Multiple Regression

Paul M. Magwene

Variable space view of multiple regression

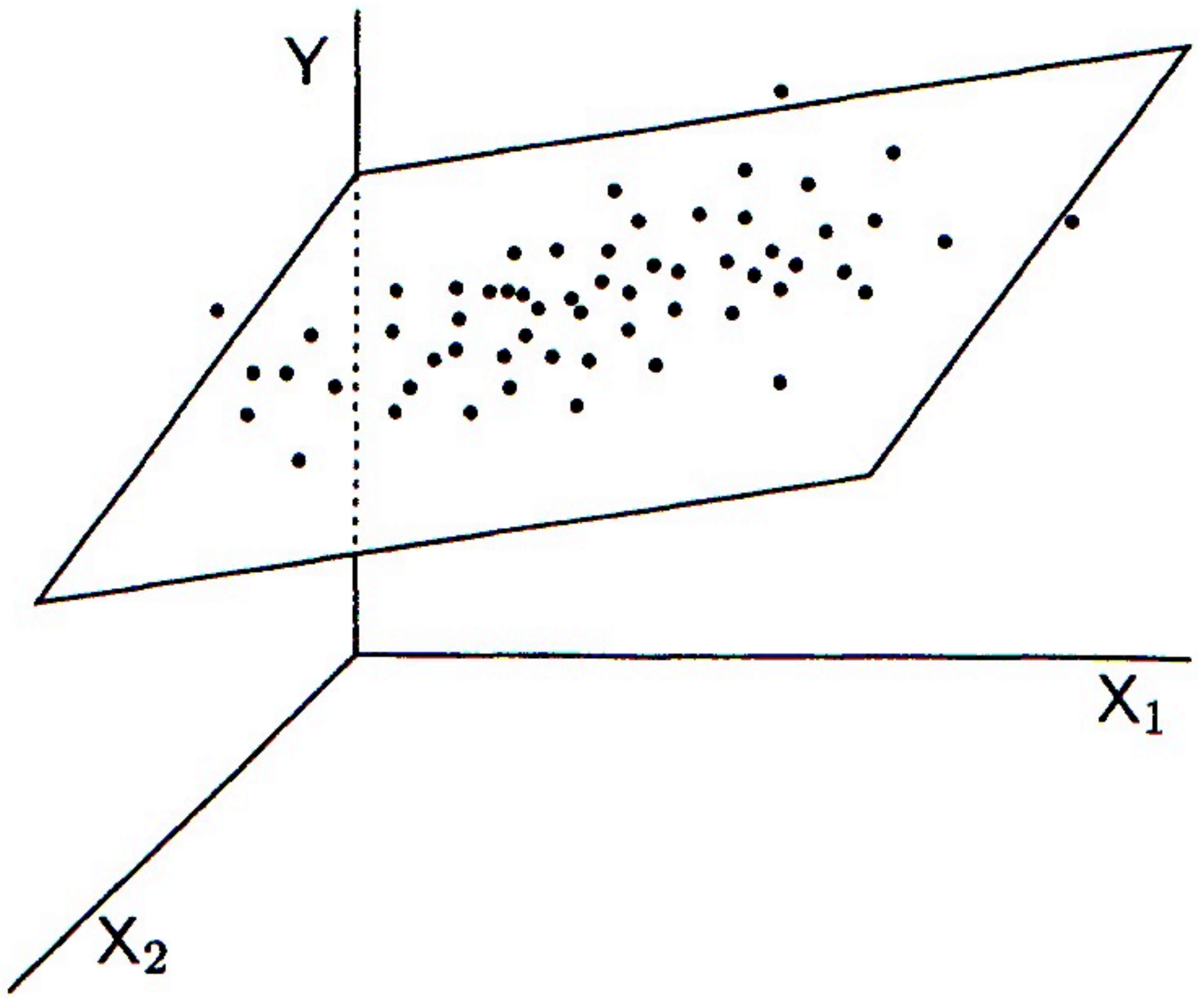
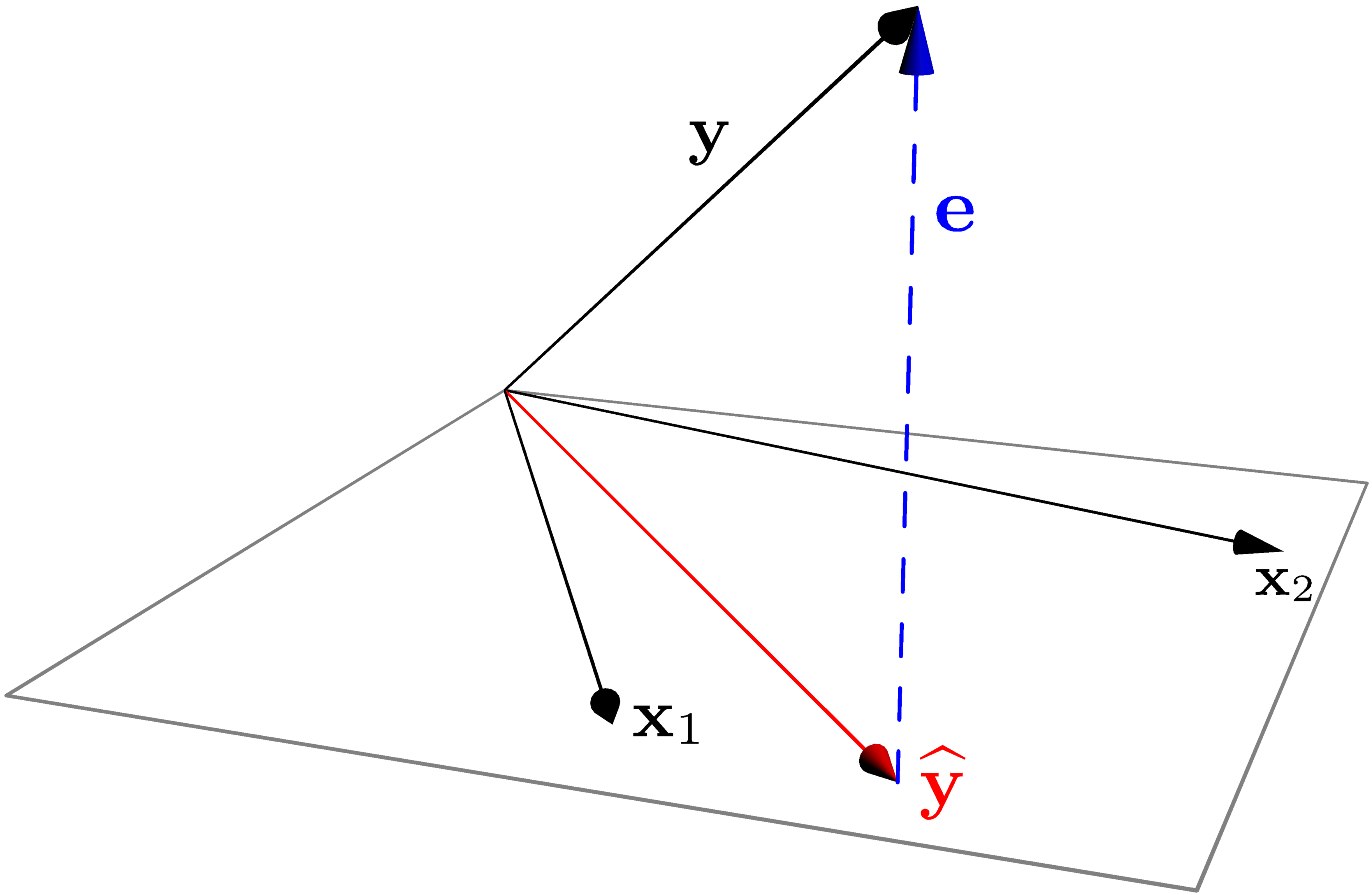


