Simple Linear Regression - Part Deux

Just the Essentials of Linear Algebra

Vector: A set of *ordered* numbers, written, by default, as a column. Example: $\boldsymbol{v} = \begin{bmatrix} 2 \\ -5 \\ 1 \\ 3 \end{bmatrix}$

Transpose of a Vector: The transpose of vector $\boldsymbol{v}$, written $\boldsymbol{v}^T$, is defined as the corresponding ordered row of numbers. Example: The transpose of the vector $\boldsymbol{v}$ above is $\boldsymbol{v}^T = \begin{bmatrix} 2 & -5 & 1 & 3 \end{bmatrix}$

Scalar Multiplication: Scalar (real number) multiplication distributes over components. Ex: $4\boldsymbol{v} = 4\begin{bmatrix} 2 \\ -5 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 8 \\ -20 \\ 4 \\ 12 \end{bmatrix}$

Vector Addition: Addition is defined componentwise. Example: For vectors $\boldsymbol{v} = \begin{bmatrix} 2 \\ -5 \\ 1 \\ 3 \end{bmatrix}$ and $\boldsymbol{u} = \begin{bmatrix} -1 \\ 3 \\ -2 \\ 4 \end{bmatrix}$,

$$\boldsymbol{v} + \boldsymbol{u} = \begin{bmatrix} 2 \\ -5 \\ 1 \\ 3 \end{bmatrix} + \begin{bmatrix} -1 \\ 3 \\ -2 \\ 4 \end{bmatrix} = \begin{bmatrix} 2-1 \\ -5+3 \\ 1-2 \\ 3+4 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ -1 \\ 7 \end{bmatrix} = \boldsymbol{u} + \boldsymbol{v}$$

Vector Multiplication: The inner product (or dot product) of vectors $\boldsymbol{v}$ and $\boldsymbol{u}$ is defined as

$$\boldsymbol{v}^T\boldsymbol{u} = \begin{bmatrix} 2 & -5 & 1 & 3 \end{bmatrix}\begin{bmatrix} -1 \\ 3 \\ -2 \\ 4 \end{bmatrix} = (2)(-1) + (-5)(3) + (1)(-2) + (3)(4) = -2 - 15 - 2 + 12 = -7$$

Note 1: The inner product of two vectors is a number!

Note 2: $\boldsymbol{v}^T\boldsymbol{u} = \boldsymbol{u}^T\boldsymbol{v}$

Orthogonal: Vectors $\boldsymbol{v}$ and $\boldsymbol{u}$ are *orthogonal* if $\boldsymbol{v}^T\boldsymbol{u} = 0$. The geometrical interpretation is that the vectors are perpendicular. Orthogonality of vectors plays a big role in linear models.

Length of a Vector: The length of vector $\boldsymbol{v}$ is given by $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}^T\boldsymbol{v}}$. This extends the Pythagorean theorem to vectors with an arbitrary number of components. Example: For $\boldsymbol{v}^T = \begin{bmatrix} 2 & -5 & 1 & 3 \end{bmatrix}$, $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}^T\boldsymbol{v}} = \sqrt{39}$.

Matrices: $\boldsymbol{A}$ is an *m* by *n* matrix if it is an array of numbers with *m* rows and *n* columns. Example: $\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ -1 & 3 \\ 1 & 3 \\ 4 & 2 \end{bmatrix}$ is a 4×2 matrix.

Scalar Multiplication: Scalar (real number) multiplication distributes over components, as for vectors.

Matrix Addition: Matrix addition is componentwise, as for vectors.

Note: A vector is a special case of a matrix with a single column (or single row in the case of its transpose).

Matrix Multiplication: If $\boldsymbol{A}$ is $m \times n$ and $\boldsymbol{B}$ is $n \times p$, the product $\boldsymbol{AB}$ is defined as the matrix of all inner products of row vectors of $\boldsymbol{A}$ with column vectors of $\boldsymbol{B}$. Example: If $\boldsymbol{A}_{4\times 2} = \begin{bmatrix} 2 & 1 \\ -1 & 3 \\ 1 & 3 \\ 4 & 2 \end{bmatrix}$ and $\boldsymbol{B}_{2\times 3} = \begin{bmatrix} 1 & 7 & 3 \\ -1 & 2 & -1 \end{bmatrix}$, then

$$(\boldsymbol{AB})_{4\times 3} = \begin{bmatrix} 1 & 16 & 5 \\ -4 & -1 & -6 \\ -2 & 13 & 0 \\ 2 & 32 & 10 \end{bmatrix}.$$

