bY=-c	both X and Y are dependent/independent variables
8.1	X=-nY-p	X: dependent variable; Y: independent variable
8.2	Y=-mX-p	Y: dependent variable; X: independent variable
8.3	mX+nY=-1	particular case of 8.0

Note that we usually do (with any of well-known software’s) the variants 4.1 & 4.2 and 8.1 & 8.2. So, the most interesting to analyze are 4.0 and 8.3, which are particular cases of 8.0.

Usually, we assign as “dependent variable” a variable which comes from experiment and which is affected by experimental random errors, and we assign as “independent variable” a variable (which comes from experiment (?) - this statement open a discussion) and which is not affected by experimental random errors.

Nevertheless, what we must to do when both variables are affected by errors? Definitely, as we already seen in table 3, is not a good idea to use any of 4.1, 4.2, 8.1 or 8.2 assumptions.

Let us go back to our hypothesis (1) and let us set all our cases variants as f(vars,coefs) = 0 (table 4, as in table 2):

Table 4. Linear regression equation as function

Variant	Equation	Function for f = 0
4.0	aX+bY=0	f({X,Y},{a,b}) = aX+bY
4.1	X+nY=0	f({X,Y},{b}) = X+bY
4.2	mX+Y=0	f({X,Y},{a}) = aX+Y
8.0	aX+bY+c=0	f({X,Y},{a,b,c}) = aX+bY+c
8.1	X+nY+p=0	f({X,Y},{b,c}) = X+bY+c
8.2	mX+Y+p=0	f({X,Y},{a,c}) = aX+Y+c
8.3	mX+nY+1=0	f({X,Y},{a,b}) = aX+bY+1

As it can be observed from table 3, all other are particular cases of 8.0. So, we will discuss all in general related to 8.0.

A sum function can be constructed in terms of deviations from the model:

$$S = \sum|f(\{x_i,y_i\},\{a,b,c\})|^k,\ k > 0 \qquad (3)$$

where sum are applied for all experimental measurements (from 1 to n). Note that in order to be consistent the definition (3), the modulus function must be used.

In terms of estimation, a, b, and c are called model parameters, and X, Y are dependent and/or independent variables (see also [1]). In terms of analysis, S is a function that depends on a, b and c as variables (unknown values) and X, Y and k as fixed (known) values. In terms of algebra, not all variables are allowed to vary in order to find a non-banal solution (null values - see also Eq. 4). Therefore, we must set one parameter.

Let us rewrite (3) in general case 8.0:

$$S(a,b,c) = \sum|aX+bY+c|^k,\ k>0\text{, one of a, b and c is set} \qquad (4)$$

How will affect a single measurement error the value of S? - see (5). Let us take a term of S:

$$S_i(a,b,c) = |ax_i+by_i+c|^k \qquad (5)$$

If $x_i = x_{0i}+erx_i$, then $S_i = |ax_{0i}+aerx_i+by_i+c|^k = |aerx_i+(ax_{0i}+by_i+c)|^k$. Thus, an absolute error of xi (erx_i, not absolute in term of modulus, absolute in term of error, with same measurement unit with X) will be propagated as absolute error of S.

In some cases, we know more about our experimental errors.

Let's say, if we use an absolute method of measurement (such as mass measurement) then our error is an absolute one (in terms of measurement scale), and it remains the same as long as we use the same scale. In these cases, our preferred error expressions must be the absolute error. The opposite case, if we use a relative method of measurement (such as instrumental methods) then our error is a relative one (in terms of calibration accuracy), and it remains the same as long as we use the same calibration. In these cases, our preferred error expressions must be the relative error.

Nevertheless, we have two measured variables! What we have to do? - See (4, 6)

Coming back to the equation (4), we can weight the terms:

$$S(a,b,c) = \sum|a\xi X+b\eta Y+c|^k,\ k>0,\ \xi,\eta \text{ weights (known values)} \quad (6)$$

where $\xi = 1$ if X has absolute errors and $\xi = 1/M(X)$ if X has revalive errors and M(X) is the aritmetic mean of X values, and the same for η and Y.

What we have to do now? - To find the values of a, b, and c by imposing to S to be lowest:

$$S(a,b,c) = \min \quad (7)$$

First pure mathematical problem comes now. Why? - Because we have modulus function $f(\cdot) = |\cdot|$ in our formula, which is a continue and derivable function but with discontinue derivative. This is the reason for which we prefer even values for k (usually 2, we will see why). Anyway, (7) can be expressed in terms of derivatives:

$$\partial S/\partial coefs = 0 \quad (8)$$

The equation (8) is a system of equations, which for $k\neq2$ is not linear. Taking as example the case 4.2 for k=4 and solving of (9) is equivalent to solving of:

$$m^3(\sum X^4)+3m^2(\sum X^3Y)+3m(\sum X^2Y^2)+(\sum XY^3)=0 \quad (9)$$

For cases that are more general or for higher k values, solving of (8) leads to equations that are far more complicated. This is the reason for which we prefer k = 2. So, we fit now on well known "minimizing of sum of partial least squares method". Rewriting of (6) for k = 2 leads to:

$$\sum\xi X(a\xi X+b\eta Y+c) = \sum\eta Y(a\xi X+b\eta Y+c) = \sum(a\xi X+b\eta Y+c) = 0 \quad (10)$$

By using M(·) as average operator $M(\cdot) = \sum(\cdot)/n$ equations (10) became:

$$\begin{aligned} &a\cdot\xi^2\cdot M(X^2)+b\cdot\xi\cdot\eta\cdot M(XY)+c\cdot\xi\cdot M(X) = 0 \\ &a\cdot\xi\cdot\eta\cdot M(XY)+b\cdot\eta^2\cdot M(Y^2)+c\cdot\eta\cdot M(Y) = 0 \\ &a\cdot\xi\cdot M(X)+b\cdot\eta\cdot M(Y)+c = 0 \end{aligned} \tag{11}$$

The system (11) is not intended to be solved in its actual form, which it provides assuming the varying of all three parameters only the banal solution (0, 0, 0).

