HOW TO TEACH THE SOLUTION OF THE PROBLEM OF LINEAR REGRESSION, DIRECT APPROACH IN A CLOSED FORM

Helmut Maier, Berlin School of Economics, Germany

The paper refers to the situation teaching basic statistics to students with little background in mathematics, especially no knowledge in differential calculus. Goal of the paper is to show that knowledge of basic linear algebra is sufficient to explain the problem and the solution of linear regression, and this in a closed form. This approach avoids black box teaching where students learn how to get results and don't learn why, and it replaces the so-called ***Gaussian normal equations*** *which use partial derivatives. Starting point is the so-called* ***residual variance*** *of the dependent variable. The paper presents a decomposition of this variance in three non-negative terms, hence the solution of the problem of linear regression is obvious. Besides, this decomposition may be used to explain properties of the correlation coefficient and the coefficient of determination.*

OPENING REMARKS

The problem of linear regression and its solution by means of the differential calculus leading to the so-called Gaussian normal equations are well known and classical components of basic courses of statistics in education. Indeed, there seems to be no need for a further contribution. Just this situation might be the reason why a fundamental formula, presented in this paper, is not written in formula collectives, and is not being used in descriptions of linear regression analysis, in opinion of the author. Viewing this formula, a decomposition of the residual variance of the dependent variable, the solution of the problem of linear regression is obvious, indeed. There is no need of knowledge of differential calculus. The formula itself, and her diversion use basic knowledge of linear algebra. Thus, the same methods are used for the formulation and the solution of the problem of linear regression, and not different ones as done by Gauss. A need for such an approach arised from education at Berlin School of Economics in the late 80th years. Because of changes in the education pattern, many students did not have enough knowledge of the differential calculus during current courses of statistics (absence of partial derivatives). The formula was developed in 1988. Her use saves lecture time, this is empirical result of her application in the 90th years. Viewing this educational aspect, the explanations in this paper are separated in the simple case of one dependent variable and one independent one, and the general case of one dependent variable and a set of independent variables. The presentation excludes the proof of the formula of decomposition, this proof is given with Maier (1998).

THE SIMPLE CASE OF LINEAR REGRESSION OF A VARIABLE Y AGAINST A VARIABLE X

We *denote* by X and Y two stochastic variables, we assume n given observations of X and Y, denoted by x_i and y_i with i = 1,2,..,n , and we assume that Y depends on X. The symbols m and b denote limited slope and intercept of any straight line $y(x) = m x + b$ in a plane defined by rectangular x and y coordinates. The cursive symbols x and y denote the means of the observations x_i and y_i, that is $x = (x_1 +...+ x_n)/n$ and $y = (y_1 +...+ y_n)/n$, the symbols s_x^2 and s_y^2 denote the estimations of the variances of X and Y, with $s_x^2 = ((x_1-x)^2 +...+ (x_n-x)^2)/n$ and $s_y^2 = ((y_1-y)^2 +...+ (y_n-y)^2)/n$, their positive roots s_x and s_y estimations of the standard deviations of X and Y. The symbol s_{xy} denotes the estimation of the covariance between X and Y, that is $s_{xy} = ((x_1-x) (y_1-y) +...+ (x_n-x)(y_n-y))/n$, and the symbols r and r^2 denote estimations of the correlation coefficient and the coefficient of determination between X and Y, the latter is the square of the first, that is $r = s_{xy} /(s_x s_y)$. Furthermore, we denote by s_{Ly}^2 the estimation of the residual variance of Y referring to the assumption that Y depends on X described by y(x) = m x + b, that is $s_{Ly}^2 = ((y_1-y(x_1))^2 +... + (y_n-y(x_n))^2)/n$. The value s_{Ly}^2 is equal to the mean of the squares of the distances of the observation points (x_i , y_i) to the corresponding points of any straight line y(x) measured parallel to the y-axis.

