Regression *

Oscar García

Regression methods are fundamental in Forest Mensuration. For a more concise and general presentation, we shall first review some matrix concepts.

1 Matrices

An order $n \times m$ *matrix* is simply a table of numbers with n rows and m columns:

$$\mathrm{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = [a_{ij}] \ .$$

The a_{ij} are the matrix *elements*. Instead of the square brackets, round parenthesis or double vertical lines are also used: $\|a_{ij}\|$.

A *vector* is a list of numbers. In matrix algebra they are taken as one-row matrices (row vector) or one-column matrices (column vector). Unless stated otherwise, we shall assume columns. They are usually represented by lower-case letters, often underlined or in bold-face:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [x_i] \ .$$

The *transpose* matrix is the matrix obtained exchanging rows and columns. The transpose of A is denoted as A' or A^T.

The *sum* of two matrices is the matrix of sums of their elements:

$$\mathrm{A} + \mathrm{B} = [a_{ij}] + [b_{ij}] = [a_{ij} + b_{ij}] \ .$$

*Translated from Appendix B in *Apuntes de Mensura Forestal – Estática*, Universidad Austral de Chile, Facultad de Ciencias Forestales, 1995

Obviously, A and B must be of the same order.

A single number, to distinguish it from a vector o matrix, is called a *scalar*. The product of a scalar and a matrix is obtained by multiplying the scalar and each of the elements of the matrix:

$$k\mathrm{A} = k[a_{ij}] = [ka_{ij}] \ .$$

From this, the subtraction or difference of matrices is

$$\mathrm{A} - \mathrm{B} = \mathrm{A} + (-1)\mathrm{B} = [a_{ij} - b_{ij}] \ .$$

The *matrix product* $\mathrm{AB} = \mathrm{C}$ is obtained in the following way:

$$[c_{ij}] = [\sum_{k} a_{ik} b_{kj}] \ .$$

That is, the element ij in the product is the sum of products of the elements from row i of A and those from column j of B. Clearly, for the product to be defined the number of columns in the first matrix must equal the number of rows in the second one.

Defining the product in this way is useful, for example, in handling systems of linear equations. The system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots & \quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

can be written simply as

$$\mathrm{A}\mathbf{x} = \mathbf{b} \ .$$

A sum of squares is

$$e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} \ ,$$

where

$$\mathbf{e} = [e_1 e_2 \cdots e_n]' \ .$$

Even if two matrices have the proper dimensions for calculating the products AB and BA, in general the results are different (the matrix product is not commutative). Other than this, and that the operations are not always possible (certain relationships between numbers of rows and columns must

be satisfied), the sum, difference, and product of matrices behave as the corresponding operations on scalars. For instance,

$$\begin{aligned} \mathrm{A}(\mathrm{B}+\mathrm{C}) &= [a_{ij}][b_{ij}+c_{ij}] = [\sum_k a_{ik}(b_{kj}+c_{kj})] = [\sum_k a_{ik}b_{kj} + \sum_k a_{ik}c_{kj}] \\ &= \mathrm{AB}+\mathrm{AC}\ . \end{aligned}$$

Questions, exercises

1. Show that the sum is commutative, $\mathrm{A}+\mathrm{B} = \mathrm{B}+\mathrm{A}$, and associative, $(\mathrm{A}+\mathrm{B})+\mathrm{C} = \mathrm{A}+(\mathrm{B}+\mathrm{C})$.

2. Show that $(\mathrm{AB})' = \mathrm{B}'\mathrm{A}'$.

3. Compute AB and BA, where

$$\mathrm{A} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}\ ,\ \ \mathrm{B} = \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix}\ .$$

4. Compute $\mathbf{x}'\mathbf{y}$ and $\mathbf{y}'\mathbf{x}$, where

$$\mathbf{x} = \begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix}\ ,\ \ \mathbf{y} = \begin{bmatrix} 2 \\ 1 \\ -3 \end{bmatrix}\ .$$

   Note: Often a matrix with just one element is considered as a scalar.