The case 4.1 (X+bY) are obtained from (11.2) when c=0 and a=1:

$$b = -(\xi/\eta)\cdot M(XY)/M(Y^2) \tag{12}$$

The case 4.2 (aX+b) are obtained from (11.1) when c=0 and b=1:

$$a = -(\eta/\xi)\cdot M(XY)/M(X^2) \tag{13}$$

Two remarks are immediate (from and for 12 and 13):

- weighting (ξ and η) does not affect the formulas for coefficients - identic est - the obtained formulas are transparent to weighting;
- is possible to construct another formula which it combine (12) and (13).

Rewriting of (12) and (13) without weighting and including the equation formulas goes to:

$$\begin{aligned} &f(\{X,Y\},\{b\}) = X\text{-}Y\cdot M(XY)/M(Y^2),\ f(\{X,Y\},\{a\}) = X\cdot M(XY)/M(X^2)\text{-}Y, \text{ or} \\ &X(Y) = -\ Y\cdot M(XY)/M(Y^2)\ \&\&\ Y(X) = -\ X\cdot M(XY)/M(X^2) \end{aligned} \tag{14}$$

Inversing the X(Y) and Y(X) functions:

$$X(Y)^{-1} = -X\cdot M(Y^2)/M(XY)\ \&\&\ Y(X)^{-1} = -\ Y\cdot M(X^2)/M(XY) \tag{15}$$

From (14.1 & 15.2) and (14.2 & 15.1) it results that the coefficient can be obtained by applying of a mean function:

$$X(Y) = -\ Y\cdot \text{Mean}(M(XY)/M(Y^2),\ M(X^2)/M(XY)) \tag{16}$$

$$Y(X) = -\ X\cdot \text{Mean}(M(Y^2)/M(XY),\ M(XY)/M(X^2)) \tag{17}$$

But which mean is suitable? - The geometric mean provides same reversed result for:

$$X(Y) = -\ Y\cdot M^{0.5}(X^2)/M^{0.5}(Y^2)\ \&\&\ Y(X) = -\ X\cdot M^{0.5}(Y^2)/M^{0.5}(X^2) \tag{18}$$

Formula (18) it represents a new formula for coefficients calculation. Which case can be assigned to (18)? - Only the remaining one, 4.0:

$$a=\pm M^{0.5}(Y^2),\ b=\mp M^{0.5}(X^2) \tag{19}$$

The cases 8.1 & 8.2 are well known; it will not be discussed here.

The case 8.3 are obtained from (11.1) & (11.2) when c = 1:

$$a\cdot M(X^2)+b\cdot M(XY)+M(X) = a\cdot M(XY)+b\cdot M(Y^2)+M(Y) = 0 \tag{20}$$

Equation (20) leads to:

$$\begin{aligned} a &= (M(Y^2)M(X)-M(Y)M(XY))/(M(X^2)M(Y^2)-M^2(XY)) \\ b &= (M(X^2)M(Y)-M(X)M(XY))/(M(X^2)M(Y^2)-M^2(XY)) \end{aligned} \quad (21)$$

Also from (21) through extension, the following can be assigned to (8.0):

$$\begin{aligned} &X\cdot(M(Y^2)M(X)-M(Y)M(XY))+ \\ &+ Y\cdot(M(X^2)M(Y)-M(X)M(XY))+ \\ &+ (M(X^2)M(Y^2)-M^2(XY))=0 \end{aligned} \quad (22)$$

Null intercept regression

Let us look more closely on case 4 with its sub-cases 4.0, 4.1 and 4.2.

- What we want? - We want a linear regression between X and Y.
- What we know? - We know at least that intercept coefficient is null.
- What we have? - We have at least a equation of type aX + bY = 0.
- What we cannot have? - We cannot have both parameters unknown.

Let us start from 4.0 (aX+bY=0) and apply the average operator. This leads to:

$$aM(X)+bM(Y)=0, \text{ for } aX+bY=0 \quad (23)$$

What is wrong in our suppositions? - Remember, we already obtained some formulas for a and b (eq. 16-19). Answer: nothing is wrong! - Let us go back to table 3 and look more carefully to dependence/independence suppositions - here are the inconsistencies.

