at $\boldsymbol{X}\hat{\beta}$ instead of at zero. The right angle formed by $\boldsymbol{y} - \boldsymbol{X}\hat{\beta}$ and $\mathcal{S}(\boldsymbol{X})$ is the key feature of least squares. At any other point in $\mathcal{S}(\boldsymbol{X})$, such as $\boldsymbol{X}\beta'$ in the figure, $\boldsymbol{y} - \boldsymbol{X}\beta'$ does not form a right angle with $\mathcal{S}(\boldsymbol{X})$ and, as a consequence, $\|\boldsymbol{y} - \boldsymbol{X}\beta'\|$ must necessarily be larger than $\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|$.

The vector of derivatives of the SSR (1.02) with respect to the elements of $\boldsymbol{\beta}$ is

$$-2\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

which must equal $\mathbf{0}$ at a minimum. Since we have assumed that the columns of $\boldsymbol{X}$ are linearly independent, the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ must have full rank. This, combined with that fact that any matrix of the form $\boldsymbol{X}^\top\boldsymbol{X}$ is necessarily nonnegative definite, implies that the sum of squared residuals is a strictly convex function of $\boldsymbol{\beta}$ and must therefore have a unique minimum. Thus $\hat{\boldsymbol{\beta}}$ is uniquely determined by the **normal equations**

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{1.03}$$

These normal equations say that the vector $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ must be orthogonal to all of the columns of $\boldsymbol{X}$ and hence to any vector that lies in the space spanned by those columns. The normal equations (1.03) are thus simply a way of stating algebraically what Figure 1.2 showed geometrically, namely, that $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ must form a right angle with $\mathcal{S}(\boldsymbol{X})$.

Since the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ has full rank, we can always invert it to solve the normal equations for $\hat{\boldsymbol{\beta}}$. We obtain the standard formula:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \tag{1.04}$$

Even if $\boldsymbol{X}$ is not of full rank, the fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ are uniquely defined, because $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is simply the point in $\mathcal{S}(\boldsymbol{X})$ that is closest to $\boldsymbol{y}$. Look again at Figure 1.2 and suppose that $\boldsymbol{X}$ is an $n \times 2$ matrix, but of rank only one. The geometrical point $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is still uniquely defined. However, since $\boldsymbol{\beta}$ is now a 2–vector and $\mathcal{S}(\boldsymbol{X})$ is just one-dimensional, the vector $\hat{\boldsymbol{\beta}}$ is not uniquely defined. Thus the requirement that $\boldsymbol{X}$ have full rank is a purely algebraic requirement that is needed to obtain unique estimates $\hat{\boldsymbol{\beta}}$.

If we substitute the right-hand side of (1.04) for $\hat{\boldsymbol{\beta}}$ into $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, we obtain

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \equiv \boldsymbol{P}_X\boldsymbol{y}. \tag{1.05}$$

This equation defines the $n \times n$ matrix $\boldsymbol{P}_X \equiv \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, which **projects** the vector $\boldsymbol{y}$ orthogonally onto $\mathcal{S}(\boldsymbol{X})$. The matrix $\boldsymbol{P}_X$ is an example of an **orthogonal projection matrix**. Associated with every linear subspace of E^n are two such matrices, one of which projects any point in E^n onto that subspace, and one of which projects any point in E^n onto its orthogonal complement. The matrix that projects onto $\mathcal{S}^\perp(\boldsymbol{X})$ is

$$\boldsymbol{M}_X \equiv \mathbf{I} - \boldsymbol{P}_X \equiv \mathbf{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top,$$