5. Show that $p(\mathrm{A}+\mathrm{B}) = p\mathrm{A}+p\mathrm{B}$ y $(p+q)\mathrm{A} = p\mathrm{A}+q\mathrm{A}$.

6. Show that the product is associative: $(\mathrm{AB})\mathrm{C} = \mathrm{A}(\mathrm{BC})$.

The matrices with ones on the diagonal and zeroes elsewhere,

$$\mathrm{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}\ ,$$

are known as *identity*. They act as the number 1 among the scalars; multiplying any matrix and the identity (of the proper order) does not change it:

$$\mathrm{IA} = \mathrm{AI} = \mathrm{A}\ .$$

Now for an analogue to scalar division. In the same way as subtraction may be seen as summing a negative, $a - b = a + (-b)$, division may be seen as multiplication with a reciprocal: $a/b = a(1/b) = ab^{-1}$, where $b^{-1}b = 1$. With matrices, the analogue of a reciprocal is the *inverse*,

$$\mathrm{A}^{-1}\mathrm{A} = \mathrm{A}\mathrm{A}^{-1} = \mathrm{I}\,.$$

Note that for this to make sense, A must be square (same number of rows and columns). Even thus, not all square matrices have an inverse. Those that do not have one are called *singular*.

Using the inverse we could write the solution of the equation system $\mathrm{A}\mathbf{x} = \mathbf{b}$ given earlier:

$$\mathbf{x} = \mathrm{A}^{-1}\mathbf{b}\,.$$

For this solution to exist, A must be square ($m = n$, that is, the number of equations must equal the number of unknowns). In addition, for A not to be singular, the equations must be "linearly independent" (there must be no redundant equations). There are various methods for inverting matrices, one of the most common being Gaussian elimination. This may be also used to solve equation systems without computing the full inverse.

It is not difficult to verify the following properties:

$$(\mathrm{AB})^{-1} = \mathrm{B}^{-1}\mathrm{A}^{-1}$$

$$(\mathrm{A}')^{-1} = (A^{-1})'\,.$$

Finally, one can define vector and matrix derivatives. The derivative of a matrix with respect to a scalar, and the derivative of a scalar with respect to a matrix, are defined simply as the matrix of derivatives. It is then easy to verify results like these (A and $\mathbf{a}$ contain constants):

$$\frac{d\mathrm{A}\mathbf{x}}{dt} = \mathrm{A}\frac{d\mathbf{x}}{dt}$$

$$\frac{d\mathbf{a}'\mathbf{x}}{d\mathbf{x}} = \mathbf{a}$$

$$\frac{d\mathbf{x}'\mathbf{x}}{d\mathbf{x}} = 2\mathbf{x}$$

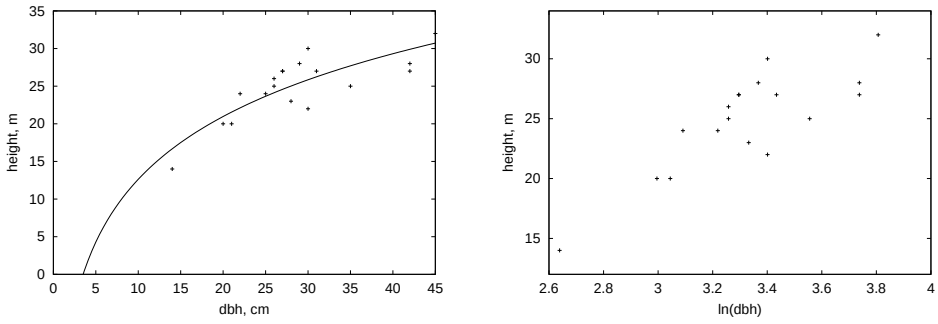
$$\frac{d\mathbf{x}'\mathrm{A}\mathbf{x}}{d\mathbf{x}} = (\mathrm{A} + \mathrm{A}')\mathbf{x}$$

etc. In general, the results are similar to those for scalars, taking into account the no commutativity of products.

2 The least-squares method

Many mensurational methods are based on relationships between a *dependent variable* and one or more *independent variables*. One is interested in describing the relationship between the variables, or in estimating or predicting the value of the dependent variable knowing the value of the predictors. For instance, the relationship between heights and diameters may be used to estimate the height of a tree knowing its dbh, which is more easily measured. Or estimate the volume knowing its dbh and height. Or predict the volume of a stand at a given age.