Transpose of a Matrix: The transpose of $m \times n$ matrix $\boldsymbol{A}$ is the $n \times m$ matrix $\boldsymbol{A}^T$ obtained by swapping rows for columns in $\boldsymbol{A}$. Example: If $\boldsymbol{A}_{4\times 2} = \begin{bmatrix} 2 & 1 \\ -1 & 3 \\ 1 & 3 \\ 4 & 2 \end{bmatrix}$, then $\boldsymbol{A}^T_{2\times 4} = \begin{bmatrix} 2 & -1 & 1 & 4 \\ 1 & 3 & 3 & 2 \end{bmatrix}$

Transpose of a Product: $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$

Symmetric Matrix: The square matrix $\boldsymbol{A}_{n\times n}$ is *symmetric* if $a_{ij} = a_{ji}$ for all $1 \le i \le n$, $1 \le j \le n$ where a_{ij} is the entry in the i^th^ row, j^th^ column of $\boldsymbol{A}_{n\times n}$. Example: In $\boldsymbol{A}_{3\times 3} = \begin{bmatrix} 1 & -3 & 1 \\ -3 & 7 & -2 \\ 1 & -2 & 5 \end{bmatrix}$: $a_{12} = a_{21} = -3$; $a_{13} = a_{31} = 1$; $a_{23} = a_{32} = -2$.

(**Note:** If $\boldsymbol{A}_{n\times n}$ is symmetric, then $\boldsymbol{A}^T = \boldsymbol{A}$)

Theorem: for any $m \times n$ matrix $\boldsymbol{A}$, $\boldsymbol{A}^T\boldsymbol{A}$ is an $n \times n$ symmetric matrix, and $\boldsymbol{AA}^T$ is an $m \times m$ symmetric matrix.

Linear Independence: The set of *n* vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \cdots, \boldsymbol{v}_{\boldsymbol{n}}$ form a linearly independent set if the linear combination $c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + c_3\boldsymbol{v}_3 + \cdots + c_n\boldsymbol{v}_n = 0$ implies that all of the constant coefficients c_i equal zero.

rank(*A*): The rank of matrix ***A***, rank(***A***), is the number of linearly independent columns (or rows) in the matrix. A matrix with all columns, or all rows, independent is said to be "full rank." Full rank matrices play an important role in linear models, especially regression.

Theorem about Rank: rank($\boldsymbol{A}^T\boldsymbol{A}$) = rank($\boldsymbol{A}\boldsymbol{A}^T$) = rank(***A***). This theorem will be used in regression.

Inverse of a Matrix: The inverse of the $n \times n$ square matrix ***A*** is the $n \times n$ square matrix $\boldsymbol{A}^{-1}$ such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} =$ ***I***, where ***I*** is the $n \times n$ "Identity" matrix, $\boldsymbol{I}_{n\times n} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{n\times n}$

Theorem: The $n \times n$ square matrix ***A*** has an inverse if and only if it is full rank, i.e., rank(***A***) = *n*.

Theorem: The invertible 2×2 matrix $\boldsymbol{A}_{2\times 2} = \begin{bmatrix} \boldsymbol{a} & \boldsymbol{b} \\ \boldsymbol{c} & \boldsymbol{d} \end{bmatrix}$ has 2×2 inverse $\boldsymbol{A}^{-1}_{2\times 2} = \dfrac{1}{\boldsymbol{ad} - \boldsymbol{bc}} \begin{bmatrix} \boldsymbol{d} & -\boldsymbol{b} \\ -\boldsymbol{c} & \boldsymbol{a} \end{bmatrix}$. Example: If $\boldsymbol{A} = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}$, then $\boldsymbol{A}^{-1} = \dfrac{1}{(2)(1)-(4)(3)} \begin{bmatrix} 1 & -4 \\ -3 & 2 \end{bmatrix} = \dfrac{1}{-10} \begin{bmatrix} 1 & -4 \\ -3 & 2 \end{bmatrix} = -\dfrac{1}{10} \begin{bmatrix} 1 & -4 \\ -3 & 2 \end{bmatrix}$. You should carry out the products $\boldsymbol{A}\boldsymbol{A}^{-1}$ and $\boldsymbol{A}^{-1}\boldsymbol{A}$ to confirm that they equal $\boldsymbol{I}_{2\times 2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Solving a System of *n* Linear Equations in *n* Unknowns

