The Geometry of Least Squares

Mathematical Basics

▶ Inner / dot product: a and b column vectors

$$a \cdot b = a^T b = \sum a_i b_i$$

$$a \perp b \Leftrightarrow a^T b = 0$$

▶ Matrix Product: A is $r \times s$ B is $s \times t$

$$(AB)_{rt} = \sum_{s} A_{rs} B_{st}$$





Partitioned Matrices

- Partitioned matrices are like ordinary matrices but the entries are matrices themselves.
- ► They add and multiply (if the dimensions match properly) just like regular matrices but(!) you must remember that matrix multiplication is **not** commutative.
- Here is an example

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{bmatrix}$$

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \\ B_{31} & B_{32} \end{bmatrix}$$





- ▶ Think of A as a 2×3 matrix and B as a 3×2 matrix.
- ▶ multiply them to get C = AB a 2 × 2 matrix as follows:

$$AB = \left[\begin{array}{c|c|c} A_{11}B_{11} + A_{12}B_{21} + A_{13}B_{31} & A_{11}B_{12} + A_{12}B_{22} + A_{13}B_{32} \\ \hline A_{21}B_{11} + A_{22}B_{21} + A_{23}B_{31} & A_{21}B_{12} + A_{22}B_{22} + A_{23}B_{32} \end{array} \right]$$

- ▶ BUT: this only works if each of the matrix products in the formulas makes sense.
- ▶ So, A_{11} must have the same number of columns as B_{11} has rows and many other similar restrictions apply.





First application:

$$X = [X_1 | X_2 | \cdots | X_p]$$

where each X_i is a column of X. Then

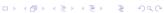
$$X\beta = [X_1|X_2|\cdots|X_p] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p$$

which is a linear combination of the columns of X.

Definition: The column space of X, written $\operatorname{col}(X)$ is the (vector space of) set of all linear combinations of columns of X also called the space "spanned" by the columns of X.

SO:
$$\hat{\mu} = X\beta$$
 is in $col(X)$.





Back to normal equations:

$$X^TY = X^TX\hat{\beta}$$

or

$$X^{T}\left[Y-X\hat{\beta}\right]=0$$

or

$$\begin{bmatrix} X_1^T \\ \vdots \\ X_p^T \end{bmatrix} \begin{bmatrix} Y - X\hat{\beta} \end{bmatrix} = 0$$

or

$$X_i^T \left[Y - X \hat{\beta} \right] = 0$$
 $i = 1, \dots, p$

or

$$Y - X\hat{\beta} \perp$$
 every vector in $col(X)$





Definition: $\hat{\epsilon} = Y - X\hat{\beta}$ is the fitted residual vector.

SO: $\hat{\epsilon} \perp \operatorname{col}(X)$ and $\hat{\epsilon} \perp \hat{\mu}$

Pythagoras' Theorem: If $a \perp b$ then

$$||a||^2 + ||b||^2 = ||a + b||^2$$

Definition: ||a|| is the "length" or "norm" of a:

$$||a|| = \sqrt{\sum a_i^2} = \sqrt{a^T a}$$

Moreover, if a, b, c, \ldots are all perpendicular then

$$||a||^2 + ||b||^2 + \dots = ||a+b+\dots||^2$$





Application

$$Y = Y - X\hat{\beta} + X\hat{\beta}$$
$$= \hat{\epsilon} + \hat{\mu}$$

SO

$$||Y||^2 = ||\hat{\epsilon}||^2 + ||\hat{\mu}||^2$$

or

$$\sum Y_i^2 = \sum \hat{\epsilon}_i^2 + \sum \hat{\mu}_i^2$$

Definitions:

$$\sum Y_i^2 = \text{Total Sum of Squares (unadjusted)}$$

$$\sum \hat{\epsilon}_i^2 = \text{Error or Residual Sum of Squares}$$

$$\sum \hat{\mu}_i^2 = \text{Regression Sum of Squares}$$





Alternative formulas for the Regression SS

$$\sum \hat{\mu}_i^2 = \hat{\mu}^T \hat{\mu}$$
$$= (X\hat{\beta})^T (X\hat{\beta})$$
$$= \hat{\beta}^T X^T X \hat{\beta}$$

Notice the matrix identity which I will use regularly:

$$(AB)^T = B^T A^T.$$





What is least squares?