As an example, take a relationship between two variables. It is useful to make a *scatter diagram*, plotting the available observations with the predictor in the abcissa ("x-axis"), and the dependent variable in the ordinate ("y-axis"). The graph on the left shows observations of height and dbh in a stand of *Eucalyptus nitens* taken by the 1994 class.



A curve like the one shown may be used for estimating the heights of trees in the stand for which only the dbh is known. Clearly, knowing the dbh helps in estimating the height, that is, contributes to reduce the uncertainty about its value. The curve is a "model" that provides height values to be used in place of the unknown ones, or that can serve as a summary description of the observations. At any rate, it is convenient to have an equation for the curve to facilitate its use, and the curve should pass "close" to the observations.

In some instances there are theoretical reasons that suggest a specific kind of equation. In others, as in this example, the equation is purely empirical, chosen with convenience and data-fitting criteria. In general, there will be a class of equations or models $y = f(\mathbf{x}, \mathbf{b})$, where y is the dependent variable, $\mathbf{x}$ is a vector of independent variables, and $\mathbf{b}$ is a vector of parameters whose values will be determined for producing a good fit. With a two-dimensional $\mathbf{x}$ we obtain a surface instead of a curve, and for

higher dimensions a hypersurface. To choose the equation form one may use experience with similar problems, trial and error, graphs with transformations producing linear data trends, considerations about the form that the curve should take for the extremes, etc. In the example we have used $H = f(D, b_1, b_2) = b_1 + b_2 \ln D$, seeing in the right-hand-side scatter diagram that the relationship between H and $\ln D$ is roughly linear (note in passing that extrapolation to small diameters outside the range of the data eventually produces negative heights). It would be always possible to choose a curve that passes close to each one of the observations. Although in some sense this would describe perfectly the observed data, in general much less irregular curves, with a small number of parameters, will produce better estimates for future or unobserved values.

Once the form of the equation to be tried is decided, it is necessary to choose parameter values that result in a good fit. It can be assumed that, for a given D, the difference between the unknown H and $f(D, b_1, b_2)$ would tend to be smaller if these differences are small for the observed values. That is, $\mathbf{b}$ should be such that the absolute values of the *deviations*, *residuals* or "errors" $e_i = H_i - f(D_i, b_1, b_2)$ are small for all the observations (D_i, H_i). Obviously, if we try to reduce one e_i beyond some point the other e_i will increase, so that we need some criterion that takes into account the whole set of these. A possible criterion would be to minimize the sum of absolute values $\sum |e_i|$ ("L_1-norm regression"). Another possibility would be to minimize the largest error ($\min \max |e_i|$, the *minimax* criterion). The criterion most commonly used, because of mathematical convenience and of possessing in some instances certain statistical justifications that we will examine later, is that of *least-squares*, which consists of minimizing $\sum e_i^2$.

We have then a model $y = f(\mathbf{x}, \mathbf{b})$, n observations $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots n$, and we look for a $\mathbf{b}$ such that it minimizes

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - f(\mathbf{x}, \mathbf{b})]^2 \ .$$

Equivalently, we minimize the *root mean square error* (RMSE) $\sqrt{\frac{1}{n} \sum e_i^2}$, which is a useful measure of goodness-of-fit. In general, this optimization problem cannot be solved analytically, and it is necessary to resort to iterative numerical optimization methods. An important exception occurs when the model is a linear function of the parameters $\mathbf{b}$. In this *linear regression* situation, it is possible to obtain explicit solutions for the optimal (least-squares) values of the parameters or *coefficients*.

Our example of *H vs D* is an instance of linear regression. It can be

written

$$y = b_1 + b_2 x \ ,$$

with $y = H,\ x = \ln D$. This is a straight line, taking here the variable x as predictor. In general, both y and x can be transformations of the original variables. Ideally, the data would satisfy the n equations system

$$\begin{aligned} y_1 &= b_1 + b_2 x_1 \\ y_2 &= b_1 + b_2 x_2 \\ \vdots &\quad \vdots \\ y_n &= b_1 + b_1 x_n \end{aligned}$$

which in matrix notation can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$\mathbf{y} = \mathrm{X}\mathbf{b} \ .$$

If we had $n = 2$, we would have a system of two equations in two unknowns (b_1 y b_2), usually with a unique solution. In matrix terms, $\mathbf{y} = \mathrm{X}\mathbf{b}$ with X square and invertible has the solution $\mathbf{b} = \mathrm{X}^{-1}\mathbf{y}$.

