holds for all n-vectors x, y, and all scalars α, β that satisfy $\alpha + \beta = 1$. In other words, superposition holds for affine combinations of vectors. (For linear functions, superposition holds for any linear combinations of vectors.)

The matrix A and the vector b in the representation of an affine function as $f(x) = Ax + b$ are unique. These parameters can be obtained by evaluating f at the vectors $0, e_1, \ldots, e_n$, where e_k is the kth unit vector in $\mathbf{R}^n$. We have

$$A = \left[\begin{array}{cccc} f(e_1) - f(0) & f(e_2) - f(0) & \cdots & f(e_n) - f(0) \end{array} \right], \qquad b = f(0).$$

Just like affine scalar-valued functions, affine vector-valued functions are often called linear, even though they are linear only when the vector b is zero.

8.2 Linear function models

Many functions or relations between variables that arise in natural science, engineering, and social sciences can be *approximated* as linear or affine functions. In these cases we refer to the linear function relating the two sets of variables as a *model* or an *approximation*, to remind us that the relation is only an approximation, and not exact. We give a few examples here.

- *Price elasticity of demand.* Consider n goods or services with prices given by the n-vector p, and demands for the goods given by the n-vector d. A change in prices will induce a change in demands. We let $\delta^{\rm price}$ be the n-vector that gives the fractional change in the prices, *i.e.*, $\delta_i^{\rm price} = (p_i^{\rm new} - p_i)/p_i$, where $p^{\rm new}$ is the n-vector of new (changed) prices. We let $\delta^{\rm dem}$ be the n-vector that gives the fractional change in the product demands, *i.e.*, $\delta_i^{\rm dem} = (d_i^{\rm new} - d_i)/d_i$, where $d^{\rm new}$ is the n-vector of new demands. A linear demand elasticity model relates these vectors as $\delta^{\rm dem} = E^{\rm d}\delta^{\rm price}$, where $E^{\rm d}$ is the $n \times n$ *demand elasticity matrix.* For example, suppose $E_{11}^{\rm d} = -0.4$ and $E_{21}^{\rm d} = 0.2$. This means that a 1% increase in the price of the first good, with other prices kept the same, will cause demand for the first good to drop by 0.4%, and demand for the second good to increase by 0.2%. (In this example, the second good is acting as a *partial substitute* for the first good.)

- *Elastic deformation.* Consider a steel structure like a bridge or the structural frame of a building. Let f be an n-vector that gives the forces applied to the structure at n specific places (and in n specific directions), sometimes called a *loading.* The structure will deform slightly due to the loading. Let d be an m-vector that gives the displacements (in specific directions) of m points in the structure, due to the load, *e.g.*, the amount of sag at a specific point on a bridge. For small displacements, the relation between displacement and loading is well approximated as linear: $d = Cf$, where C is the $m \times n$ *compliance matrix.* The units of the entries of C are m/N.

8.2.1 Taylor approximation

Suppose $f : \mathbf{R}^n \to \mathbf{R}^m$ is differentiable, *i.e.*, has partial derivatives, and z is an n-vector. The first-order Taylor approximation of f near z is given by

$$\begin{array}{rcl}\hat{f}(x)_i & = & f_i(z) + \dfrac{\partial f_i}{\partial x_1}(z)(x_1 - z_1) + \cdots + \dfrac{\partial f_i}{\partial x_n}(z)(x_n - z_n)\\ & = & f_i(z) + \nabla f_i(z)^T(x-z),\end{array}$$

for $i = 1, \ldots, m$. (This is just the first-order Taylor approximation of each of the scalar-valued functions f_i, described in §2.2.) For x near z, $\hat{f}(x)$ is a very good approximation of $f(x)$. We can express this approximation in compact notation, using matrix-vector multiplication, as

$$\hat{f}(x) = f(z) + Df(z)(x-z), \tag{8.3}$$

where the $m \times n$ matrix $Df(z)$ is the *derivative* or *Jacobian* matrix of f at z (see §C.1). Its components are the partial derivatives of f,

$$Df(z)_{ij} = \frac{\partial f_i}{\partial x_j}(z), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n,$$

evaluated at the point z. The rows of the Jacobian are $\nabla f_i(z)^T$, for $i = 1, \ldots, m$. The Jacobian matrix is named for the mathematician Carl Gustav Jacob Jacobi.