Now let us review our results for aX+bY=0:

- aM(X)+bM(Y) = 0 - main result, eq. (23)
  - o if X and Y are independent variables, then from eq. 23 solution is immediate:

$$a = \pm M(Y),\ b = \mp M(X), \text{ for X, Y independent variables} \quad (24)$$

  - o if X is the dependent variable and Y is the independent variable, then (see also 17):

$$a = \text{Mean}(M(Y^2)/M(XY), M(XY)/M(X^2)),\ b = -1, \text{ for } Y=Y(X) \quad (25)$$

  - o if Y is the dependent variable and X is the independent variable, then (see also 16):

$$b = \text{Mean}(M(XY)/M(Y^2), M(X^2)/M(XY)),\ a = -1, \text{ for } X=X(Y) \quad (26)$$

  - o if X and Y are both dependent variables (see also 19):

$$a = \pm M^{0.5}(Y^2),\ b=\mp M^{0.5}(X^2) \quad (27)$$

Few remarks can be made:

- The equations (25)-(27) assume that at least one of followings is (or a transformation applied to the data make it) true: M(X) = 0, M(Y) = 0, Mean(M(Y^2)/M(XY), M(XY)/M(X^2))M(X)-M(Y) = 0, Mean(M(Y^2)/M(XY), M(XY)/M(X^2))M(Y)-M(X) = 0.
- The mean function can be a weighted mean such as:

$$a = (1-f)\cdot M(Y^2)/M(XY) + (f)\cdot M(XY)/M(X^2),\ b = -1 \tag{28}$$

f portion (fraction) of X dependence in Y and (1-f) vice versa (1 ≥ f > 0.5)

- or

$$b = (1-f)\cdot M(X^2)/M(XY) + (f)\cdot M(XY)/M(Y^2),\ a = -1 \tag{29}$$

f portion (fraction) of Y dependence in X and (1-f) vice versa (1 ≥ f > 0.5)

- Near to middle region (f ≈ 0.5) we can use any un-weighted mean. Followings are for Y=Y(X):

$$a = \mathrm{AM}\left(\frac{M(Y^2)}{M(XY)}, \frac{M(XY)}{M(X^2)}\right) = \frac{M(X^2)M(Y^2) + M^2(XY)}{2M(XY)M(X^2)}$$

$$a = \mathrm{GM}\left(\frac{M(Y^2)}{M(XY)}, \frac{M(XY)}{M(X^2)}\right) = \sqrt{\frac{M(Y^2)}{M(X^2)}}$$

$$a = \mathrm{EM}\left(\frac{M(Y^2)}{M(XY)}, \frac{M(XY)}{M(X^2)}\right) = \sqrt{\frac{\left(\frac{M(Y^2)}{M(XY)}\right)^2 + \left(\frac{M(XY)}{M(X^2)}\right)^2}{2}}$$

$$a = \mathrm{HM}\left(\frac{M(Y^2)}{M(XY)}, \frac{M(XY)}{M(X^2)}\right) = \frac{1}{\mathrm{AM}\left(\frac{M(XY)}{M(Y^2)}, \frac{M(X^2)}{M(XY)}\right)} \tag{30}$$

$$a = \mathrm{PM}\left(\frac{M(Y^2)}{M(XY)}, \frac{M(XY)}{M(X^2)}, p\right) = \left(\frac{\left(\frac{M(Y^2)}{M(XY)}\right)^p + \left(\frac{M(XY)}{M(X^2)}\right)^p}{2}\right)^{1/p},\ p \neq 0$$

$$a = \mathrm{AGM}\left(\frac{M(Y^2)}{M(XY)}, \frac{M(XY)}{M(X^2)}\right) = \lim_{m\to\infty} c_m = \lim_{m\to\infty} d_m, \quad \text{where}$$

$$c_m = \mathrm{AM}(c_{m-1}, d_{m-1}),\ d_m = \mathrm{GM}(c_{m-1}, d_{m-1}), \quad \text{and}$$

$$c_0 = \mathrm{AM}\left(\frac{M(Y^2)}{M(XY)}, \frac{M(XY)}{M(X^2)}\right),\ d_0 = \mathrm{GM}\left(\frac{M(Y^2)}{M(XY)}, \frac{M(XY)}{M(X^2)}\right)$$

and note the followings:

$$\min(\bullet,\bullet) \le \mathrm{HM}(\bullet,\bullet) \le \mathrm{GM}(\bullet,\bullet) \le \mathrm{AGM}(\bullet,\bullet) \le \mathrm{AM}(\bullet,\bullet) \le \mathrm{EM}(\bullet,\bullet) \le \max(\bullet,\bullet) \quad (31)$$

$$\min(\bullet,\bullet) \le \mathrm{PM}(\bullet,\bullet,\bullet) \le \max(\bullet,\bullet) \quad (32)$$

$$\begin{aligned} &\min(\bullet,\bullet) = \lim_{p\to-\infty} \mathrm{PM}(\bullet,\bullet,p),\ \mathrm{HM}(\bullet,\bullet) = \mathrm{PM}(\bullet,\bullet,-1),\\ &\mathrm{GM}(\bullet,\bullet) = \lim_{p\to 0} \mathrm{PM}(\bullet,\bullet,p),\ \mathrm{AM}(\bullet,\bullet) = \mathrm{PM}(\bullet,\bullet,1),\\ &\mathrm{EM}(\bullet,\bullet) = \mathrm{PM}(\bullet,\bullet,2),\ \max(\bullet,\bullet) = \lim_{p\to+\infty} \mathrm{PM}(\bullet,\bullet,p) \end{aligned} \quad (33)$$

More, a definition of PPM(·,·) similarly to AGM(·,·) leads to GM(·,·):

$$\begin{aligned} &c_0 = \mathrm{PM}\left(\frac{M(Y^2)}{M(XY)},\frac{M(XY)}{M(X^2)},1\right),\ d_0 = \mathrm{PM}\left(\frac{M(Y^2)}{M(XY)},\frac{M(XY)}{M(X^2)},-1\right),\\ &c_m = \mathrm{PM}(c_{m-1},d_{m-1},p),\ d_m = \mathrm{PM}(c_{m-1},d_{m-1},-p),\\ &a = \mathrm{PPM}\left(\frac{M(Y^2)}{M(XY)},\frac{M(XY)}{M(X^2)}\right) = \lim_{m\to\infty} c_m, d_m = \mathrm{GM}\left(\frac{M(Y^2)}{M(XY)},\frac{M(XY)}{M(X^2)}\right) \end{aligned} \quad (34)$$