The system of *n* linear equations in *n* unknowns,

$$\begin{aligned} \boldsymbol{a}_{11}\boldsymbol{x}_1 + \boldsymbol{a}_{12}\boldsymbol{x}_2 + \cdots + \boldsymbol{a}_{1n}\boldsymbol{x}_n &= \boldsymbol{b}_1 \\ \boldsymbol{a}_{21}\boldsymbol{x}_1 + \boldsymbol{a}_{22}\boldsymbol{x}_2 + \cdots + \boldsymbol{a}_{2n}\boldsymbol{x}_n &= \boldsymbol{b}_2 \\ &\vdots \\ \boldsymbol{a}_{n1}\boldsymbol{x}_1 + \boldsymbol{a}_{n2}\boldsymbol{x}_2 + \cdots + \boldsymbol{a}_{nn}\boldsymbol{x}_n &= \boldsymbol{b}_n \end{aligned}$$

, can be represented by the matrix equation $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A}_{n\times n} = \begin{bmatrix} \boldsymbol{a}_{11} & \boldsymbol{a}_{12} & \cdots & \boldsymbol{a}_{1n} \\ \boldsymbol{a}_{21} & \boldsymbol{a}_{22} & \cdots & \boldsymbol{a}_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{a}_{n1} & \boldsymbol{a}_{n2} & \cdots & \boldsymbol{a}_{nn} \end{bmatrix}_{n\times n}$, $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix}$, and $\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_n \end{bmatrix}$ are the matrix of coefficients, vector of unknowns, and vector of constants, respectively. The system has a unique solution *if and only if* ***A*** is invertible, i.e., $\boldsymbol{A}^{-1}$ exists. In that case, the solution is given by $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$.

Representing the Simple Linear Regression Model as a Matrix Equation

For a sample of *n* observations on the bivariate distribution of the variables ***X*** and ***Y***, the simple linear regression model $\boldsymbol{Y} = \beta_0 + \beta_1 X + \boldsymbol{\varepsilon}$ leads to the system of *n* equations $\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 + \boldsymbol{\varepsilon}_1 \\ \beta_0 + \beta_1 X_2 + \boldsymbol{\varepsilon}_2 \\ \\ \beta_0 + \beta_1 X_{\boldsymbol{n}} + \boldsymbol{\varepsilon}_{\boldsymbol{n}} \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_{\boldsymbol{n}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_{\boldsymbol{n}} \end{bmatrix}$,

which can be written $\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_{\boldsymbol{n}} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_{\boldsymbol{n}} \end{bmatrix}$. The shorthand for this is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the $n \times 2$ matrix $\mathbf{X}$ is called the "design" matrix because the values of the variable ***X*** are often fixed by the researcher in the design of the experiment used to investigate the relationship between the variables ***X*** and ***Y***.

Fitting the Best Least Squares Regression Line to the *n* Observations

Ideally, we would solve the matrix equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for the vector of regression coefficients $\boldsymbol{\beta}$, i.e., the true intercept β_0 and true slope β_1 in the model $\boldsymbol{Y} = \beta_0 + \beta_1 X + \boldsymbol{\varepsilon}$. In practice, however, this is never possible because the equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *has no knowable solution*! (Why?) When faced with an equation that we cannot solve, we do what mathematicians usually do: we find a related equation that we *can* solve. First, since we don't know the errors in our observations, we forget about the vector of the errors $\boldsymbol{\varepsilon}$. (Here, it is important that we not confuse the errors $\boldsymbol{\varepsilon_i}$, which we never know, with the residuals e_{i} associated with our estimated regression line.)

Unable to determine the parameters β_0 and β_1, we look to estimate them, i.e., solve for $\hat{\beta}_0$ and $\hat{\beta}_1$ in the matrix equation $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$, but the system involves *n* equations in the two unknowns $\hat{\beta}_0$ and $\hat{\beta}_1$. If you remember your algebra, overdetermined systems of linear equations rarely have solutions. What we need is a *related* system of linear equations that we *can* solve. (Of course, we'll have to show that the solution to the new system has relevance to our problem of estimating β_0 and β_1.) Finally (drum roll please), the system we'll actually solve is,

(1.1)
$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$$

Next, we show what the system above looks like, and, as we go, we'll see why it's solvable and why it's relevant.