The *decomposition of the residual variance*, s_{Ly}^2, is

$$s_{Ly}^2 = (s_x m - r s_y)^2 + (y - m x - b)^2 + s_y^2 (1 - r^2) ,$$

that is s_{Ly}^2 is a function of the data parameters x, y, s_x, s_y, r, and of the parameters m and b of the straight line y(x) = m x + b . This function is a sum of three non-negative parts. For the first two parts, this is obvious because these terms are squares. For the third part, this is obvious if we make use of the knowledge that $(1 - r^2)$ is non-negative. Given fixed data sets of X and Y, the data parameters are constants, and the parameters of the straight line are variables. This decomposition holds for any limited values of m and b.

The *problem of linear regression* of Y against X is: Find a straight line y(x) = m x + b with minimum residual variance s_{Ly}^2 . To realize such an optimal *solution*, we choose the values of slope m and intercept b so that the first two non-negative parts in the decomposition are zero. Obviously, this choice is a necessary as well as a sufficient condition to get the absolute minimum, and it leads to a canonical form of a linear equation system for the unknown values of the variables slope m and intercept b,

$$s_x m - r s_y = 0 \qquad \text{and} \qquad y - m x - b = 0 ,$$

hence to the solutions $m = r s_y / s_x$ and $b = y - x r s_y / s_x$, and to the straight line of regression $y(x) = y + (x - x) r s_y / s_x$, which describes the assumed dependence between X and Y in quantitative terms. The *graphical interpretation* of this solution process is: Turn a given straight line in this way that its slope is equal $(r s_y / s_x)$, and adjust it in this way that the point (x , y) representing the means of the data sets, is on this straight line. Then you will get the straight line with minimum residual variance. This solution uses elementary methods of linear algebra. Besides, as the third part of the partition of the residual variance s_{Ly}^2 is independent from m and b, this part, $s_y^2 (1 - r^2)$, is the minimum of the residual variance.

Characteristics of the correlation coefficient may be derived from the formula of decompo-sition. In case of minimum of the residual variance, the equations $s_{Ly}^2 = s_y^2 (1- r^2)$ and $m = r s_y / s_x$ hold. As s_y and s_x are positive, the slope of the regression line is positive when r is positive, and the slope is negative when r is negative. As s_{Ly}^2 is non-negative and s_y^2 is positive, we conclude $0 \leq s_{Ly}^2 = s_y^2 (1 - r^2)$, hence $0 \leq 1 - r^2$, and $-1 \leq r \leq 1$. The case $r^2 = 1$ is equivalent to $s_{Ly}^2 = 0$ and equivalent to $y_i - y(x_i) = 0$ for $i = 1,...,n$ by definition of s_{Ly}^2. The latter means that all points (x_i , y_i) lie on the regression line and satisfy the dependence between X and Y. The case $r = 0$ is equivalent to $s_{Ly}^2 = s_y^2$ and means that the minimum residual variance is as big as the variance of Y, hence we conclude that there is no dependence between X and Y.

The *interpretation of the coefficient of determination* may be derived from the formula of decomposition. In case of minimum of the residual variance, the equation $s_{Ly}^2 = s_y^2 (1-r^2)$ holds. Hence we derive $r^2 = (s_y^2 - s_{Ly}^2)/ s_y^2$. As s_{Ly}^2 is the residual variance of Y (after linear regression analysis), and s_y^2 is the (original) variance of Y, we interpret the term $(s_y^2 - s_{Ly}^2)$ as this part of the variance of Y which can be explained by the assumption that Y depends on X in the computed way. Hence the ratio $(s_y^2 - s_{Ly}^2) / s_y^2$ is the percentage of the variance of Y which can be explained by the assumption that Y depends on X in this way, and thus by definition this ratio is the coefficient of determination.