With $n > 2$, in general not all the observations are co-linear, and the equation system is incompatible. The objective is to find a $\mathbf{b}$ such that the approximation $\mathbf{y} \approx \mathrm{X}\mathbf{b}$ is the best possible, in the sense of minimizing the length $|\mathbf{e}|$ of the vector $\mathbf{e} = \mathbf{y} - \mathrm{X}\mathbf{b}$ computed from a generalization to n dimensions of Pithagoras Theorem:

$$|\mathbf{e}|^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} \ .$$

There are algorithms, based on matrix factorization, that produce directly the least-squares solution of $\mathbf{y} \approx \mathrm{X}\mathbf{b}$. These are used in the better statistical packages. Sometimes, *pseudoinverses* or *generalized inverses* X^- are used, in terms of which the solution can be written as $\mathbf{b} = \mathrm{X}^-\mathbf{y}$. The APL computer language to be used in the laboratories has a generalized inversion and generalized matrix division operator that makes very simple the computation of linear regressions. In APL notation, the matrix product $\mathrm{X}\mathbf{b}$ is (indicating that we are dealing with sums of products). The

coefficients can be obtained with the generalized inverse, or, preferably, with the generalized matrix division .

Before presenting the least-squares solution most commonly used in textbooks and manual calculations, let us examine the more general *multiple* linear regression situation, where in contrast to the previous *simple* linear regression example in which there was just one predictor x there are now p predictors. The model is

$$y = b_1 x_1 + b_2 x_2 + \ldots b_p x_p = \mathbf{b}'\mathbf{x} = \mathbf{x}'\mathbf{b} \; .$$

Simple linear regression is the special case $p = 2,\ \mathbf{b} = (b_1, b_2),\ \mathbf{x} = (1, x)$. The system of equations, including now the deviations e_i, is

$$\begin{aligned} y_1 &= b_1 x_{11} + b_2 x_{12} + \cdots + b_p x_{1p} + e_1 \\ y_2 &= b_1 x_{21} + b_2 x_{22} + \cdots + b_p x_{2p} + e_2 \\ \vdots &\quad\ \vdots \\ y_n &= b_1 x_{n1} + b_2 x_{n2} + \cdots + b_p x_{np} + e_n \end{aligned}$$

that is,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$\mathbf{y} = \mathrm{X}\mathbf{b} + \mathbf{e} \; .$$

The matrix equation is the same as before, again we have to minimize $\mathbf{e}'\mathbf{e}$, and the direct factorization and APL solutions do not change. Almost always a constant is included in the model, and then x_1 and the x_{i1} equal 1.

The most usual explicit solution form is obtained as follows. To minimize the sum of squares $Q = \mathbf{e}'\mathbf{e}$, we make the derivative equal to zero:

$$\begin{aligned} Q &= \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathrm{X}\mathbf{b})'(\mathbf{y} - \mathrm{X}\mathbf{b}) \\ &\quad\ \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathrm{X}\mathbf{b} + \mathbf{b}'\mathrm{X}'\mathrm{X}\mathbf{b} \\ \frac{dQ}{d\mathbf{b}} &= -2\mathrm{X}'\mathbf{y} + 2\mathrm{X}'\mathrm{X}\mathbf{b} = 0 \; , \end{aligned}$$

what gives us the *normal equations*:

$$\mathrm{X}'\mathrm{X}\mathbf{b} = \mathrm{X}'\mathbf{y} \; .$$

The p equations may be solved numerically for the p unknowns $\mathbf{b}$. The solution may also be written explicitly:

$$\mathbf{b} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} \ .$$

The goodness of fit can be evaluated through the sum of squares $\mathbf{e}'\mathbf{e}$, through the root mean square error RMSE $= \sqrt{\mathbf{e}'\mathbf{e}/n}$, or the standard error SE $= \sqrt{\mathbf{e}'\mathbf{e}/(n-p)}$. The number of parameters p in the SE penalizes somewhat the complexity of the model when comparing alternatives, and has also a statistical justification explained in the following section.