Choose $\hat{\beta}$ to minimize

$$\sum (Y_i - \hat{\mu}_i)^2 = ||Y - \hat{\mu}||^2$$

That is, to minimize $||\hat{\epsilon}||^2$. The resulting $\hat{\mu}$ is called the **Orthogonal Projection** of Y onto the column space of X. **Extension**:

$$X = [X_1|X_2]$$
 $\beta = \left[\frac{\beta_1}{\beta_2}\right]$ $p = p_1 + p_2$

Imagine we fit 2 models:

1. The FULL model:

$$Y = X\beta + \epsilon (= X_1\beta_1 + X_2\beta_2 + \epsilon)$$

2. The REDUCED model:

$$Y = X_1 \beta_1 + \epsilon$$



If we fit the full model we get

$$\hat{\beta}_F \quad \hat{\mu}_F \quad \hat{\epsilon}_F \qquad \hat{\epsilon}_F \perp \operatorname{col}(X)$$
 (1)

If we fit the reduced model we get

$$\hat{\beta}_R \quad \hat{\mu}_R \quad \hat{\epsilon}_R \qquad \hat{\mu}_R \in \operatorname{col}(X_1) \subset \operatorname{col}(X)$$
 (2)

Notice that

$$\hat{\epsilon}_F \perp \hat{\mu}_R$$
. (3)

(The vector $\hat{\mu}_R$ is in the column space of X_1 so it is in the column space of X and $\hat{\epsilon}_F$ is orthogonal to **everything** in the column space of X.) So:

$$Y = \hat{\epsilon}_F + \hat{\mu}_F$$

= $\hat{\epsilon}_F + \hat{\mu}_R + (\hat{\mu}_F - \hat{\mu}_R) = \epsilon_R + \hat{\mu}_R$





You know $\hat{\epsilon}_F \perp \hat{\mu}_R$ (from (3) above) and $\hat{\epsilon}_F \perp \hat{\mu}_F$ (from (1) above). So

$$\hat{\epsilon}_F \perp \hat{\mu}_F - \hat{\mu}_R$$

Also

$$\hat{\mu}_R \perp \hat{\epsilon}_R = \hat{\epsilon}_F + (\hat{\mu}_F - \hat{\mu}_R)$$

So

$$0 = (\hat{\epsilon}_F + \hat{\mu}_F - \hat{\mu}_R)^T \hat{\mu}_R$$
$$= \underbrace{\hat{\epsilon}_F^T \hat{\mu}_R}_{0} + (\hat{\mu}_F - \hat{\mu}_R)^T \hat{\mu}_R$$

SO

$$\hat{\mu}_F - \hat{\mu}_R \perp \hat{\mu}_R$$





Summary

We have

$$Y = \hat{\mu}_R + (\hat{\mu}_F - \hat{\mu}_R) + \hat{\epsilon}_F$$

All three vectors on the Right Hand Side are perpendicular to each other.

This gives:

$$||Y||^2 = ||\hat{\mu}_R||^2 + ||\hat{\mu}_F - \hat{\mu}_R||^2 + ||\hat{\epsilon}_F||^2$$

which is an Analysis of Variance (ANOVA) table!





Here is the most basic version of the above:

$$X = [\mathbf{1}|X_1]$$
 $Y_i = \beta_0 + \cdots + \epsilon_i$

The notation here is that

$$\mathbf{1} = \left[egin{array}{c} 1 \ dots \ 1 \end{array}
ight]$$

is a column vector with all entries equal to 1. The coefficient of this column, β_0 , is called the "intercept" term in the model.





To find $\hat{\mu}_R$ we minimize

$$\sum (Y_i - \hat{\beta}_0)^2$$

and get simply

$$\hat{\beta}_0 = \bar{Y}$$

and

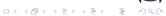
$$\hat{\mu}_R = \left[egin{array}{c} ar{Y} \ dots \ ar{Y} \end{array}
ight]$$

Our ANOVA identity is now

$$||Y||^2 = ||\hat{\mu}_R||^2 + ||\hat{\mu}_F - \hat{\mu}_R||^2 + ||\hat{\epsilon}_F||^2$$

= $n\bar{Y}^2 + ||\hat{\mu}_F - \hat{\mu}_R||^2 + ||\hat{\epsilon}_F||^2$





This identity is usually rewritten in subtracted form:

$$||Y||^2 - n\bar{Y}^2 = ||\hat{\mu}_F - \hat{\mu}_R||^2 + ||\hat{\epsilon}_F||^2$$

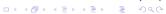
Remembering the identity $\sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n \bar{Y}^2$ we find

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{\mu}_{F,i} - \bar{Y})^2 + \sum \hat{\epsilon}_{F,i}^2$$

These terms are respectively:

- the Adjusted or Corrected Total Sum of Squares,
- ▶ the Regression or Model Sum of Squares and
- the Error Sum of Squares.





Simple Linear Regression

- Filled Gas tank 107 times.
- Record distance since last fill, gas needed to fill.
- Question for discussion: natural model?
- Look at JMP analysis.





