1.2.1 THE LINEAR REGRESSION MODEL

The data consist of n sets of observations $\{x_{1i}, x_{2i}, \ldots, x_{pi}, y_i\}$, which represent a random sample from a larger population. It is assumed that these observations satisfy a linear relationship,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \tag{1.1}$$

where the β coefficients are unknown parameters, and the ε_i are random error terms. By a *linear* model, it is meant that the model is linear in the *parameters*; a quadratic model,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

*

The errors are normally distributed. This is needed if we want to construct any confidence or prediction intervals, or hypothesis tests, which we usually do. If this assumption is violated, hypothesis tests and confidence and prediction intervals can be very misleading.

$$(E(\varepsilon_i) = 0 \text{ for all } i).$$

$$(V(\varepsilon_i) = \sigma^2 \text{ for all } i).$$

The errors are uncorrelated with each other.

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix} \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The regression model (1.1) is then

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The normal equations [which determine the minimizer of (1.2)] can be shown (using multivariate calculus) to be

$$(X'X)\hat{\boldsymbol{\beta}} = X'\mathbf{y},$$

which implies that the least squares estimates satisfy

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}.$$

The fitted values are then

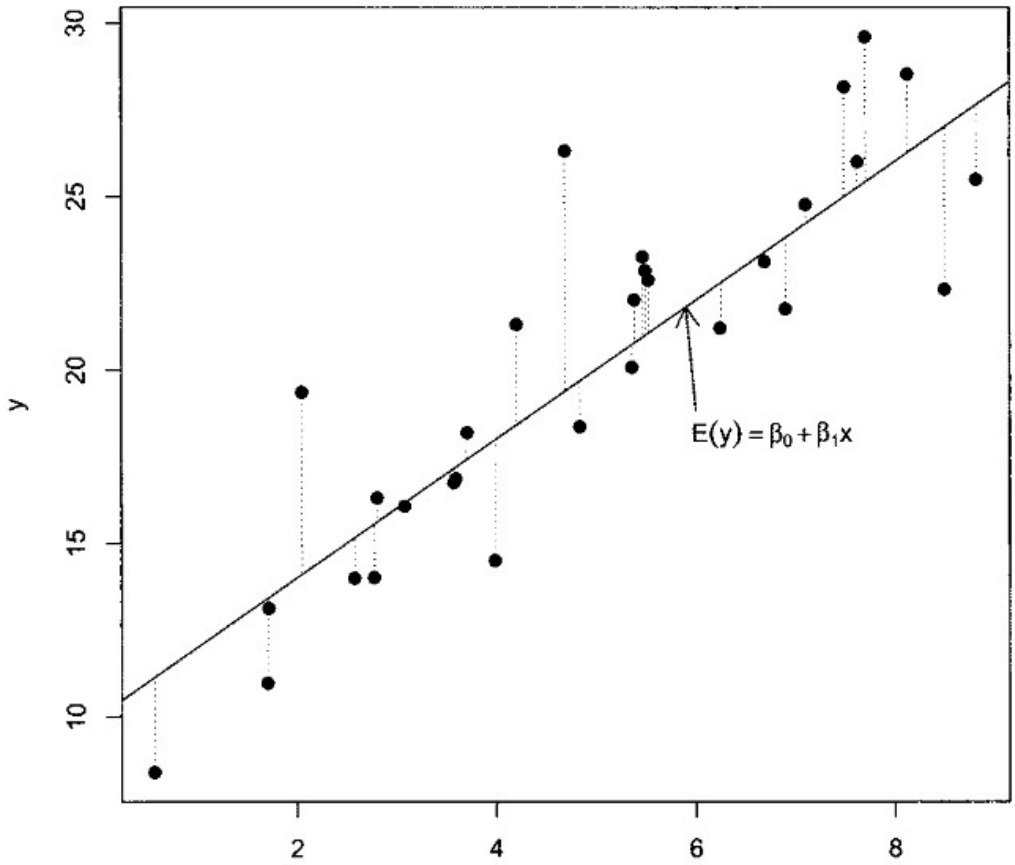
$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\mathbf{y} \equiv H\mathbf{y}, \tag{1.3}$$

where $H = X(X'X)^{-1}X'$ is the so-called "hat" matrix (since it takes $\mathbf{y}$ to $\hat{\mathbf{y}}$). The residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ thus satisfy

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X(X'X)^{-1}X'\mathbf{y} = (I - X(X'X)^{-1}X')\mathbf{y}, \tag{1.4}$$