Figure 4.1: *The regression of* Y *onto* X_1 *and* X_2 *as a scatterplot in variable space.*

Subject Space Geometry of Multiple Regression



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

Matrix Representation of Multiple Regression

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

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

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

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

Multiple Regression

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

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

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

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

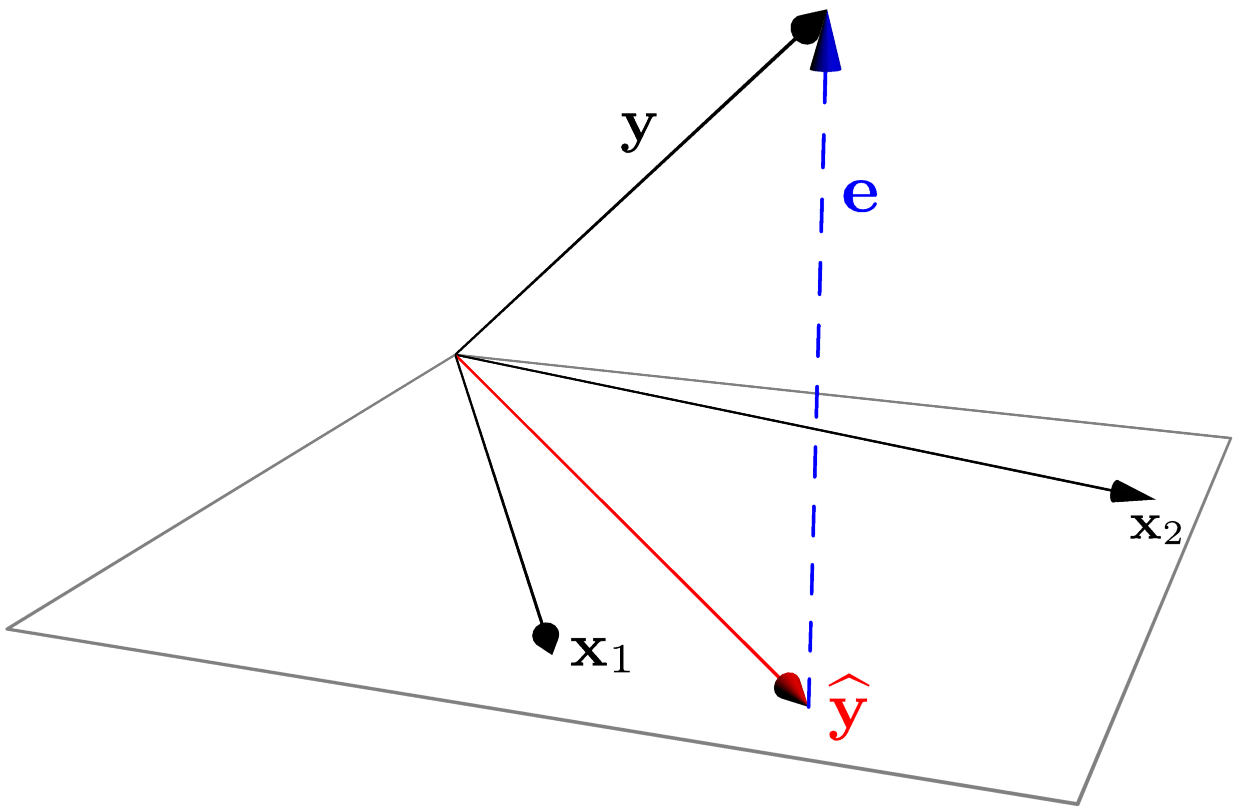
$$z = b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \cdots + b_p\boldsymbol{x}_p$$

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

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

Estimating the Coefficients for Multiple Regression

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

Estimate $\vec{\boldsymbol{b}}$ as:

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

Regression coefficients

$$\vec{\boldsymbol{b}} = \begin{bmatrix} a & b_1 & b_2 & \cdots & b_p \end{bmatrix}^T$$

- 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:

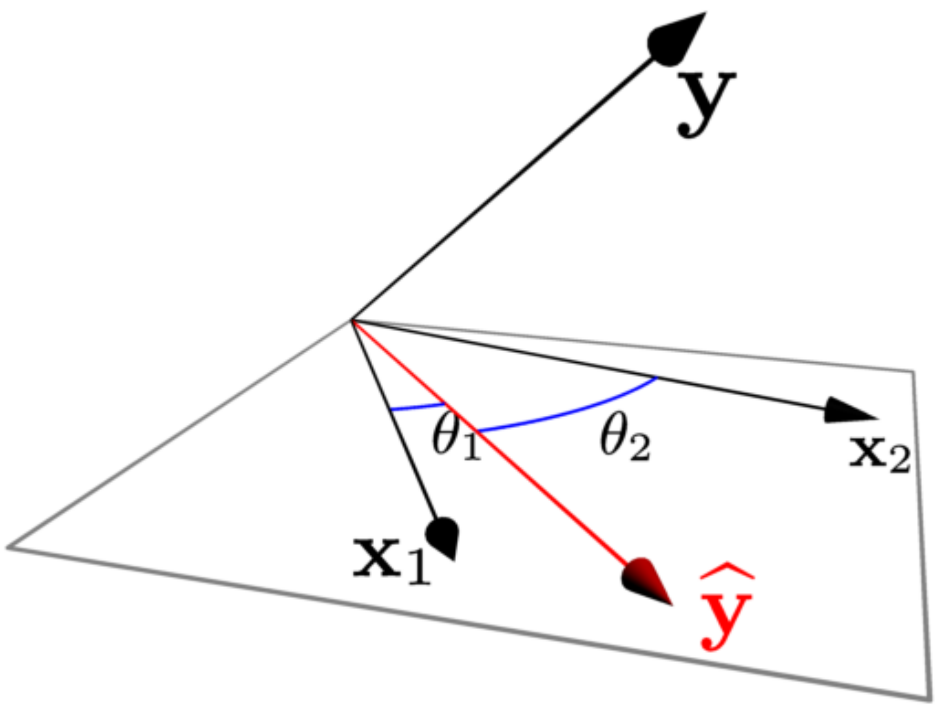
$$b'_j = \frac{|\vec{\boldsymbol{x}}_j|}{|\vec{\boldsymbol{y}}|} b_j$$