Although this expression is useful in theoretical derivations, in general it is not the most advisable from the numeric point of view. First, the normal equations system can be solved with less work that that necessary for computing the inverse and the matrix product. Second, important catastrophic cancellation errors can occur, similar to those for the computation of variances demonstrated in the error propagation section. As already mentioned, the most accurate procedures are based on the factorization of X.

When the model includes a constant (column of ones in X), cancellation errors in the normal equations can be much reduced by "centering" the variables, as in the case of the variance, using deviations from the means instead of the original variables. For a model $y = b_0 + \mathbf{x}'\mathbf{b} + e$ it is seen that with the least-square parameters the means satisfy $\bar{y} = b_0 + \bar{\mathbf{x}}'\mathbf{b}$, since the first of the normal equations ensures that the sum of residuals is zero: $0 = \mathrm{X}'(\mathbf{y} - \mathrm{X}\mathbf{b}) = \mathrm{X}'\mathbf{e}$. Subtracting, we have the equivalent model $y - \bar{y} = (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{b} + e$. We estimate $\mathbf{b}$ with this model, and the constant is obtained from $b_0 = \bar{y} - \bar{\mathbf{x}}'\mathbf{b}$.

Questions, exercises

1. Estimate b in the model $y = b + e$. Do you recognize the result?

2. Verify that

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \ .$$

   Use this to obtain formulas for the two parameters in simple linear regression (with the original variables).

3. Obtain formulas for simple linear regression using the centered variables (deviations from the means).

4. Estimate by least squares the parameters of the model $y = b_1 + b_2 x + b_3 x^2$. Use centered variables.

5. Research the solution of linear equation systems and matrix inversion by Gaussian elimination.

3 Statistical considerations

We have presented least squares as a more or less reasonable and mathematically convenient method for "fitting" functions to observed data. Under certain probabilistic models for the deviations, the least squares criterion can also be justified by statistical arguments.

Assume first that the observations y_i are generated according to a model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \ ,$$

where the ε_i are uncorrelated random variables with mean 0 and unknown variance σ^2. That is,

$$E[\varepsilon_i] = 0 \, , \ \ E[\varepsilon_i^2] = \sigma^2 \, , \ \ E[\varepsilon_i \varepsilon_j] = 0 \text{ si } i \neq j \ ,$$

or, with matrix notation,

$$E[\boldsymbol{\varepsilon}] = \mathbf{0}, \ \ V[\boldsymbol{\varepsilon}] = \sigma^2 \mathrm{I} \, ,$$

where $V[\cdot]$ is the covariance matrix. The $\mathbf{x}_i$ are known predictor vectors, and $\boldsymbol{\beta}$ is a vector of unknown parameters to be estimated.

We look for an estimator $\hat{\boldsymbol{\beta}} = \mathbf{b}$ unbiased, i. e. $E[\mathbf{b}] = \boldsymbol{\beta}$, and with a variance as small as possible. Let us restrict the search also to estimators that are linear functions on the observations, $\mathbf{b} = \mathrm{A}\mathbf{y}$ for some matrix A. Then, the Gauss-Markov theorem says that for the linear minimum variance unbiased estimator $\mathrm{A} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$. This is the least-squares estimator.

The restriction to estimators that are linear on the observations may seem somewhat arbitrary. If we add the assumption that the deviations follow a normal distribution, the least squares criterion is obtained through a different route. Let the model, not necessarily linear, be

$$y_i = f(\mathbf{x}_i, \beta) + \varepsilon_i$$

with the ε_i normal, with mean 0, variance σ^2, and independent. That is,

$$\mathbf{y} = \mathbf{f}(\mathrm{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \ ,$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathrm{I}) \ .$$

The *likelihood function* is the probability of the model generating data like the observed. The *maximum likelihood* (ML) estimation method consists of estimating the unknown parameters as the values that maximize this function. Besides being intuitively reasonable, the MV estimators have a number of desirable statistical properties, especially in large samples.