As in the scalar-valued case, Taylor approximation is sometimes written with a second argument as $\hat{f}(x; z)$ to show the point z around which the approximation is made. Evidently the Taylor series approximation $\hat{f}$ is an affine function of x. (It is often called a linear approximation of f, even though it is not, in general, a linear function.)

8.2.2 Regression model

Recall the regression model (2.7)

$$\hat{y} = x^T\beta + v, \tag{8.4}$$

where the n-vector x is a feature vector for some object, β is an n-vector of weights, v is a constant (the offset), and $\hat{y}$ is the (scalar) value of the regression model prediction.

Now suppose we have a set of N objects (also called *samples* or *examples*), with feature vectors $x^{(1)}, \ldots, x^{(N)}$. The regression model predictions associated with the examples are given by

$$\hat{y}^{(i)} = (x^{(i)})^T\beta + v, \quad i = 1, \ldots, N.$$

These numbers usually correspond to predictions of the value of the outputs or responses. If in addition to the example feature vectors $x^{(i)}$ we are also given the actual value of the associated response variables, $y^{(1)}, \ldots, y^{(N)}$, then our *prediction errors* or *residuals* are

$$r^{(i)} = y^{(i)} - \hat{y}^{(i)}, \quad i = 1, \ldots, N.$$

(Some authors define the prediction errors as $\hat{y}^{(i)} - y^{(i)}$.)

We can express this using compact matrix-vector notation. We form the $n \times N$ feature matrix X with columns $x^{(1)}, \ldots, x^{(N)}$. We let $y^{\rm d}$ denote the N-vector whose entries are the actual values of the response for the N examples. (The superscript 'd' stands for 'data'.) We let $\hat{y}^{\rm d}$ denote the N-vector of regression model predictions for the N examples, and we let $r^{\rm d}$ denote the N-vector of residuals or prediction errors. We can then express the regression model predictions for this data set in matrix-vector form as

$$\hat{y}^{\rm d} = X^T \beta + v\mathbf{1}.$$

The vector of N prediction errors for the examples is given by

$$r^{\rm d} = y^{\rm d} - \hat{y}^{\rm d} = y^{\rm d} - X^T\beta - v\mathbf{1}.$$

We can include the offset v in the regression model by including an additional feature equal to one as the first entry of each feature vector:

$$\hat{y}^{\rm d} = \left[\begin{array}{c} \mathbf{1}^T \\ X \end{array}\right]^T \left[\begin{array}{c} v \\ \beta \end{array}\right] = \tilde{X}^T \tilde{\beta},$$

where $\tilde{X}$ is the new feature matrix, with a new first row of ones, and $\tilde{\beta} = (v, \beta)$ is the vector of regression model parameters. This is often written without the tildes, as $\hat{y}^{\rm d} = X^T\beta$, by simply including the feature one as the first feature.

The equation above shows that the N-vector of predictions for the N examples is a linear function of the model parameters (v, β). The N-vector of prediction errors is an affine function of the model parameters.

8.3 Systems of linear equations

Consider a set (also called a system) of m linear equations in n variables or unknowns $x_1, \ldots, x_n$:

$$\begin{array}{rcl} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n & = & b_1 \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n & = & b_2 \\ & \vdots & \\ A_{m1}x_1 + A_{m2}x_2 + \cdots + A_{mn}x_n & = & b_m. \end{array}$$

The numbers A_{ij} are called the *coefficients* in the linear equations, and the numbers b_i are called the *right-hand sides* (since by tradition, they appear on the right-hand