Regression loadings

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

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

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

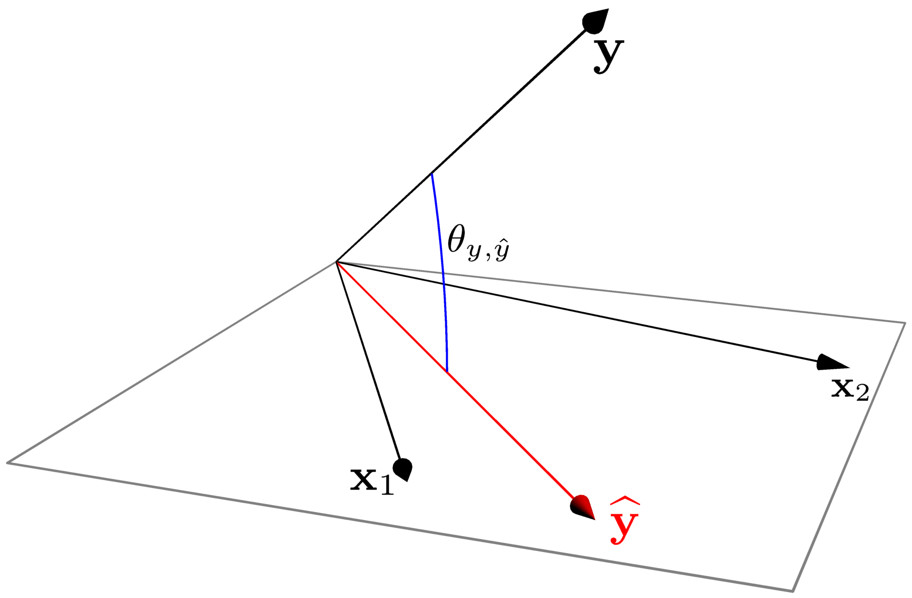


Coefficient of determination

The coefficient of determination, R^2, is the standard measure of the fraction of variation 'explained' by the regression model.
This is equivalent to the cosine of angle between the outcome vector and the predicted outcome vector, squared.

$$R^2 = (\cos\theta_{\vec{\boldsymbol{y}},\vec{\hat{\boldsymbol{y}}}})^2 = \frac{|\vec{\hat{\boldsymbol{y}}}|^2}{|\vec{\boldsymbol{y}}|^2} = \frac{\vec{\hat{\boldsymbol{y}}}\cdot\vec{\hat{\boldsymbol{y}}}}{\vec{\boldsymbol{y}}\cdot\vec{\boldsymbol{y}}}$$

Note that the vector formulation above assumes mean-centered $\vec{\boldsymbol{y}}$ and $\vec{\hat{\boldsymbol{y}}}$.



F-statistic

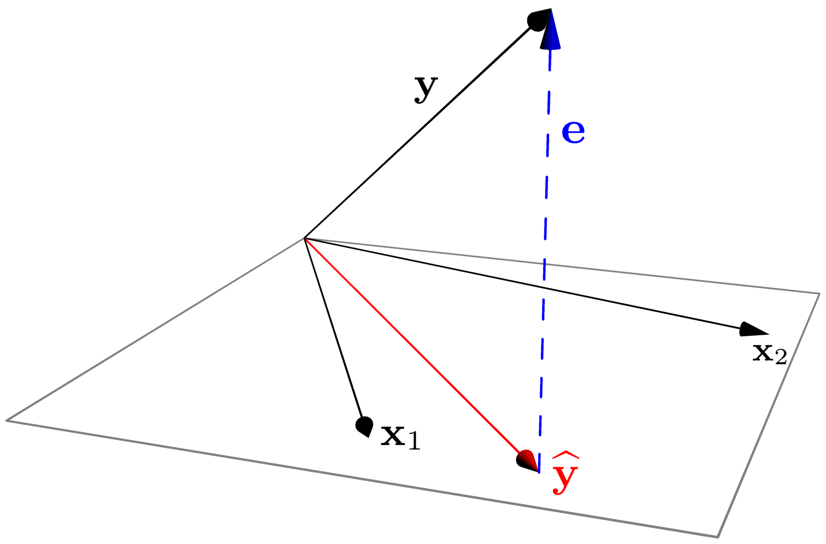
To compare the squared length of $|\vec{\hat{y}}|^2$ and $|\vec{e}|^2$ we divide them by the dimension of the subspaces in which they lie.