Here the likelihood function equals the joint probability density of the y_i, considered as a function of $\boldsymbol{\beta}$ and σ^2. From the independence assumption, the joint density is the product of the (normal) densities of each y_i:

$$L = f_1(y_1)f_2(y_2)\cdots f_n(y_n) \ ,$$

with

$$f_i(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{\{y_i - f(\mathbf{x}_i, \boldsymbol{\beta})\}^2}{2\sigma^2}] \ .$$

The likelihood is then

$$L = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp[-\frac{\sum \varepsilon_i^2}{2\sigma^2}] \ .$$

Clearly, the $\boldsymbol{\beta}$ that maximizes L is that which minimizes the sum $\sum \varepsilon_i^2$. We conclude that, under this model, the ML estimator of $\boldsymbol{\beta}$ is the least-squares estimator.

It is also found, taking the derivative of L with respect to σ^2 and making it equal to zero, that the ML estimator of σ^2 is *mean square error* (MSE) $\sum \hat{\varepsilon}_i^2/n = \sum e_i^2/n$, the square of the RMSE. The expected value of $\sum e_i^2$, for linear models, turns out to be $(n-p)\sigma^2$, so that the MSE is biased. It is customary to use the unbiased estimator SE^2 for the residual variance σ^2, and the standard error SE as estimator for σ.

Another goodness-of-fit indicator often used, incorrectly, is the coefficient of determination $R^2 = 1 - \text{MSE}/S_y^2$, where $S_y^2 = \sum(y_i - \bar{y})^2/n$ is the variance of the observations y_i when the predictors are ignored. For comparing models with the same data, R^2 provides the same information as the MSE or RMSE. With different data sets, however, an R^2 close to one does not imply necessarily a tight relationship or a good model. Among other things, the total variance depends of how the sample has been selected, and unless this can be considered as a random sample from a multivariate distribution, it does not represent a characteristic of the population.

Questions, exercises

1. Compute a linear regression between y and x with the following data:

x	1	2	3	4	5	6	7	8	9	10
y	1	4	9	16	25	36	49	64	81	100

2. Compute R^2.

3. Plot the data and the regression line.

♡ It is seen that for the linear regression

$$E[\mathbf{b}] = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'E[\mathbf{y}] = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}\boldsymbol{\beta} = \boldsymbol{\beta}\ ,$$

so that $\mathbf{b}$ is an unbiased estimator. The same happens with any function linear on the parameters, and, in particular, the prediction expected value $\hat{y}(\mathbf{x}) = \mathbf{x}'\mathbf{b}$ equals $y(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ for any $\mathbf{x}$.

Because the covariance matrix $V[\mathrm{A}\mathbf{z}]$ for a linear transformation is $\mathrm{A}V[\mathbf{z}]A'$, it is found that

$$V[\mathbf{b}] = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}\ .$$

If $\boldsymbol{\varepsilon}$ is normal, this and the fact that any linear transformation of a normal vector is normal allow us to obtain confidence intervals and hypothesis tests for linear functions of b.

Obviously, in real life these statistical models cannot be expected to be fulfilled exactly. But it can be expected that the more we approach the assumptions, the better the estimators will be. For instance, if it is seen that the scatter of the residuals is not quite uniform (*heterocedasticity*), it would be advisable to employ some transformation that changes this situation. Another possible problem is the presence of autocorrelation (correlation among consecutive measurements). In particular, hypothesis tests are subject to the plausibility of the statistical model.

♡**Generalized least squares** Assume that in the linear model the covariance matrix for $\boldsymbol{\varepsilon}$ has the form $\sigma^2\mathrm{W}$, with a known matrix $\mathrm{W} \neq \mathrm{I}$. Maintaining the other assumptions, it is then found that both the minimum variance unbiased and the ML estimator are obtained by minimizing $\mathbf{e}'\mathrm{W}^{-1}\mathbf{e}$. The solution is $\mathbf{b} = (\mathrm{X}'\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{W}^{-1}\mathbf{y}$.

A good introduction to statistical inference is found in Chapter 2 of Graybill, for which there is a Spanish translation among the course materials. A general text with a good treatment of linear regression is Peña Sánchez de Rivera, D. "Estadística, Modelos y Métodos" (2 Vols.), Alianza Editorial, Madrid, 1992.