or

$$\mathbf{e} = (I - H)\mathbf{y}.$$



*

*

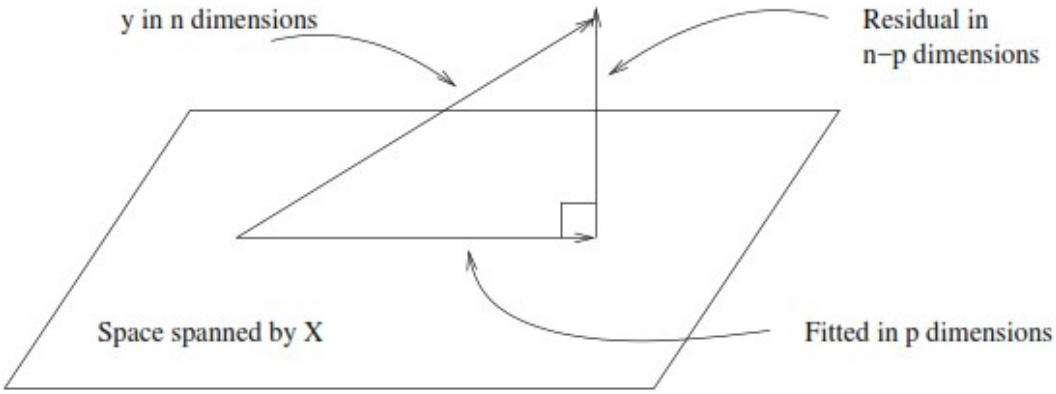


Figure 2.1: Geometric representation of the estimation β. The data vector Y is projected orthogonally onto the model space spanned by X. The fit is represented by projection $\hat{y} = X\hat{\beta}$ with the difference between the fit and the data represented by the residual vector $\hat{\varepsilon}$.

- Predicted values: $\hat{y} = Hy = X\hat{\beta}$.
- Residuals: $\hat{\varepsilon} = y - X\hat{\beta} = y - \hat{y} = (I - H)y$
- Residual sum of squares: $\hat{\varepsilon}^T\hat{\varepsilon} = y^T(I - H)(I - H)y = y^T(I - H)y$

*

2.7 Why is $\hat{\beta}$ a good estimate?

1. It results from an orthogonal projection onto the model space. It makes sense geometrically.
2. 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 β that maximizes the probability of the data that was observed.
3. The Gauss-Markov theorem states that it is best linear unbiased estimate. (BLUE).

First we need to understand the concept of an *estimable function*. A linear combination of the parameters $\psi = c^T\beta$ is estimable if and only if there exists a linear combination $a^T y$ such that

$$Ea^T y = c^T\beta \qquad \forall\beta$$

ering. If X is of full rank (which it usually is for observational data), then all linear combinations are estimable.

Situations where estimators other than ordinary least squares should be considered are

1. When the errors are correlated or have unequal variance, generalized least squares should be used.
2. When the error distribution is long-tailed, then robust estimates might be used. Robust estimates are typically not linear in y.
3. When the predictors are highly correlated (collinear), then biased estimators such as ridge regression might be preferable.

2.9 Mean and Variance of $\hat{\beta}$

Now $\hat{\beta} = (X^T X)^{-1} X^T y$ so

- Mean $E\hat{\beta} = (X^T X)^{-1} X^T X\beta = \beta$ (unbiased)
- var $\hat{\beta} = (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} = (X^T X)^{-1}\sigma^2$

Note that since $\hat{\beta}$ is a vector, $(X^T X)^{-1}\sigma^2$ is a variance-covariance matrix. Sometimes you want the standard error for a particular component which can be picked out as in $se(\hat{\beta}_i) = \sqrt{(X^T X)^{-1}_{ii}}\hat{\sigma}$.

2.10 Estimating σ^2

Recall that the residual sum of squares was $\hat{\varepsilon}^T\hat{\varepsilon} = y^T(I-H)y$. Now after some calculation, one can show that $E\hat{\varepsilon}^T\hat{\varepsilon} = \sigma^2(n-p)$ which shows that

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-p}$$

is an unbiased estimate of σ^2. $n-p$ is the *degrees of freedom* of the model. Actually a theorem parallel to the Gauss-Markov theorem shows that it has the minimum variance among all quadratic unbiased estimators of σ^2.

2.11 Goodness of Fit

How well does the model fit the data? One measure is R^2, the so-called *coefficient of determination* or *percentage of variance explained*

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{Total SS(corrected for mean)}}$$

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

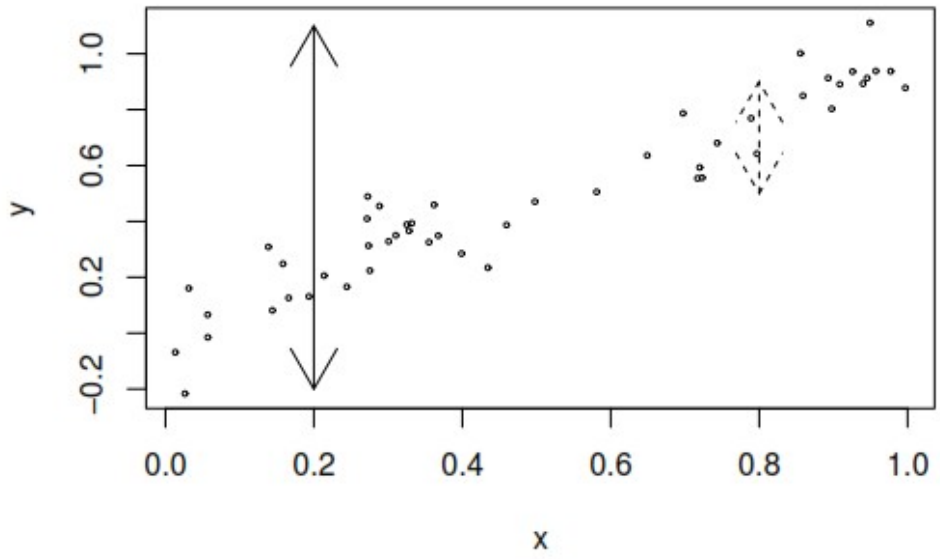


Figure 2.2: Variation in the response y when x is known is denoted by dotted arrows while variation in y when x is unknown is shown with the solid arrows