The sum of squares decomposition in one example

- Example discussed in *Introduction*.
- Consider model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

with
$$\alpha_4 = -(\alpha_1 + \alpha_2 + \alpha_3)$$
.

- ▶ Data consist of blood coagulation times for 24 animals fed one of 4 different diets.
- Now I write the data in a table and decompose the table into a sum of several tables.
- ► The 4 columns of the table correspond to Diets A, B, C and D.
- You should think of the entries in each table as being stacked up into a column vector, but the tables save space.





- ► The design matrix can be partitioned into a column of 1s and 3 other columns.
- \triangleright You should compute the product X^TX and get

$$\begin{bmatrix} 24 & -4 & -2 & -2 \\ -4 & 12 & 8 & 8 \\ -2 & 8 & 14 & 8 \\ -2 & 8 & 8 & 14 \end{bmatrix}$$

▶ The matrix X^TY is just

$$\left[\sum_{ij} Y_{ij}, \sum_{j} Y_{1j} - \sum_{j} Y_{4j}, \sum_{j} Y_{2j} - \sum_{j} Y_{4j}, \sum_{j} Y_{3j} - \sum_{j} Y_{4j}\right]$$





- ▶ The matrix X^TX can be inverted using a program like Maple.
- ▶ I found that

$$384(X^TX)^{-1} = \begin{bmatrix} 17 & 7 & -1 & -1 \\ 7 & 65 & -23 & -23 \\ -1 & -23 & 49 & -15 \\ -1 & -23 & -15 & 49 \end{bmatrix}$$

▶ It now takes quite a bit of algebra to verify that the vector of fitted values can be computed by simply averaging the data in each column.





That is, the fitted value, $\hat{\mu}$ is the table





On the other hand fitting the model with a design matrix consisting only of a column of 1s just leads to $\hat{\mu}_R$ (notation from the lecture) given by





Earlier I gave identity:

$$Y = \hat{\mu}_R + (\hat{\mu}_F - \hat{\mu}_R) + \hat{\epsilon}_F$$

which corresponds to the following identity:





Pythagoras identity: ANOVA

- ► The sums of squares of the entries of each of these arrays are as follows.
- ▶ Uncorrected total sum of squares: On the left hand side $62^2 + 63^2 + \cdots = 98644$.
- ▶ The first term on the right hand side gives $24(64^2) = 98304$.
- This term is sometimes put in ANOVA tables as the Sum of Squares due to the Grand Mean.
- But it is usually subtracted from the total to produce the Total Sum of Squares which we usually put at the bottom of the table
- This is often called the Corrected (or Adjusted) Total Sum of Squares.



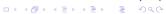


In this case the corrected sum of squares is the squared length of the table

$$\begin{bmatrix} -2 & -1 & 4 & -8 \\ -4 & 3 & 2 & -2 \\ -1 & 7 & 7 & -4 \\ -5 & 0 & 3 & -3 \\ & 1 & 4 & -1 \\ & 2 & 4 & 0 \\ & & & -1 \\ & & & -5 \end{bmatrix}$$

which is 340.





- ► Treatment Sum of Squares: The second term on the right hand side of the equation has squared length $4(-3)^2 + 6(2)^2 + 6(4)^2 + 8(-3)^2 = 228$.
- ▶ The formula for this Sum of Squares is

$$\sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^{I} n_i (\bar{X}_{i.} - \bar{X}_{..})^2$$

- ▶ but I want you to see that the formula is just the squared length of the vector of individual sample means minus the grand mean.
- ► The last vector of the decomposition is called the residual vector.
- ▶ It has squared length $1^2 + (-3)^2 + 0^2 + \cdots = 112$.





Degrees of freedom: dimensions of spaces

- ► Corresponding to the decomposition of the total squared length of the data vector is a decomposition of its dimension, 24, into the dimensions of subspaces.
- ► For instance the grand mean is always a multiple of the single vector all of whose entries are 1;
- this describes a one dimensional space
- ▶ this is just another way of saying that the reduced $\hat{\mu}_R$ is in the column space of the reduced model design matrix.
- ▶ The second vector, of deviations from a grand mean lies in the three dimensional subspace of tables which are constant in each column and have a total equal to 0.
- Similarly the vector of residuals lies in a 20 dimensional subspace – the set of all tables whose columns sum to 0.





Degrees of Freedom

- ➤ This decomposition of dimensions is the decomposition of degrees of freedom.
- ▶ So 24 = 1 + 3 + 20 and the degrees of freedom for treatment and error are 3 and 20 respectively.
- ▶ The vector whose squared length is the Corrected Total Sum of Squares lies in the 23 dimensional subspace of vectors whose entries sum to 1.
- ► This produces the 23 total degrees of freedom in the usual ANOVA table.