(1.2)
$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & \vdots \\ 1 & X_{\boldsymbol{n}} \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

(1.3)
$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} n\hat{\beta}_0 + \hat{\beta}_1\sum X_i \\ \hat{\beta}_0\sum X_i + \hat{\beta}_1\sum X_i^2 \end{bmatrix}$$

(1.4)
$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix}\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_{\boldsymbol{n}} \end{bmatrix} = \begin{bmatrix} \sum \boldsymbol{Y}_i \\ \sum X_i\boldsymbol{Y}_i \end{bmatrix}$$

Now we're in a position to write out the equations for the system $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$,

(1.5)
$$n\hat{\beta}_0 + \hat{\beta}_1\sum X_i = \sum \boldsymbol{Y}_i$$

$$\hat{\beta}_0\sum X_i + \hat{\beta}_1\sum X_i^2 = \sum X_i\boldsymbol{Y}_i$$

If these equations look familiar, they are the equations derived in the previous notes by applying the Least Squares criterion for the best fitting regression line. Thus, the solution to the matrix equation (1.1) is the least squares solution to the problem of fitting a line to data! (Although we could have established this directly using theorems from linear algebra and vector spaces, the calculus argument made in Part I of the regression notes is simpler and probably more convincing.) Finally, the system (1.6) of two equations in the unknowns $\hat{\beta}_0$ and $\hat{\beta}_1$ has a solution!

To solve the system $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$, we note that the 2×2 matrix $\mathbf{X}^T\mathbf{X}$ has an inverse. We know this since rank($\mathbf{X}^T\mathbf{X}$) = rank($\mathbf{X}$) = 2 , and full rank square matrices have inverses. (Note: rank($\mathbf{X}$) = 2 because the two columns of the design matrix $\mathbf{X}$ are linearly independent.) The solution to the matrix equation (1.1) has the form,

(1.7)
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

where $\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \dfrac{1}{n\sum X_i^2 - \left(\sum X_i\right)^2}\begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix} = \dfrac{1}{n\sum\left(X_i - \bar{X}\right)^2}\begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix}$. After a fair amount of algebra, equation (1.6) returns the estimated intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ given in Part I of the regression notes:

(1.8)
$$\hat{\beta}_1 = \frac{\sum X_i\boldsymbol{Y}_i - \frac{1}{n}\left[\left(\sum X_i\right)\left(\sum \boldsymbol{Y}_i\right)\right]}{\sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2} = \frac{\sum\left(X_i - \bar{X}\right)\left(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}}\right)}{\sum\left(X_i - \bar{X}\right)^2}$$

(1.9)
$$\hat{\beta}_0 = \bar{\boldsymbol{Y}} - \hat{\beta}_1\bar{X}$$

Example:

Although you will have plenty of opportunities to use software to determine the best fitting regression line for a sample of bivariate data, it might be useful to go through the process once by hand for the "toy" data set:

x	1	2	4	5
y	8	4	6	2

Rather than using the equations (1.7) and (1.8), start with $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$ and find (a) $\mathbf{X}^T\mathbf{X}$, (b) $\mathbf{X}^T\mathbf{Y}$, (c) $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ (using only the definition of the inverse of a 2×2 matrix given in the linear algebra review), and (d) $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$. Perform the analysis again using software or a calculator to confirm the answer.

The (Least Squares) Solution to the Regression Model

The equation $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$ is the estimate to the simple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$ is the least squares estimate of the intercept and slope of the *true* regression line $y = \beta_0 + \beta_1 x$, and $\mathbf{e}$ is the $n\times 1$ vector of residuals. Now, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ can be written $\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e}$, where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is the $n\times 1$ vector of predictions made for the observations by the *estimated* regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The n points $\left(X_i, \hat{Y}_i\right)$ will, of course, all lie on the estimated regression line.

The Hat Matrix

A useful matrix that shows up again and again in regression analysis is the $n\times n$ matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X}$, called the "hat" matrix. To see how it gets its name, note that $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X}\mathbf{Y} = \mathbf{H}\mathbf{Y}$. Thus, the matrix $\mathbf{H}$ puts a "hat" on the vector $\mathbf{Y}$.