Not null intercept formulas

The following table contains the obtained formulas (see algebraic approach section):

Table 5. Linear regression coefficients formulas

Variant	Equation	Function for f = 0	Coefficients

8.0	aX+bY+c=0	f({X,Y},{a,b,c}) = aX+bY+c	a = $M(Y^2)M(X)-M(Y)M(XY)$
			b = $M(X^2)M(Y)-M(X)M(XY)$
			c = $M(X^2)M(Y^2)-M^2(XY)$

8.1	X+nY+p=0	f({X,Y},{b,c}) = X+bY+c	b = - $(M(XY)-M(X)M(Y))$ $/(M(Y^2)-M^2(Y))$
			c = - $(M(Y^2)M(X)-M(Y)M(XY))$ $/(M(Y^2)-M^2(Y))$

8.2	mX+Y+p=0	f({X,Y},{a,c}) = aX+Y+c	a = - $(M(XY)-M(X)M(Y))$ $/(M(X^2)-M^2(X))$
			c = - $(M(X^2)M(Y)-M(X)M(XY))$ $/(M(X^2)-M^2(X))$

8.3	mX+nY+1=0	f({X,Y},{a,b}) = aX+bY+1	a = $(M(Y^2)M(X)-M(Y)M(XY))$ $/(M(X^2)M(Y^2)-M^2(XY))$
			b = $(M(X^2)M(Y)-M(X)M(XY))$ $/(M(X^2)M(Y^2)-M^2(XY))$

Geometrical approach

In the following figure is depicted a (X,Y) plot, with a regression equation line (assigned with Y=aX+c), and a point P_i - of coordinates (x_i,y_i).

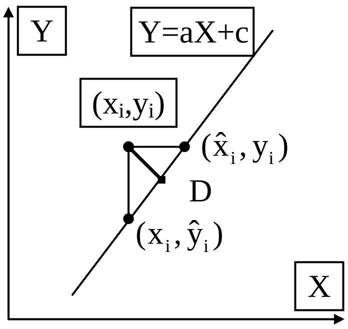

Figure 1. Geometrical interpretation of error estimates

Following are supplementary depicted:

- The intersection of X axis parallel with regression equation line - point of generic coordinates (x_i^{est},y_i) - from assumption that X=bY+c is the regression equation;
- The intersection of Y axis parallel with regression equation line - point of generic coordinates (x_i,y_i^{est}) - from assumption that Y=aX+c is the regression equation;
- The intersection between perpendicularly from P_i to regression equation line.

The followings are true:

- if $S \leftarrow \sum(\hat{y}_i - y_i)^2$ then $a \leftarrow \frac{M(XY)-M(X)M(Y)}{M(X^2)-M^2(X)}$, $c \leftarrow \frac{M(X^2)M(Y)-M(X)M(XY)}{M(X^2)-M^2(X)}$
- if $S \leftarrow \sum(\hat{x}_i - x_i)^2$ then $b \leftarrow \left(\frac{M(XY)-M(X)M(Y)}{M(Y^2)-M^2(Y)}\right)^{-1}$, $c \leftarrow -\frac{M(Y^2)M(X)-M(Y)M(XY)}{M(XY)-M(X)M(Y)}$

It's easy to check that:

$$b_{S\leftarrow\Sigma(\hat{x}_i-x_i)^2} \leftarrow \frac{1}{a_{S\leftarrow\Sigma(\hat{y}_i-y_i)^2}} \text{ and } c_{S\leftarrow\Sigma(\hat{x}_i-x_i)^2} \leftarrow -\frac{c_{S\leftarrow\Sigma(\hat{y}_i-y_i)^2}}{a_{S\leftarrow\Sigma(\hat{y}_i-y_i)^2}} \text{ when } X \leftrightarrow Y \tag{35}$$

which comes also from:

$$Y = aX + c \leftrightarrow X = \frac{1}{a}Y - \frac{c}{a} \tag{36}$$

The equations (35) and (36) prove that the chousing of S = $\sum\left(\hat{x}_i - x_i\right)^2$ for Y = aX + c is equivalent to chousing of S = $\sum\left(\hat{y}_i - y_i\right)^2$ for X = bY + c. So, the use of square $(x_i^{est}-x_i)^2$ is equivalent to case 8.2, and the use of square $(y_i^{est}-y_i)^2$ is equivalent to case 8.1.

