Scientific Computing for Biologists

Multiple Regression

Paul M. Magwene

Variable space view of multiple regression

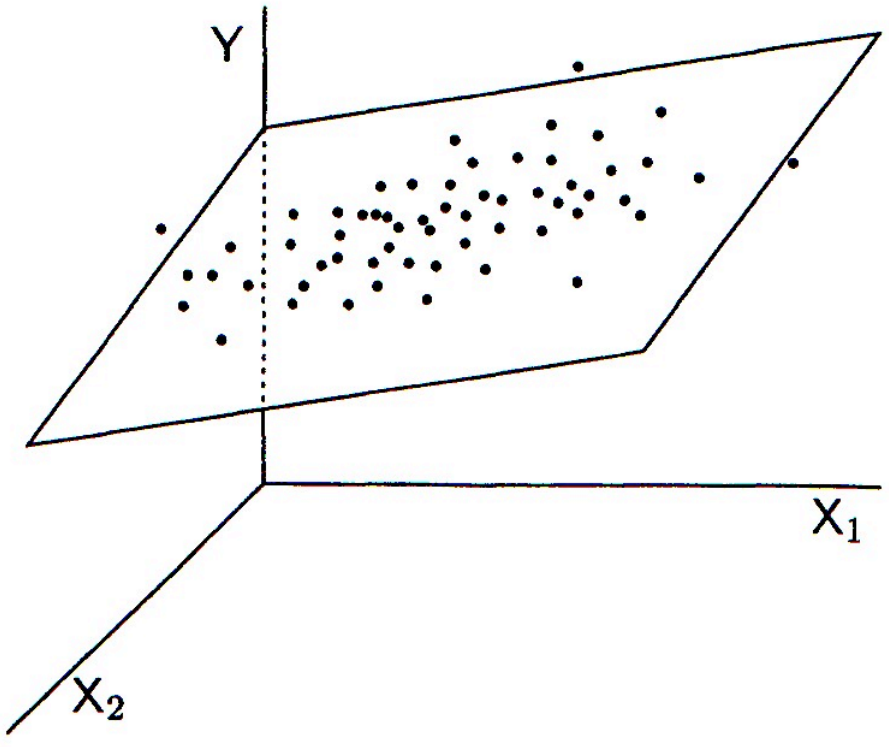
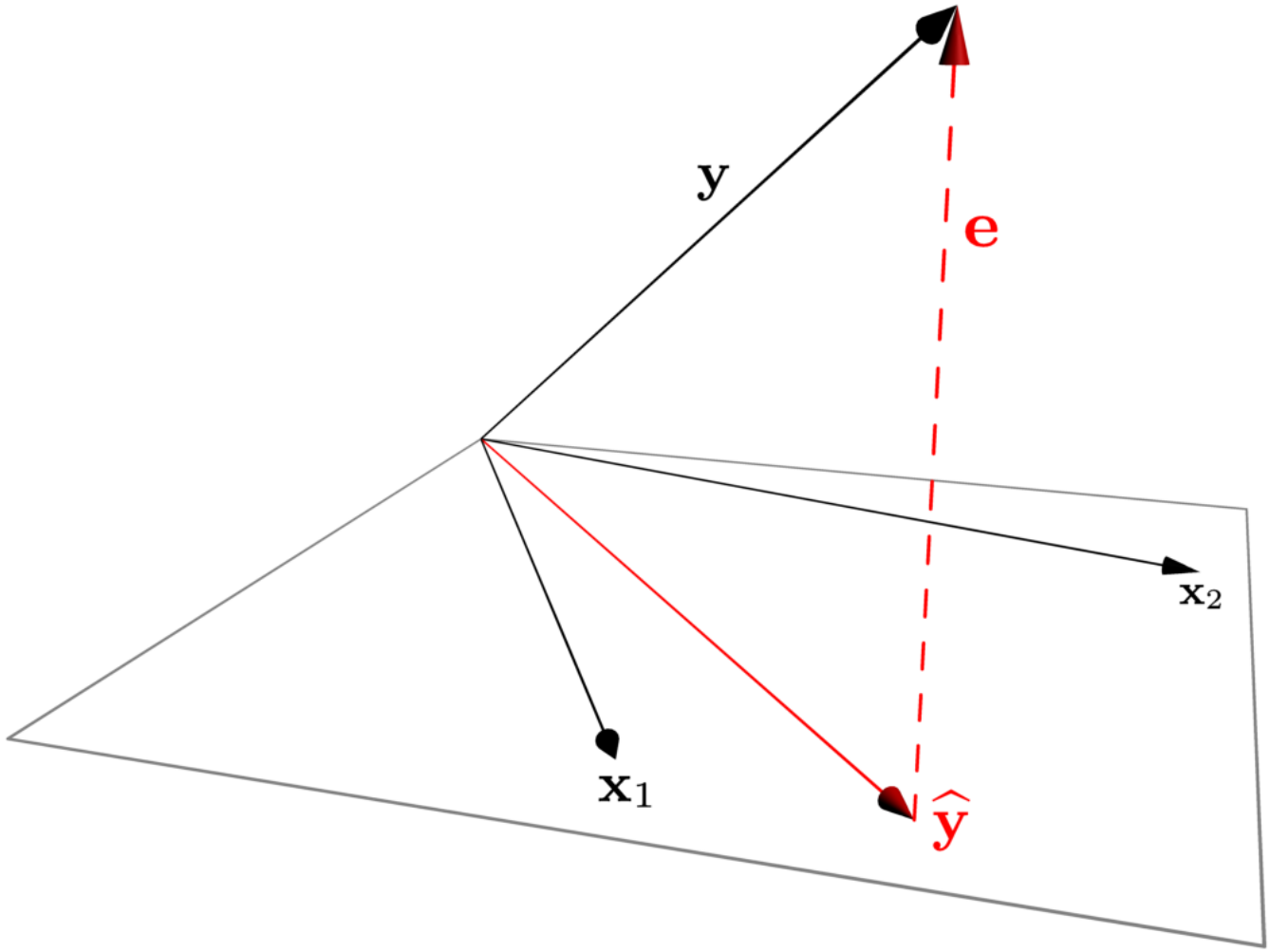