The hat matrix $\mathbf{H}$ has many nice properties. These include:

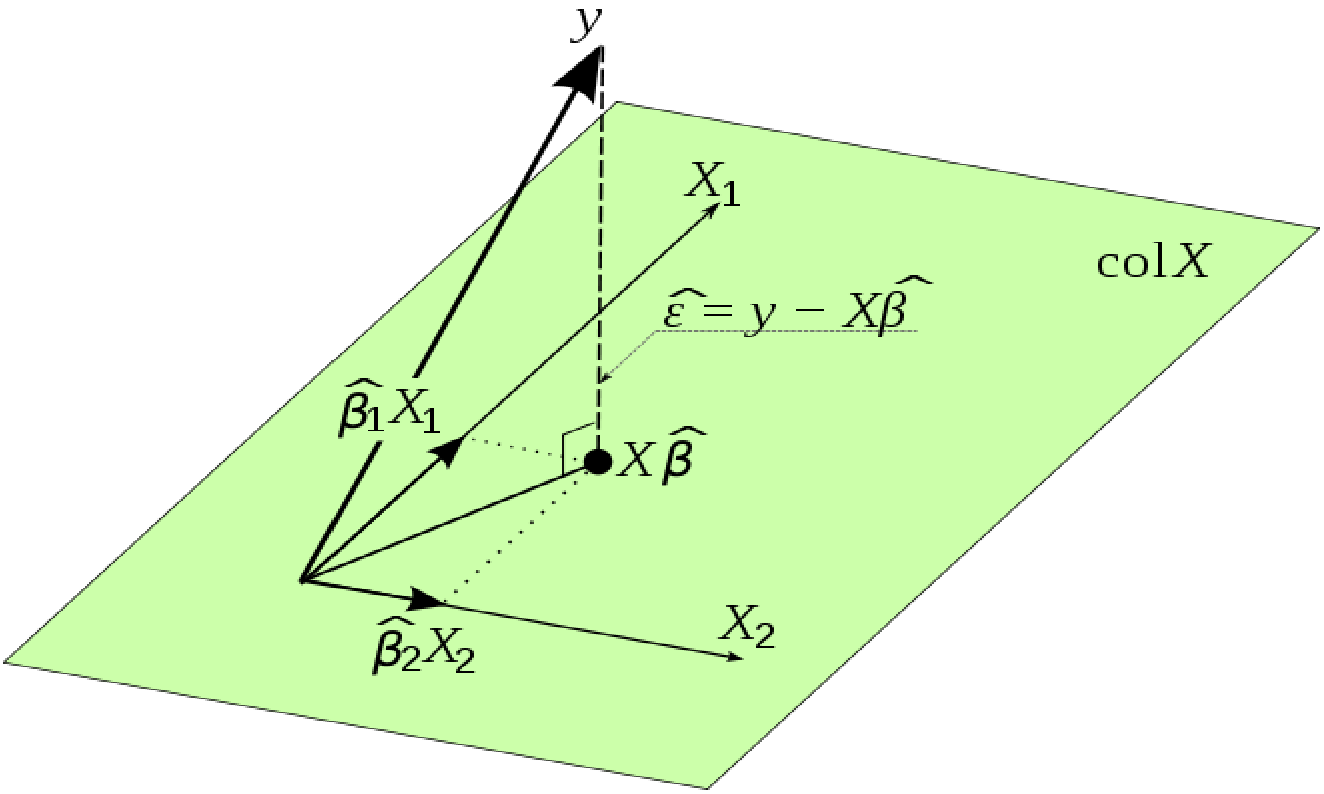
- $\mathbf{H}$ is symmetric, i.e., $\mathbf{H}^T = \mathbf{H}$, as is easily proven.
- $\mathbf{H}$ is idempotent, which is a fancy way of saying that $\mathbf{H}\mathbf{H} = \mathbf{H}$. The shorthand for this is $\mathbf{H}^2 = \mathbf{H}$. This is also easily proven.

- The matrix $(\mathbf{I} - \mathbf{H})$, where $\mathbf{I}$ is the $n \times n$ identity matrix, is both symmetric and idempotent as well.

The Error Sum of Squares, *SSE*, in Matrix Form

From the equation $\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e}$, we get $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. Using the hat matrix, this becomes $\mathbf{e} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, where we have right-factored the vector $\mathbf{Y}$. The error sum of squares, *SSE*, is just the squared length of the vector of residuals, i.e., $SSE = \sum e_i^2 = \mathbf{e}^T\mathbf{e}$. In terms of the hat matrix this becomes $SSE = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$, where we have used the symmetry and idempotency of the matrix $(\mathbf{I} - \mathbf{H})$ to simplify the result.