If S ← P_iD^2 then:

$$S = \sum(ax_i-y_i+c)/(a^2+1)$$

After calculation of ∂S/∂a and ∂S/∂c it results:

$$a = -\frac{\left(M(X^2)-M^2(X)\right)-\left(M(Y^2)-M^2(Y)\right)}{2\left(M(XY)-M(X)M(Y)\right)}$$

$$\pm\sqrt{\frac{\left(M(X^2)-M^2(X)\right)-\left(M(Y^2)-M^2(Y)\right)}{2\left(M(XY)-M(X)M(Y)\right)}+1} \tag{37}$$

and c = M(Y) - aM(X) for S ← $\Sigma\frac{\left(\hat{y}_i-y_i\right)^2\left(\hat{x}_i-x_i\right)^2}{\left(\hat{y}_i-y_i\right)^2+\left(\hat{x}_i-x_i\right)^2}$

A formula that is even more complicated is obtained when offsets from S are choused to be with a slope of m. In this case, the formula for S is:

$$S \leftarrow \frac{m^2+1}{(a-m)^2}\Sigma(ax_i-y_i+c)^2 \tag{38}$$

When m is independent to both a and c it results:

$$a = \frac{m^2\left(M(X^2)+M^2(X)\right)-2m\left(M(XY)-M(X)M(Y)\right)+\left(M(Y^2)-M^2(Y)\right)\pm\sqrt{\Delta}}{2\left(m\left(3M^2(X)-M(X^2)\right)+\left(M(XY)-M(X)M(Y)\right)\right)}$$

$$\begin{aligned}\Delta = {} & 2m^2\left(4\left(M(XY)-M(X)M(Y)\right)^2-\left(M(Y^2)-M^2(Y)\right)\left(3M(X^2)-5M^2(X)\right)\right) \\ & -8m\left(M(XY)-M(X)M(Y)\right)\left(m^2\left(M(X^2)-M^2(X)\right)+\left(M(Y^2)-M^2(Y)\right)\right) \\ & +m^4\left(M(X^2)+M^2(X)\right)^2+\left(M(Y^2)-M^2(Y)\right)^2, \quad c = M(Y)-aM(X)\end{aligned} \tag{39}$$

Note that for the case of dependence between m and a, ∂S/∂m = 0 can be solved in two ways:

- if m is not a function of a then:

$$\frac{\partial}{\partial m}\left(\frac{m^2+1}{(a-m)^2}\right) = 0 \leftrightarrow m = -\frac{1}{a} \text{ (perpendicular offsets)} \tag{40}$$

- if m is a function of a then:

$$\frac{\partial}{\partial m}\left(\frac{m^2+1}{(a(m)-m)^2}\right)=0 \leftrightarrow a=m+D\sqrt{m^2+1}\text{, C any constant} \tag{41}$$

By replacing of (27) in (24) it results:

$$S \leftarrow \frac{1}{D^2}\Sigma(ax_i-y_i+b)^2 \tag{42}$$

Is the value of Pearson affected by how the slope and intercept are calculated?

The full question is: assuming that we have a measured X and Y and we want to estimate Y by using of regression equation which we obtained, how the calculated slope and intercept affects the Pearson r between measured Y and estimated Ŷ? - The answer is No, see below.

Let us take the squared Pearson coefficient between Y and Ŷ:

$$r^2(\hat{Y},Y)=\frac{\left(M(\hat{Y}Y)-M(\hat{Y})M(Y)\right)^2}{\left(M(\hat{Y}^2)-M^2(\hat{Y})\right)\left(M(Y^2)-M^2(Y)\right)} \tag{43}$$

By substituting of Ŷ = aX+c in (43) it results:

$$\begin{aligned} r^2(\hat{Y},Y)&=\frac{\left(M((aX+b)Y)-M(aX+c)M(Y)\right)^2}{\left(M(aX+c)^2-M^2(aX+c)\right)\left(M(Y^2)-M^2(Y)\right)}= \\ &=\frac{\left(M(aXY+cY)-M(aX+c)M(Y)\right)^2}{\left(M(a^2X^2+2acX+c^2)-M^2(aX+c)\right)\left(M(Y^2)-M^2(Y)\right)}= \\ &=\frac{\left(aM(XY)+cM(Y)-(aM(X)+c)M(Y)\right)^2}{\left((a^2M(X^2)+2acM(X)+c^2)-(aM(X)+c)^2\right)\left(M(Y^2)-M^2(Y)\right)}= \\ &=\frac{\left(aM(XY)-aM(X)M(Y)\right)^2}{\left(a^2M(X^2)-a^2M^2(X)\right)\left(M(Y^2)-M^2(Y)\right)}= \\ &=\frac{\left(M(XY)-M(X)M(Y)\right)^2}{\left(M(X^2)-M^2(X)\right)\left(M(Y^2)-M^2(Y)\right)} \end{aligned} \tag{44}$$

Relation (30) prove that the values of a and c does not affect the correlation between measured Y and estimated Y, Ŷ.

About standard errors for regression parameters calculation

Standard error of the estimate (SEE) is used in association with t-test to see if a significant linear correlation exists. The SEE is related to r^2 through:

$$\text{SEE} = \sqrt{\frac{1-r^2}{n-k}}\text{, k = the number of parameters in regression model} \qquad (45)$$

As we seen (Eq. 44), r^2 does not depend on parameters values, so, also SEE does not depend on parameters values.

For two parameters type linear regression, standard errors for parameters are:

$$\text{SE(a)} = \frac{\text{SEE}}{\sqrt{N}}\sqrt{\frac{1}{M(X^2)-M^2(X)}}\text{, SE(c) = }\text{SE(a)}\sqrt{M(X^2)}\text{, for Y=aX+c} \qquad (46)$$

For one type linear regression, standard errors for parameters are:

$$\text{SE(a)} = \frac{\text{SEE}}{\sqrt{N}}\sqrt{\frac{1}{M(X^2)}}\text{, for Y=aX} \qquad (47)$$

Gauss-Markov theorem implications for linear regression

The Gauss-Markov theorem states that parameters that are obtained from minimizing the sum of the squared errors are Best Linear Unbiased Estimate (called BLUE). Of course, this conclusion comes in some assumptions. If the errors are independent and identically normally distributed, it is the maximum likelihood estimator. Loosely put, the maximum likelihood estimate is the value of parameters that maximizes the probability of the data that was observed.