Figure 4.1: *The regression of* Y *onto* X_1 *and* X_2 *as a scatterplot in variable space.*

Subject Space Geometry of Multiple Regression



$$\vec{y} = \hat{\vec{y}} + \vec{e}$$

Multiple Regression

$$\hat{\vec{y}} = a\mathbf{1} + b_1\boldsymbol{X}_1 + b_2\boldsymbol{X}_2 + \cdots + b_p\boldsymbol{X}_p$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = a \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + b_1 \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} + b_2 \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} + \cdots + b_p \begin{bmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{bmatrix}$$

$\hat{y}$ is linear combination of the predictor variables

Recall that a *linear combination* of vectors is an equation of the form:

$$z = b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \cdots + b_p\boldsymbol{x}_p$$

In regression, the vector of fitted values, $\hat{\vec{y}}$, is a linear combination of the predictor variables.

Matrix Representation of Multiple Regression

Let $\vec{y}$ be a vector of values for the outcome variable. Let X_i be explanatory variables. In matrix form:

$$\vec{y} = X\vec{b} + \vec{e}$$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix};$$

$$b = \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}; e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Estimating the Coefficients for Multiple Regression

$$\vec{y} = X\vec{b} + \vec{e}$$

Estimate $\vec{b}$ as:

$$\vec{b} = (X^T X)^{-1} X^T \vec{y}$$

Matrix Inverses

- If A is a *square matrix* and C is a matrix of the same size where $AC = I$ and $CA = I$ than C is the inverse of A and we denote is A^{-1}.

$$AA^{-1} = A^{-1}A = I$$

- Rules for inverses:
  - Only square matrices are invertible
  - A matrix for which we can find an inverse is called ***invertible*** (non-singular)
  - A matrix for which no inverse exists is ***singular*** (non-invertible)
  - If A and B are both invertible $p \times p$ matrices than $(AB)^{-1} = B^{-1}A^{-1}$ (note change in order).

More facts about Matrix Inverses

- Not every square matrix is invertible
- Every orthogonal matrix is invertible
- Any diagonal matrix, A, where the a_{ii} are non-zero, is invertible

Regression coefficients

$$\vec{\boldsymbol{b}} = \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}$$

- The regression coefficients can be thought of as "weightings" of the predictor variables
- Comparing the magnitude of regression coefficients only makes sense if all the predictor variables have the same scale
- If predictor variables have different scales, then can can first standardize the variables (subtract means and divide by variances), before fitting regression to get standardized regression coefficients.
- Alternately, calculate standardized regression coefficients as:

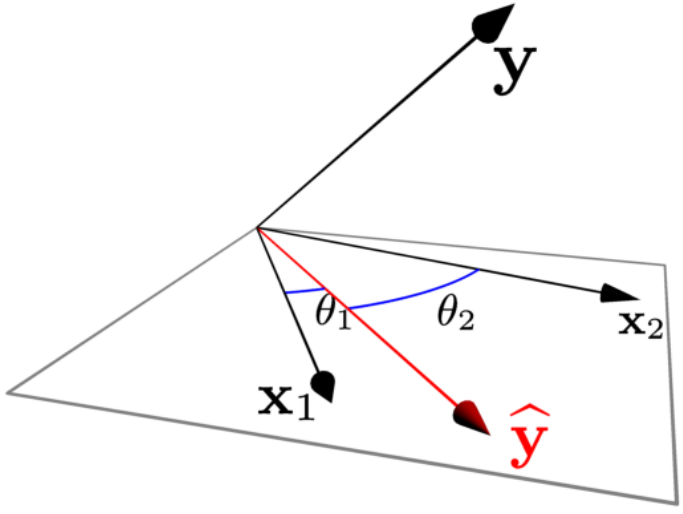
$$b'_j = \frac{|\vec{\boldsymbol{x}}_j|}{|\vec{\boldsymbol{y}}|} b_j$$

Regression loadings

The correlation between each predictor variable, x_j and the prediction, $\hat{\vec{y}}$, is given by the angle between them:

$$\cos\theta_{\overrightarrow{x_j},\overrightarrow{\hat{y}}} = \frac{\overrightarrow{x_j} \cdot \overrightarrow{\hat{y}}}{|\overrightarrow{x_j}||\overrightarrow{\hat{y}}|}$$

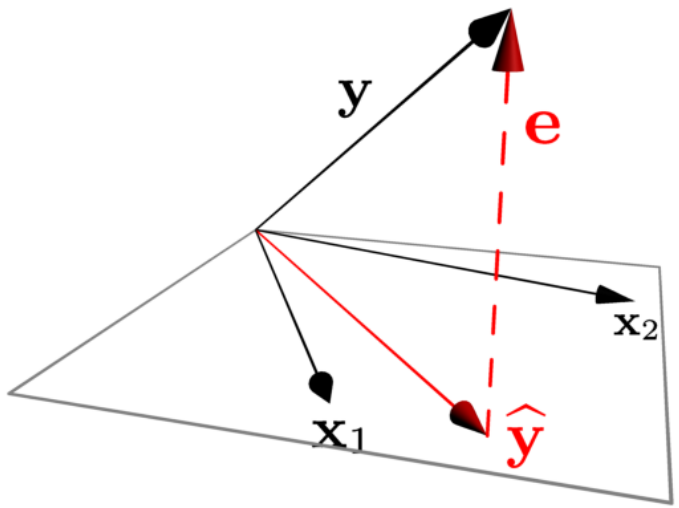
These are sometimes called "loadings" and should be examined in combination with the regression coefficients to understand the model implied by the multiple regression.



Space Spanned by a List of Vectors

Definition

Let X be a finite list of n-vectors. The **space spanned** by X is the set of all vectors that can be written as linear combinations of the vectors in X.



The prediction, $\hat{\vec{y}}$ is the closest vector to $\vec{y}$ in the space spanned by the predictor variables, X.

The predictor variables must be linearly independent

To solve the multiple regression equation, the predictor variables, $\boldsymbol{X}$, must be linearly independent.

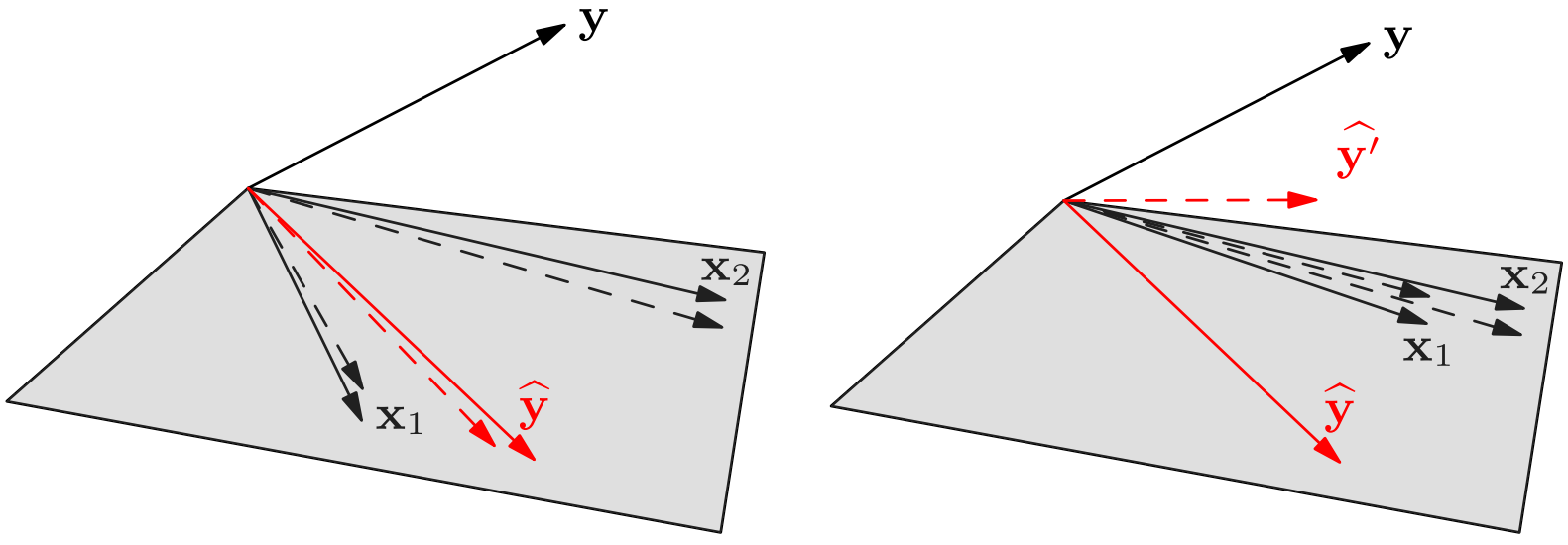
- A list of vectors, $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_p$, is said the be ***linearly dependent*** if there is a non-trivial combination of them which is equal to the zero vector.