The Geometry of Least Squares



$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e} \tag{1.10}$$

One of the attractions of linear models, and especially of the least squares solutions to them, is the wealth of geometrical interpretations that spring from them. The $n \times 1$ vectors $\mathbf{Y}$, $\hat{\mathbf{Y}}$, and $\mathbf{e}$ are vectors in the vector space $\mathbb{R}^n$, (an *n*-dimensional space whose components are real numbers, as opposed to complex numbers). From equation (1.9) above, we know that the vectors $\hat{\mathbf{Y}}$ and $\mathbf{e}$ sum to $\mathbf{Y}$, but we can show a much more surprising result that will eventually lead to powerful conclusions about the sums of squares *SSE*, *SSR*, and *SST*.

First, we have to know where each of the vectors $\mathbf{Y}$, $\hat{\mathbf{Y}}$, and $\mathbf{e}$ "live":

- The *n*-vector of observations $\mathbf{Y}$ has no restrictions placed on it and therefore can lie anywhere in $\mathbb{R}^n$.

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\beta}_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ is restricted to the two-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of the design matrix $\mathbf{X}$, called (appropriately) the **column space** of $\mathbf{X}$.

- We will shortly show that $\mathbf{e}$ lives in a subspace of $\mathbb{R}^n$ **orthogonal** to the column space of $\mathbf{X}$.

Next, we derive the critical result that the vectors $\hat{\mathbf{Y}}$ and $\mathbf{e}$ are orthogonal (perpendicular) in $\mathbb{R}^n$. (Remember: vectors $\boldsymbol{v}$ and $\boldsymbol{u}$ are *orthogonal* if and only if $\boldsymbol{v}^T\boldsymbol{u} = 0$.)

- $\hat{\mathbf{Y}}^T\mathbf{e} = (\mathbf{HY})^T(\mathbf{I}-\mathbf{H})\mathbf{Y} = \mathbf{Y}^T\mathbf{H}^T(\mathbf{I}-\mathbf{H})\mathbf{Y} = \mathbf{Y}^T(\mathbf{H}-\mathbf{H})\mathbf{Y} = 0$

(Where we made repeated use of the fact that $\mathbf{H}$ is symmetric and idempotent.) Therefore, $\mathbf{e}$ is restricted to an *n* - 2 dimensional subspace of $\mathbb{R}^n$ (because it must be orthogonal to every vector in the column space of $\mathbf{X}$).

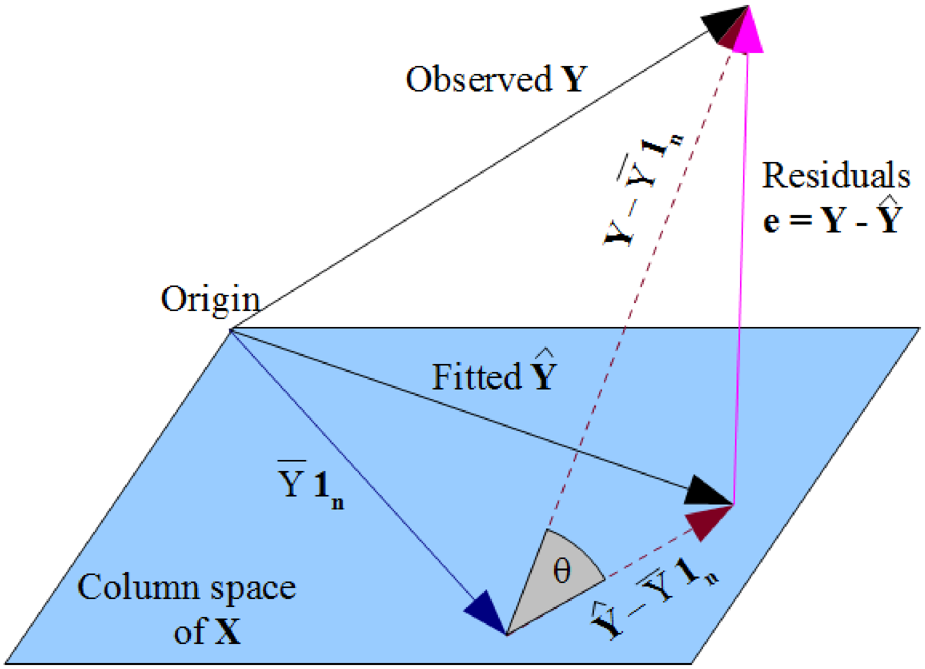
Combining the facts that the vectors $\hat{\mathbf{Y}}$ and $\mathbf{e}$ sum to $\mathbf{Y}$ *and* are orthogonal to each other, we conclude that $\hat{\mathbf{Y}}$ and $\mathbf{e}$ form the legs of a right triangle (in $\mathbb{R}^n$) with hypotenuse $\mathbf{Y}$. By the Pythagorean Theorem for right triangles, $\left\|\hat{\mathbf{Y}}\right\|^2 + \|\mathbf{e}\|^2 = \|\mathbf{Y}\|^2$ or equivalently,