The Gauss-Markov theorem shows that the least squares estimate is a good choice, but if the errors are correlated or have unequal variance, there will be better estimators. Even if the errors behave but are non- normal then non-linear or biased estimates may work better in some sense. So this theorem does not tell one to use least squares all the time, it just strongly suggests it unless there is some strong reason to do otherwise. Situations where estimators other than ordinary least squares should be considered are:

- When the errors are correlated or have unequal variance, generalized least squares should be used.
- When the error distribution is long-tailed, then robust estimates might be used. Robust estimates are typically not linear in *y*.

- When the predictors are highly correlated (collinear), then biased estimators such as ridge regression might be preferable.

We have described linear models. Parameters (a, b, and c) may be estimated using least squares. If we further assume that errors of estimation are normally distributed then we can test any linear hypothesis about parameters, construct confidence regions for parameters (from standard errors), make predictions with confidence intervals.

What can go wrong? - many things, unfortunately; we try to categorize them below:

- Source and quality of the data - how the data was collected directly effects what conclusions we can draw. We may have a biased sample, such as a sample of convenience, from the population of interest. This makes it very difficult to extrapolate from what we see in the sample to general statements about the population. Important predictors may not have been observed. This means that our predictions may be poor or we may misinterpret the relationship between the predictors and the response. Observational data make causal conclusions problematic - lack of orthogonality makes disentangling effects difficult; missing predictors add to this problem. The range and qualitative nature of the data may limit effective predictions. It is unsafe to extrapolate too much. Carcinogen trials may apply large doses to mice. What do the results say about small doses applied to humans? Much of the evidence for harm from substances such as asbestos and radon comes from people exposed to much larger amounts than that encountered in a normal life. It's clear that workers in older asbestos manufacturing plants and uranium miners suffered from their respective exposures to these substances, but what does that say about the danger to you or me?
- We hope that errors are normally distributed; but errors may be heterogeneous (unequal variance), may be correlated, and/or may not be normally distributed. The last defect is less serious than the first two because even if the errors are not normal, the parameters will tend to normality due to the power of the central limit theorem [A]. With larger datasets, normality of the data is not much of a problem.

[A] from Wikipedia [http://en.wikipedia.org/wiki/Central_limit_theorem] - A central limit theorem is any of a set of weak-convergence results in probability theory. They all express the fact that any sum of many independent identically-distributed random variables will tend to be distributed according to a particular *attractor distribution*. The most important and famous result is called The Central Limit Theorem which states that if the sum of the variables has a finite variance, then it will be approximately normally distributed.

- The structural part of y = aX+c model may be incorrect. The model we use may come from different sources:
    - Physical theory may suggest a model; for example, Hooke's law says that the extension of a spring is proportional to the weight attached. Models like these usually arise in the physical sciences and engineering.
    - Experience with past data; similar data used in the past was modeled in a particular way. It's natural to see if the same model will work the current data. Models like these usually arise in the social sciences.
    - No prior idea - the model comes from an exploration of the data itself.

Confidence in the conclusions from a model declines as we progress through these. Models that derive directly from physical theory are relatively uncommon so that usually the linear model can only be regarded as an approximation to a reality, which is very complex. Most statistical theory rests on the assumption that the model is correct. In practice, the best one can hope for is that the model is a fair representation of reality. A model can be no more than a good portrait [2].

Rescaling of X and Y and Ridge regression

When we want to rescale the X and Y values? - When we want to make comparisons between predictors; predictors of similar magnitude are easier to compare; a change of units might aid interpretability; numerical stability is enhanced when all the predictors are on a similar scale.

Rescaling X and Y leaves the t, F tests and r^2 unchanged, and obtained new parameters are linear in rescaling. We already prove this for r^2.

Ridge regression makes the assumption that the regression coefficients (after normalization) are not likely to be very large.

Let us go back to our model (2) rewritten in term of estimator and to be estimate. Then our *estimator* Ê and our *to be estimated* E are:

$$\hat{E}(X,Y) = aX + bY + c,\ E(X,Y) = 0,\ \hat{E} \text{ estimator of } E \tag{48}$$

The sum function S from (3) became:

$$S = \sum(E(X,Y)-\hat{E}(X,Y))^2 = \sum(ax_i+by_i+c)^2 \tag{49}$$

The sum S can be averaged, when are called means-quared-error, MSE:

$$\text{MSE} = \text{M}((\text{E}(X,Y)-\hat{\text{E}}(X,Y))^2) \tag{50}$$

Let us assign another value to MSE:

$$\text{MSE} = (\text{E}(X,Y)-\text{M}(\hat{\text{E}}(X,Y)))^2 + \text{M}((\hat{\text{E}}(X,Y)-\text{M}(\hat{\text{E}}(X,Y)))^2) \tag{51}$$

The formulas (50) and (51) are equivalent only if E(X,Y) is assumed to be constant (independent of X and Y).

In formula (51) two interesting terms appears [3]:

$$\text{bias} = (\text{E}(X,Y)-\text{M}(\hat{\text{E}}(X,Y)))^2,\ \text{variance} = \text{M}((\hat{\text{E}}(X,Y)-\text{M}(\hat{\text{E}}(X,Y)))^2) \tag{52}$$

Note that occurs (53) and then model is unbiased.

$$\text{if E}(X,Y) = \text{M}(\hat{\text{E}}(X,Y)) \text{ for all } (X,Y) \text{ pairs} \Leftrightarrow \text{bias} = 0 \tag{53}$$

So, in the classifying of (53) our linear models can be splitted into biased (such as (28), (29), (30), (8.0), and (8.3)) and unbiased (such as (24), (8.1), (8.2), (37), and (38-39)).