$$b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \cdots + b_p\boldsymbol{x}_p = 0$$

- When a set of vectors are linearly dependent we also called them ***multicollinear***
- A list of vectors that are *not* linearly dependent are said to be ***linearly independent***
- If a matrix is invertible than it's columns form a linearly independent list of vectors!

Near multicollinearity of the predictors leads to unstable regression solutions

Predictor variables that are **nearly multicollinear** are, perhaps, even more difficult to deal with:



(a) Non-collinear predictors (b) Nearly collinear predictors

Figure: When predictors are nearly collinear, small differences in the vectors can result in large differences in the estimated regression.

What can I do if my predictors are (nearly) collinear?

- Drop some of the linearly dependent sets of predictors.
- Replace the linearly dependent predictors with a combined variable.
- Define orthogonal predictors, via linear combinations of the original variables (PC regression approach)
- 'Tweak' the predictor variables so that they're no longer multicollinear (Ridge regression).

Curvilinear Regression

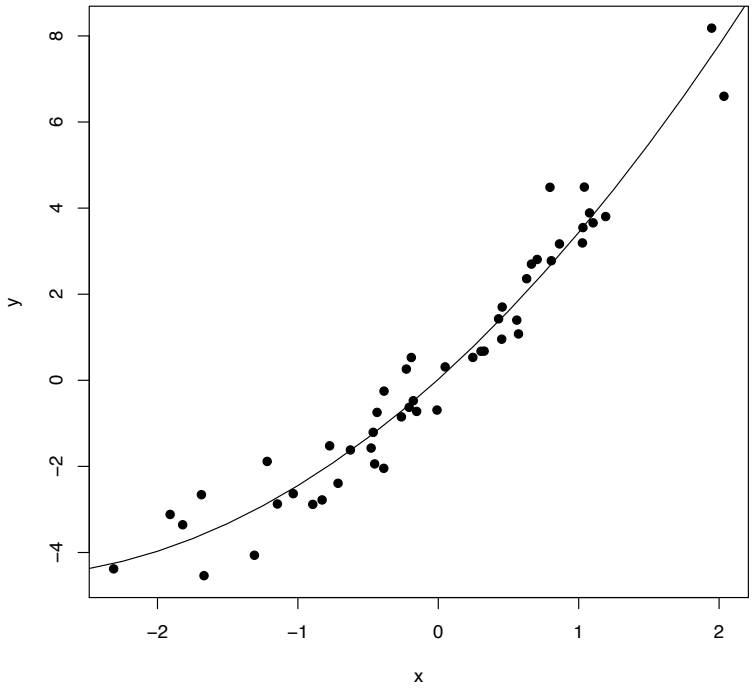
Curvilinear regression using **polynomial models** is simply multiple regression with the x_i replace by powers of x.

$$\hat{y} = b_1 \boldsymbol{x} + b_2 \boldsymbol{x}^2 + \cdots + b_p \boldsymbol{x}^n$$

Note:

- this is still a *linear* regression (linear in the coefficients)
- best applied when a specific hypothesis justifies their use
- generally not higher than quadratic or cubic

Example of Curvilinear Regression



$$y = 3x + 0.5x^2 + e$$

```
lm(formula = y ~ x + I(x^2))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02229    0.11651   0.191    0.849
x            2.94001    0.09693  30.331  < 2e-16 ***
I(x^2)       0.47146    0.07685   6.135 1.68e-07 ***
```