Where $e_i = y_i - \hat{y}_i$ are the residuals for each data point, and $\hat{y}_i$ is the predicted value.
In matrix form, this becomes:

$$RSS = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

3. Minimizing the RSS

We need to differentiate the RSS with respect to $\boldsymbol{\beta}$ and set it equal to zero to find the $\boldsymbol{\beta}$ that minimizes the RSS. To do this, we first expand the RSS expression:

$$RSS = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

Now, differentiate the RSS with respect to $\boldsymbol{\beta}$:

$$\frac{\partial}{\partial \boldsymbol{\beta}} RSS = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

Set this derivative equal to zero:

$$-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0$$