Let us rewrite (52) using (49):

$$\text{bias} = (a\text{M}(X)+b\text{M}(Y)+c)^2,\ \text{variance} = \text{M}((a(X-\text{M}(X))+b(Y-\text{M}(Y)))^2) \tag{54}$$

So, the bias occurs when $(\text{M}(X),\text{M}(Y)) \not\in aX+bY+c = 0$. However, an unbiased model may still have a large mean-squared-error if $\hat{\text{E}}(X,Y)$ it has a large variance. This will be the case if $\hat{\text{E}}(X,Y)$ is highly sensitive to the peculiarities (such as noise and the choice of sample points) of each particular training set and it is this sensitivity which causes regression problems to be ill-posed in the Tikhonov [4] sense. Often, however, the variance can be significantly reduced by deliberately introducing a small amount of bias so that the net effect is a reduction in mean-squared-error. This is the job of Ridge regression [5], a method for solving badly conditioned linear regression problems.

Bad conditioning means numerical difficulties in performing the matrix inverse necessary to obtain the variance matrix. It is also a symptom of an ill-posed regression problem in Tikhonov's sense and Hoerl & Kennard's method was in fact a crude form of regularization, known now as zero-order regularization [6].

Introducing bias is equivalent to restricting the range of functions for which a model can account. Typically, this is achieved by removing degrees of freedom. Examples would be lowering the order of a polynomial or reducing the number of weights in a neural network. Ridge regression does not explicitly remove degrees of freedom but instead reduces the effective number of parameters. The resulting loss of flexibility makes the model less sensitive. A convenient, if somewhat arbitrary, method of restricting the flexibility of linear models is to augment the sum-squared-error with a term, which penalizes large weights,

$$\text{MSER} = \text{M}((\text{E}(\text{X},\text{Y})-\hat{\text{E}}(\text{X},\text{Y}))^2) + \rho^2(\text{a}^2(\text{M}(\text{X}^2)-\text{M}^2(\text{X}))+\text{b}^2(\text{M}(\text{Y}^2)-\text{M}^2(\text{Y}))) \quad (55)$$

This is ridge regression (weight decay) and the regularization parameter ρ^2 controls the balance between fitting the data and avoiding the penalty. A small value for means the data can be fit tightly without causing a large penalty; a large value for means a tight fit has to be sacrificed if it requires large weights. The bias introduced favors solutions involving small weights and the effect are to smooth the output function since large weights are usually required to produce a highly variable (rough) output function.

The use of ridge regression can be motivated in two ways. Suppose we take a Bayesian point of view and put a prior (multivariate normal) distribution on b that expresses the belief that smaller values of a and b are more likely than larger ones. Large values of ρ^2 correspond to a belief that the b are really quite small whereas smaller values of ρ^2 correspond to a more relaxed belief about a and b. Another way of looking at is to suppose we place to some upper bound on ($a^2+b^2+c^2$) and then compute the least squares estimate to this restriction. Use of Lagrange multipliers leads to ridge regression. The choice of ρ^2 corresponds to the choice of upper bound in this formulation. ρ^2 may be chosen by automatic methods but it is probably safest to plot the values of parameters as a function of ρ^2. You should pick the smallest value of ρ^2 that produces stable estimates of parameters.

Discussion

The use of PM(·,·,·), eq. (30) - Hölder's mean - it opens an interesting discussion. We already had seen that:

- if $S \leftarrow (Y-(aX+c))^2$ then:
    - if c = **0** then:
        - $a = \dfrac{M(XY)}{M(X^2)}$ (or $a = \dfrac{M(Y)}{M(X)}$ from M(Y) = aM(X))
    - else:
        - $a = \dfrac{M(XY)-M(X)M(Y)}{M(X^2)-M^2(X)}$, $c = \dfrac{M(X^2)M(Y)-M(X)M(XY)}{M(X^2)-M^2(X)}$

- if S ← (X-(Y-c)/a)^2 then:
  - if c = **0** then:
    - $a = \dfrac{M(Y^2)}{M(XY)}$ (or $a = \dfrac{M(Y)}{M(X)}$ from M(Y) = aM(X))
  - else:
    - $a = \dfrac{M(Y^2)-M^2(Y)}{M(XY)-M(X)M(Y)}$, $c = \dfrac{M(Y)M(XY)-M(X)M(Y^2)}{M(XY)-M(X)M(Y)}$

If we put back our formulas to the geometrical interpretation, following result is obtained (Figure 2):

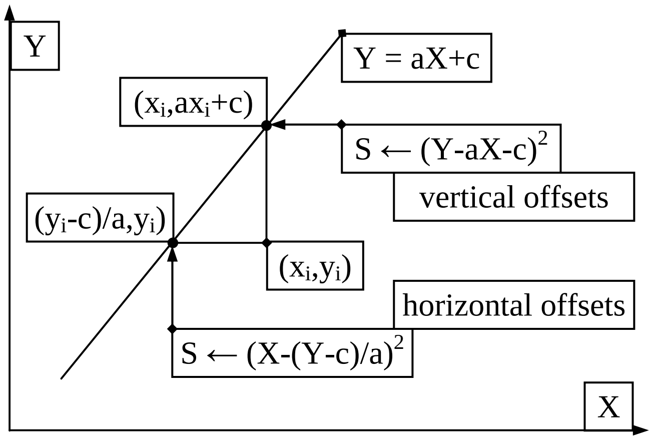

Figure 2. Penality function S vs. vertical and horizontal offsets

In addition, we had seen that (eq. 33):