THE GENERAL CASE OF LINEAR REGRESSION OF A VARIABLE Y AGAINST

A SET OF VARIABLES X_J

We *denote* by Y and X_j , $j = 1,...,p$, stochastic variables, we assume n given observations of Y and X_j , $j = 1,...,p$, denoted by y_i and x_{ji} with $i = 1,..,n$ and $j = 1,...,p$, and we assume that Y depends on the X_j . The symbols $m_1,...,m_p$ denote the different limited slopes of a p-dimensional hyperplane in the (p+1)-dimensional space of the variables Y and X_j , and the symbol b denotes its limited intercept. Using the coordinates of this space, $x_1,...,x_p$, and y, the equation of this hyperplane is $y(x_1,..,x_p) = m_1 x_1 +...+ m_p x_p + b$. With cursive symbols x_j for $j = 1,..,p$, and *y*, we denote the means of the observations of the variables X_j and Y, in formulas $\mathit{x}_j = (x_{j1} +... +x_{jn})/n$ and $\mathit{y} = (y_1+...+y_n)/n$. The symbols s_j^2 , for $j = 1,...,p$, and s_y^2 denote the estimations of the variances of the variables X_j and Y, their positive roots s_j and s_y denote the standard deviations of X_j and Y, that is $s_j^2 = ((x_{j1}-\mathit{x}_j)^2 +...+ (x_{jn}-\mathit{x}_j)^2)/n$ for $j = 1,...,p$ and $s_y^2 = ((y_1-\mathit{y})2 +...+ (y_n-\mathit{y})^2)/n$. The symbol k_{qj} denotes for $q = 1,..,p$ and $j = 1,...,p$ the estimation of the covariance between X_q and X_j , that is $k_{qj} = ((x_{q1}-\mathit{x}_q)(x_{j1}-\mathit{x}_j) +... + (x_{qn}-\mathit{x}_q)(x_{jn}-\mathit{x}_j))/n$, and the symbol r_{qj} denotes the correlation coefficient between the variables X_q and X_j , that is $r_{qj} = k_{qj} / (s_q s_j)$. For $q = 1,...,p$, the symbol k_{qy} denotes the estimation of the covariance between X_q and Y, that is $k_{qy} = ((x_{q1}-\mathit{x}_q)(y_1-\mathit{y}) +...+ (x_{qn}-\mathit{x}_q) (y_n-\mathit{y})) /n$, and the symbol r_{qy} denotes the correlation coefficient between X_q and Y, that is $r_{qy} = k_{qy} / (s_q s_y)$. We denote by s_{Hy}^2 the estimation of the residual variance of Y referring to any p-dimensional hyperplane, we measure this variance in the (p+1)-dimensional space due to Gauss as mean of the squared distances of the n observation points $(x_{1i} ,..., x_{pi} , y_i)$, for $i = 1,...,n$, and parallel to the y-axis to the *upper* or *lower* situated points of this hyper-plane. The residual variance is $s_{Hy}^2 = ((y_1 - y(x_{11} ,..., x_{p1}))^2 +...+ (y_n - y(x_{1n} ,..., x_{pn}))^2)/n$. We use *matrix and vector notation*. We write fat letters for vectors and matrices, big letters for matrices, and we denote by brackets the referring sets of values of matrices and vectors. We write the symbol * for the transposed matrix, we assume vectors to be column vectors, and we remind that the transposed vector of a row vector is a column vector. We denote the scalar product of two vectors with a point $\cdot$ in between. Thus we get as notations, for the vector with the slopes of the p-dimensional hyperplane: $\mathbf{m} = (m_j) = (m_1 ,..., m_p)^*$, for the vector variable with the coordinates $\mathbf{x} = (x_j) = (x_1 ,..., x_p)^*$, for the equation of any hyperplane $y(\mathbf{x}) = \mathbf{m} \cdot \mathbf{x} + b$, for the vector with the means of the variables X_j : Cursive $\boldsymbol{\mathit{x}} = (\mathit{x}_j) = (\mathit{x}_1 ,..., \mathit{x}_p)^*$, for the matrix of the p^2 correlation coefficients between the p variables $X_1,...,X_p$: $\mathbf{R} = (r_{qj})$, for the matrix of the p^2 covariances between the variables $X_1,...,X_p$: $\mathbf{K} = (k_{qj}) = (r_{qj}\, s_q\, s_j)$, for the vector with the covariances between Y and the variables $X_1,...,X_p$: $\mathbf{k}_y = (k_{qy}) = (k_{1y},...,k_{py})^*$, and for the vector with the correlation coefficients between Y and $X_1,...,X_p$: $\mathbf{r}_y = (r_{qy}) = (r_{1y},...,r_{py})^*$. We note, **R** and **K** are symmetric matrices. We introduce matrix **A** , defined by **A*A = K**, as a partitioning of matrix **K** of the covariances**,** and vector **ß** as a solution of $\mathbf{A}^*\mathbf{ß} = \mathbf{k}_y$. We assume, **R** and **K** are regular, then the inverse matrices $\mathbf{R}^{-1}$ and $\mathbf{K}^{-1}$ as well as **A** , $\mathbf{A}^{-1}$, and vector **ß** exist. The *decomposition of the residual variance* of Y, $s_{Hy}{}^2$, referring to any p-dimensional hyperplane $y(\mathbf{x}) = \mathbf{m} \cdot \mathbf{x} + b$ with limited slopes and intercept is

