Today we focus on the geometry of least squares via projection mappings. Specifically, each data point, $\mathbf{x} = (x_1, \ldots, x_n)$ can be regarded as a point in n dimensional space. We're interesting in the space of all linear combinations of the random variables $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_p}$. First,

$$\overline{\mathbf{X}}^* = \begin{bmatrix} \overline{x_1} & \overline{x_2} & \ldots & \overline{x_p} \\ \vdots & \vdots & \vdots & \vdots \\ \overline{x_1} & \overline{x_2} & \ldots & \overline{x_p} \end{bmatrix} \quad \boldsymbol{\beta}^* = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad \mathbf{X}^* = \begin{bmatrix} x_{11} & x_{21} & \ldots & x_{p1} \\ x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \ldots & x_{pn} \end{bmatrix}$$

Then we can rewrite the mean corrected MLR model as

$$\mathbf{Y} = \alpha \cdot 1 + (\mathbf{X}^* - \overline{\mathbf{X}}^*)\boldsymbol{\beta}^* + \mathbf{e} \tag{1}$$

where $\alpha = \beta_0 \cdot 1 + \overline{\mathbf{X}}^* \boldsymbol{\beta}^*$. One can show that $\widehat{\alpha} = \overline{y}$. So, roughly, we get

$$(y_i - \overline{y}) = (\mathbf{X}^* - \overline{\mathbf{X}}^*)\boldsymbol{\beta}^* + \mathbf{e} \tag{2}$$

Call this model now

$$\mathcal{Y} = \mathcal{X}\boldsymbol{\beta}^* + \mathbf{e} \tag{3}$$

This gives rise to the OLS estimate of $\boldsymbol{\beta}^*$ as

$$\widehat{\boldsymbol{\beta}^*} = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathcal{Y}$$

This solution solves the problem $\min_b (\mathcal{Y} - \widehat{\mathcal{Y}})^T(\mathcal{Y} - \widehat{\mathcal{Y}})$ where $\mathcal{Y}$ *must be in the column space of* $\boldsymbol{X}$. Identically, we can consider this problem as

$$\min_{\widehat{y} \in col(\mathbf{X})} ||\mathcal{Y} - \widehat{\mathcal{Y}}||_2^2 \tag{4}$$

We can achieve this minimization by choosing $\widehat{\mathcal{Y}}$ as the the point on the span of $\mathbf{X}$ closest to $\mathcal{Y}$. This corresponds to $\mathcal{Y}$'s projection onto $col(\mathbf{X})$. The projection map is given by

$$H = \mathcal{X}(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T \tag{5}$$

This gives a really nice interpretation, because then we see $e^T\widehat{\mathcal{Y}} = 0$ i.e. the residual space and the column space are orthogonal. Moreover, we have

$$SSY = ||\mathcal{Y}||_2^2 \qquad R^2 = 1 - \frac{||\mathbf{e}||_2^2}{||\mathcal{Y}||_2^2}$$

Moreover we can think of ANOVA in a much much cleaner sense. We can decompose the variance in $\mathcal{Y}$ by

$$||\mathcal{Y}||_2^2 = ||\widehat{\mathcal{Y}}||_2^2 + ||\widehat{\mathbf{e}}||_2^2 = ||\widehat{\mathcal{Y}}||_2^2 + ||(I - H)\mathcal{Y}||_2^2$$

and think of degrees of freedom as simply dimensions of subspaces.