2.3 Regression model

In this section we describe a very commonly used affine function, especially when the n-vector x represents a feature vector. The affine function of x given by

$$\hat{y} = x^T\beta + v, \tag{2.7}$$

where β is an n-vector and v is a scalar, is called a *regression model*. In this context, the entries of x are called the *regressors*, and $\hat{y}$ is called the *prediction*, since the regression model is typically an approximation or prediction of some true value y, which is called the *dependent variable*, *outcome*, or *label*.

The vector β is called the *weight vector* or *coefficient vector*, and the scalar v is called the *offset* or *intercept* in the regression model. Together, β and v are called the *parameters* in the regression model. (We will see in chapter 13 how the parameters in a regression model can be estimated or guessed, based on some past or known observations of the feature vector x and the associated outcome y.) The symbol $\hat{y}$ is used in the regression model to emphasize that it is an *estimate* or *prediction* of some outcome y.

The entries in the weight vector have a simple interpretation: β_i is the amount by which $\hat{y}$ increases (if $\beta_i > 0$) when feature i increases by one (with all other features the same). If β_i is small, the prediction $\hat{y}$ doesn't depend too strongly on feature i. The offset v is the value of $\hat{y}$ when all features have the value 0.

The regression model is very interpretable when all of the features are Boolean, *i.e.*, have values that are either 0 or 1, which occurs when the features represent which of two outcomes holds. As a simple example consider a regression model for the lifespan of a person in some group, with $x_1 = 0$ if the person is female ($x_1 = 1$ if male), $x_2 = 1$ if the person has type II diabetes, and $x_3 = 1$ if the person smokes cigarettes. In this case, v is the regression model estimate for the lifespan of a female nondiabetic nonsmoker; β_1 is the increase in estimated lifespan if the person is male, β_2 is the increase in estimated lifespan if the person is diabetic, and β_3 is the increase in estimated lifespan if the person smokes cigarettes. (In a model that fits real data, all three of these coefficients would be negative, meaning that they decrease the regression model estimate of lifespan.)

Simplified regression model notation. Vector stacking can be used to lump the weights and offset in the regression model (2.7) into a single parameter vector, which simplifies the regression model notation a bit. We create a new regressor vector $\tilde{x}$, with $n + 1$ entries, as $\tilde{x} = (1, x)$. We can think of $\tilde{x}$ as a new feature vector, consisting of all n original features, and one new feature added ($\tilde{x}_1$) at the beginning, which always has the value one. We define the parameter vector $\tilde{\beta} = (v, \beta)$, so the regression model (2.7) has the simple inner product form

$$\hat{y} = x^T\beta + v = \begin{bmatrix} 1 \\ x \end{bmatrix}^T \begin{bmatrix} v \\ \beta \end{bmatrix} = \tilde{x}^T\tilde{\beta}. \tag{2.8}$$

Often we omit the tildes, and simply write this as $\hat{y} = x^T\beta$, where we assume that the first feature in x is the constant 1. A feature that always has the value 1 is not particularly informative or interesting, but it does simplify the notation in a regression model.

House	x_1 (area)	x_2 (beds)	y (price)	$\hat{y}$ (prediction)
1	0.846	1	115.00	161.37
2	1.324	2	234.50	213.61
3	1.150	3	198.00	168.88
4	3.037	4	528.00	430.67
5	3.984	5	572.50	552.66

Table 2.3 Five houses with associated feature vectors shown in the second and third columns. The fourth and fifth column give the actual price, and the price predicted by the regression model.

House price regression model. As a simple example of a regression model, suppose that y is the selling price of a house in some neighborhood, over some time period, and the 2-vector x contains attributes of the house:

- x_1 is the house area (in 1000 square feet),
- x_2 is the number of bedrooms.

If y represents the selling price of the house, in thousands of dollars, the regression model

$$\hat{y} = x^T\beta + v = \beta_1 x_1 + \beta_2 x_2 + v$$

predicts the price in terms of the attributes or features. This regression model is not meant to describe an exact relationship between the house attributes and its selling price; it is a model or approximation. Indeed, we would expect such a model to give, at best, only a crude approximation of selling price.