$$s_{Hy}{}^2 = (\mathbf{Am} - \mathbf{ß}) \cdot (\mathbf{Am} - \mathbf{ß}) + (\, y - \mathbf{m} \cdot \boldsymbol{x} - b\,)^2 + (\, s_y{}^2 - \mathbf{k}_y \cdot \mathbf{K}^{-1}\, \mathbf{k}_y\,)\,,$$

that is $s_{Hy}{}^2$ is a function of the data parameters $\boldsymbol{x}$, y , $s_y{}^2$, $\mathbf{k}_y$, **K** ($\boldsymbol{\beta}$ depends on **A** and $\mathbf{k}_y$, and **A** depends on **K**), and of the parameters **m** and b of the p-dimensional hyperplane. This function is a sum of three non-negative parts. For the first two parts, this is obvious because these terms are squares. For the third part it turns obvious later, we note $\mathbf{k}_y \cdot \mathbf{K}^{-1}\, \mathbf{k}_y = s_y{}^2\, (\mathbf{r}_y \cdot \mathbf{R}^{-1}\, \mathbf{r}_y)$. Given fixed data sets of the variables X_j , j = 1,...,p , and Y, these data parameters are constants, and the parameters of the hyperplane are variables. The first two parts depend on the parameters of the hyperplane, and the third does not.

The *problem of linear regression* is: Find a p-dimensional hyperplane $y(\mathbf{x}) = \mathbf{m} \cdot \mathbf{x} + b$ in a (p+1)-dimensional space with minimum residual variance $s_{Hy}{}^2$. To realize such an optimal *solution*, we choose the slope vector **m** and the intercept b so that the first two parts of the decomposition of the residual variance are zero. Obviously, this choice is a necessary as well as a sufficient condition to get the absolute minimum of the residual variance $s_{Hy}{}^2$. Using symbol **0** for the zero vector with p zeros as components, **m** and b must satisfy the equations

$$\mathbf{Am} - \mathbf{ß} = \mathbf{0} \qquad \text{and} \qquad y - \mathbf{m} \cdot \boldsymbol{x} - b = 0\ .$$

This is a linear equation system with (p+1) equations for the (p+1) unknowns $m_1,...,m_p$ and b. Firstly, we estimate the slope vector **m**,

$$\mathbf{m} = \mathbf{A}^{-1}\, \mathbf{ß} = \mathbf{A}^{-1}\, (\, (\mathbf{A}^*)^{-1}\, \mathbf{k}_y\,) = (\mathbf{A}^{-1}\, (\mathbf{A}^*)^{-1}\,)\, \mathbf{k}_y = (\mathbf{A}^*\mathbf{A})^{-1}\, \mathbf{k}_y = \mathbf{K}^{-1}\, \mathbf{k}_y\ ,$$

and we note that the solution $\mathbf{m} = \mathbf{K}^{-1}\, \mathbf{k}_y$ does not depend on the chosen partitioning **A*A** of **K** , and hence is unique. Secondly and using this result, we estimate the intercept b,

$$b = y - \mathbf{m} \cdot \boldsymbol{x} = y - \mathbf{K}^{-1}\, \mathbf{k}_y \cdot \boldsymbol{x}$$

which is unique as well. Hence we get the equation for the hyperplane with minimum residual variance s_{Hy}^2 :

$$\mathrm{y}(\mathbf{x}) \;=\; \mathbf{m} \cdot \mathbf{x} \;+\; \mathrm{b} \;=\; \mathbf{K}^{-1}\,\mathbf{k}_y \cdot \mathbf{x} \;+\; y - \mathbf{K}^{-1}\,\mathbf{k}_y \cdot \boldsymbol{x} \;=\; \mathbf{K}^{-1}\,\mathbf{k}_y \cdot (\mathbf{x} - \boldsymbol{x}) \;+\; y\,.$$