$$\hat{\mathbf{Y}}^T\hat{\mathbf{Y}} + \mathbf{e}^T\mathbf{e} = \mathbf{Y}^T\mathbf{Y} \tag{1.11}$$

A slight modification of the argument above shows that the vectors $\hat{\mathbf{Y}} - \overline{\mathbf{Y}}$ and $\mathbf{e}$ sum to $\mathbf{Y} - \overline{\mathbf{Y}}$ *and* are orthogonal to each other, hence we conclude that $\hat{\mathbf{Y}} - \overline{\mathbf{Y}}$ and $\mathbf{e}$ form the legs of a right triangle (in $\mathbb{R}^n$) with hypotenuse $\mathbf{Y} - \overline{\mathbf{Y}}$. By the Pythagorean Theorem for right triangles,

$$\left(\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\right)^{\mathbf{T}}\left(\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\right) + \mathbf{e}^{\mathbf{T}}\mathbf{e} = \left(\mathbf{Y} - \overline{\mathbf{Y}}\right)^{\mathbf{T}}\left(\mathbf{Y} - \overline{\mathbf{Y}}\right) \tag{1.12}$$

or, equivalently, $\boxed{SSR + SSE = SST}$. This last equality is the famous one involving the three sums of squares in regression. (See picture below.)


Observed $\mathbf{Y}$
$\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1}_n$
Residuals
$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$
Origin
Fitted $\hat{\mathbf{Y}}$
$\bar{\mathrm{Y}}\mathbf{1}_n$
θ
$\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\mathbf{1}_n$
Column space of $\mathbf{X}$

We've actually done more than just derive the equation involving the three sums of squares. It turns out that the dimensions of the subspaces the vectors "live" in also determines their "degrees of freedom," so we've also shown that *SSE* has *n* - 2 degrees of freedom because the vector of residuals $\mathbf{e}$ is restricted to an *n* - 2 dimensional subspace of $\mathbb{R}^n$. (The term "degrees of freedom" is actually quite descriptive because the vector $\mathbf{e}$ is only "free" to assume values in this *n* - 2 dimensional subspace.)

The Analysis of Variance (ANOVA) Table

The computer output of a regression analysis always contains a table containing the sums of squares *SSR*, *SSE*, and *SST*. The table is called an analysis of variance, or ANOVA, table for reasons we will see later in the course. The table has the form displayed below. (Note: The mean square of the residuals is the "mean squared error" *MSE*, i.e., the estimate of the variance σ^2 of the error variable ε in the regression model.)

Source	Degrees of Freedom (df)	Sums of Squares (SS)	Mean Square (MS) = SS/df
Regression Model	1	$SSR = \sum\left(\hat{Y}_i - \bar{Y}\right)^2 = \left(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\right)^{\mathbf{T}}\left(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\right)$	$MSR = \sum\left(\hat{Y}_i - \bar{Y}\right)^2$
Residual Error	*n* - 2	$SSE = \sum\left(Y_i - \hat{Y}_i\right)^2 = \mathbf{e}^{\mathbf{T}}\mathbf{e}$	$MSE = \dfrac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2}$
Total	*n* - 1	$SST = \sum\left(Y_i - \bar{Y}\right)^2 = \left(\mathbf{Y} - \bar{\mathbf{Y}}\right)^{\mathbf{T}}\left(\mathbf{Y} - \bar{\mathbf{Y}}\right)$	