As a specific numerical example, consider the regression model parameters

$$\beta = (148.73, -18.85), \qquad v = 54.40. \tag{2.9}$$

These parameter values were found using the methods we will see in chapter 13, based on records of sales for 774 houses in the Sacramento area. Table 2.3 shows the feature vectors x for five houses that sold during the period, the actual sale price y, and the predicted price $\hat{y}$ from the regression model above. Figure 2.4 shows the predicted and actual sale prices for 774 houses, including the five houses in the table, on a scatter plot, with actual price on the horizontal axis and predicted price on the vertical axis.

We can see that this particular regression model gives reasonable, but not very accurate, predictions of the actual sale price. (Regression models for house prices that are used in practice use many more than two regressors, and are much more accurate.)

The model parameters in (2.9) are readily interpreted. The parameter $\beta_1 = 148.73$ is the amount the regression model price prediction increases (in thousands of dollars) when the house area increases by 1000 square feet (with the same number of bedrooms). The parameter $\beta_2 = -18.85$ is the price prediction increase with the addition of one bedroom, with the total house area held constant, in units of

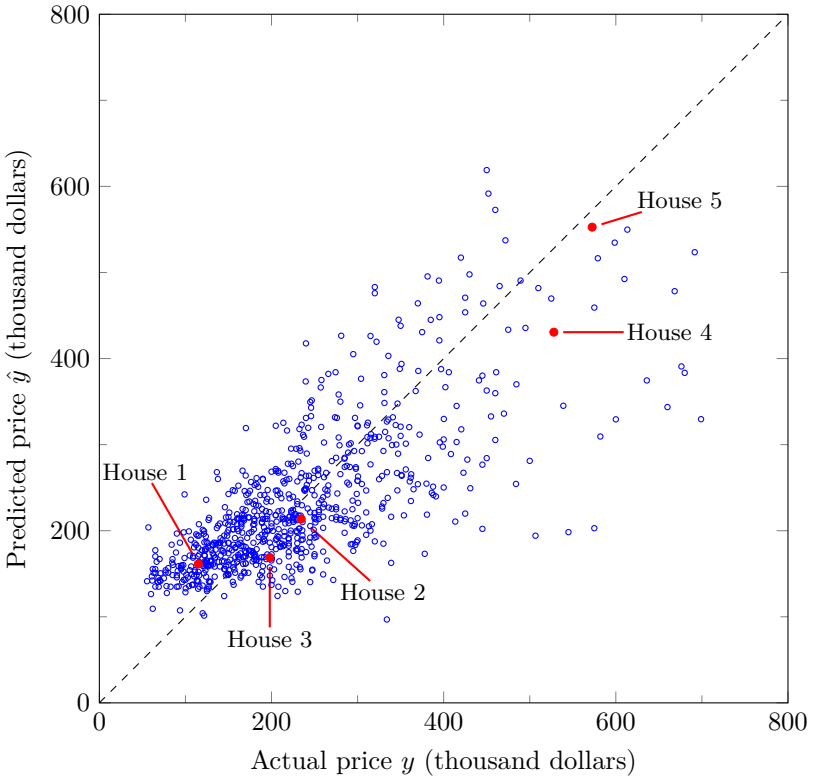

Figure 2.4 Scatter plot of actual and predicted sale prices for 774 houses sold in Sacramento during a five-day period.

thousands of dollars per bedroom. It might seem strange that β_2 is negative, since one imagines that adding a bedroom to a house would *increase* its sale price, not decrease it. To understand why β_2 might be negative, we note that it gives the change in predicted price when we add a bedroom, without adding any additional area to the house. If we remodel a house by adding a bedroom that *also* adds more than around 127 square feet to the house area, the regression model (2.9) *does* predict an increase in house sale price. The offset $v = 54.40$ is the predicted price for a house with no area and no bedrooms, which we might interpret as the model's prediction of the value of the lot. But this regression model is crude enough that these interpretations are dubious.