The *geometrical interpretation* of this solution process is: Move a given hyperplane $\mathrm{y}(\mathbf{x})$ until the slopes fit $\mathbf{Am} - \mathbf{ß} = \mathbf{0}$, and adjust it in this way that the point $(\boldsymbol{x}, y)$ representing the means of the data sets is on this hyperplane.

With respect to this hyperplane, the *minimal residual variance* is $s_{Hy}^2 = s_y^2 - \mathbf{k}_y \cdot \mathbf{K}^{-1}\,\mathbf{k}_y$. Hence we conclude that this third part of the decomposition of s_{Hy}^2 is also non-negative. The residual variance is equal zero if all observations satisfy the equation of this hyperplane. Furthermore and because of the relation $k_{qy} = r_{qy}\, s_q\, s_y$ holds for $q = 1,...,p$, the scalar product $\mathbf{k}_y \cdot \mathbf{K}^{-1}\,\mathbf{k}_y$ includes s_y^2 as a factor and may be written in different terms using correlation coefficients, and may be used to derive characteristics of the correlation coefficients. This we note. Because of $\mathbf{k}_y \cdot \mathbf{K}^{-1}\,\mathbf{k}_y = s_y^2 - s_{Hy}^2$ we interpret $\mathbf{k}_y \cdot \mathbf{K}^{-1}\,\mathbf{k}_y$ as this part of the (original) variance of Y which can be explained by the assumption that Y depends on the $X_1,...,X_p$ in the estimated way. Hence $(\mathbf{k}_y \cdot \mathbf{K}^{-1}\,\mathbf{k}_y)\,/s_y^2 = (s_y^2 - s_{Hy}^2)\,/\,s_y^2$ is the percentage of the variance of Y wich can be explained by the assumption that Y depends on the variables $X_1,...,X_p$, that is $(\mathbf{k}_y \cdot \mathbf{K}^{-1}\,\mathbf{k}_y)\,/s_y^2$ represents the *coefficient of determination*. We conclude $0 \leq (\mathbf{k}_y \cdot \mathbf{K}^{-1}\,\mathbf{k}_y)\,/s_y^2 \leq 1$, and $(\mathbf{k}_y \cdot \mathbf{K}^{-1}\,\mathbf{k}_y)\,/s_y^2 = \mathbf{r}_y \cdot \mathbf{R}^{-1}\,\mathbf{r}_y$.

CLOSING REMARKS

The use of the decomposition formula of the residual variance of the dependent variable leads to a linear equation system for the parameters of the p-dimensional hyperplane which is equivalent to the so-called Gaussian normal equations. This we note. In this direct approach, we get the solution in a closed form in addition. This means that in case of a single problem of linear regression there is no need to solve a linear equation system. We compute the covariance matrix $\mathbf{K}$ of the independent variables X_j, the inverse matrix $\mathbf{K}^{-1}$, the covariance vector $\mathbf{k}_y$, the vector $\boldsymbol{x}$ with the means of the independent variables X_j , the mean y of the dependent variable Y, and hence the hyperplane $\mathrm{y}(\mathbf{x}) = \mathbf{K}^{-1}\,\mathbf{k}_y \cdot (\mathbf{x} - \boldsymbol{x}) + y$. The different operations of linear algebra are only necessary to understand why this hyper-plane is the solution of the problem of linear regression. Furthermore, the results remain valid when we replace the biased estimations of the variances and covariances (with factor 1/n) by the unbiased estimations (with factor 1/(n-1)), see Maier (1998). In this sense, the analysis of the residual variance is a strong and suitable instrument to provide deepened knowledge of the problem of linear regression and its solution.

BIBLIOGRAPHY

Maier, H. (1998). A direct solution of the problem of linear regression by analysis of variance, *Student*, *2(3)*,1-12.