- $\min(\bullet,\bullet) = \lim\limits_{p\to-\infty} PM(\bullet,\bullet,p)$
- $\max(\bullet,\bullet) = \lim\limits_{p\to+\infty} PM(\bullet,\bullet,p)$

So, if we put our formulas for, let us see, slope obtained from this two different aproaches in previous formulas, it result that:

$$\begin{aligned}\min(a_{S\leftarrow(Y-aX-c)^2}, a_{S\leftarrow(X-(Y-c)/a)^2}) &= \lim_{p\to-\infty} PM(a_{S\leftarrow(Y-aX-c)^2}, a_{S\leftarrow(X-(Y-c)/a)^2}, p)\\ \max(a_{S\leftarrow(Y-aX-c)^2}, a_{S\leftarrow(X-(Y-c)/a)^2}) &= \lim_{p\to+\infty} PM(a_{S\leftarrow(Y-aX-c)^2}, a_{S\leftarrow(X-(Y-c)/a)^2}, p)\end{aligned} \tag{56}$$

In fact, through equation (56) we construct a function (PM) which sweeps the entire right angle (figure 3).

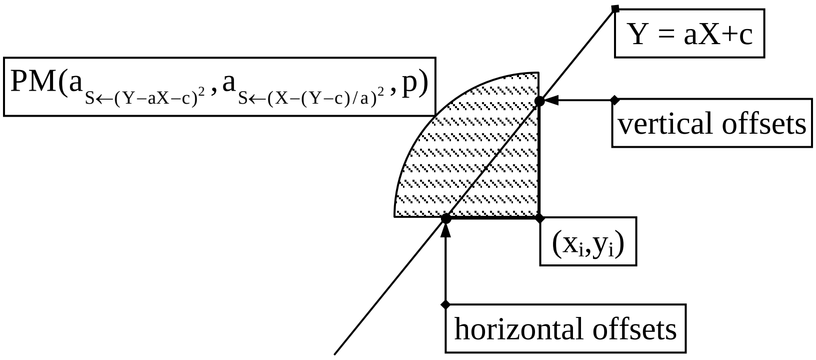
Y = aX+c
$\mathrm{PM}(a_{S\leftarrow(Y-aX-c)^2}, a_{S\leftarrow(X-(Y-c)/a)^2}, p)$
vertical offsets
(x_i,y_i)
horizontal offsets

Figure 3. Hölder mean, acting for linear regression

As consequence, all obtained formulas for slope (and for intercept, when is not set null) can be obtained from a Hölder mean:

$$\forall S\ \exists p \in \widehat{\mathbb{R}}^{*} \text{ such that}: |a_S| = \mathrm{PM}\left(\left|a_{(Y-ax-c)^2}\right|, \left|a_{(X-(Y-c)/a)^2}\right|, p\right) \tag{57}$$

where the obtained formula (57) was completed with negative slope cases. Of course, when slope is negative, then are choused the negative solution of (57).

Finally, note that equation (57) has a solution in $\widehat{\mathbb{R}}^{*}$ and this is unique if and only if $|a_S|$ are in between:

$$\begin{gathered}\exists! p \in \widehat{\mathbb{R}}^{*} : 2|a_S| = \left(\left|a_{(Y-ax-c)^2}\right|^p + \left|a_{(X-(Y-c)/a)^2}\right|^p\right)^{1/p} \\ \Leftrightarrow \\ \left|a_{(Y-ax-c)^2}\right| \le |a_S| \le \left|a_{(X-(Y-c)/a)^2}\right|\end{gathered} \tag{60}$$

and as we already seen (eq. 33), for $a_S{}^2 = a_{(Y-ax-c)^2} a_{(X-(Y-c)/a)^2}$ admits also p → 0 as limit solution.

An interesting formula results also as consequence of (8.1), (8.2), (30), (33) and construction from Figure 2: calculation of slope using GM and intercept using M(Y)=aM(X)+c, where sign of a and c respectively, are given from quadrant of scatter plot:

$$a = \pm\sqrt{\frac{M(Y^2) - M^2(Y)}{M(X^2) - M^2(X)}},\ c = M(Y) \mp M(X)\sqrt{\frac{M(Y^2) - M^2(Y)}{M(X^2) - M^2(X)}} \tag{61}$$

References

[1] Kvalseth T. O., *Cautionary Note about R^2*, The American Statistician, 1985, 39(4), p. 279-285.

[2] Faraway J. J., *Practical Regression and Anova using R*, July 2002, Copyright © 1999, 2000, 2002 Julian J. Faraway, http://www.stat.lsa.umich.edu/˜faraway/book.

[3] Geman S., Bienenstock E., Doursat R., *Neural networks and the bias/variance dilemma*, Neural Computation, 1992, 4(1), p. 1-58.

[4] Tikhonov A. N., Arsenin V. Y., Solutions of Ill-Posed Problems, Winston, Washington, 1977.

[5] Hoerl A. E., Kennard R. W., *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 1970, 12(3), p. 55-67.

[6] Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., Numerical Recipes in C. Cambridge University Press, Cambridge, UK, second edition, 1992.