$$M(\vec{\hat{y}}) = \frac{|\vec{\hat{y}}|^2}{\dim(\mathcal{V}_x)}$$
$$M(\vec{e}) = \frac{|\vec{e}|^2}{\dim(\mathcal{V}_e)}$$

We compare these by defining a statistic, F:

$$F = \frac{M(\vec{\hat{y}})}{M(\vec{e})} = \frac{\dim(\mathcal{V}_e)|\vec{\hat{y}}|^2}{\dim(\mathcal{V}_x)|\vec{e}|^2}$$
$$= \frac{(N-p-1)R^2}{p(1-R^2)}$$

When null hypothesis is true, $F \approx 1$; when it is false, $F \gg 1$.



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.

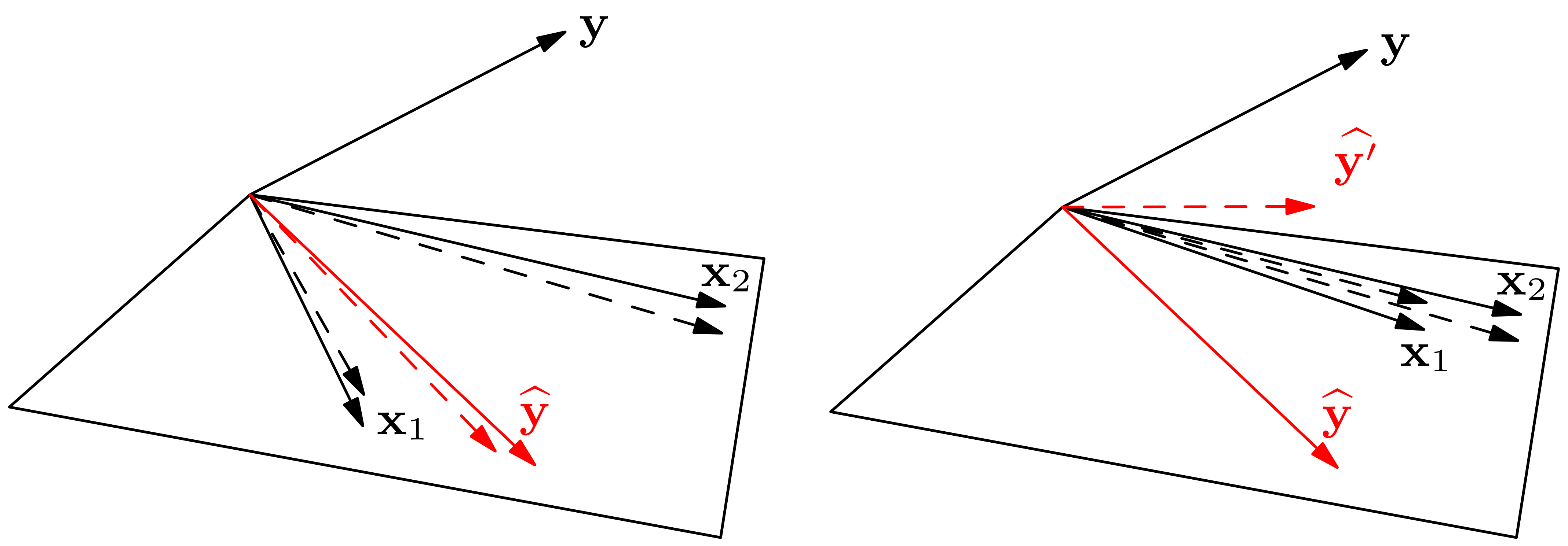
$$b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \cdots + b_p\boldsymbol{x}_p = 0$$

- 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!

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

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

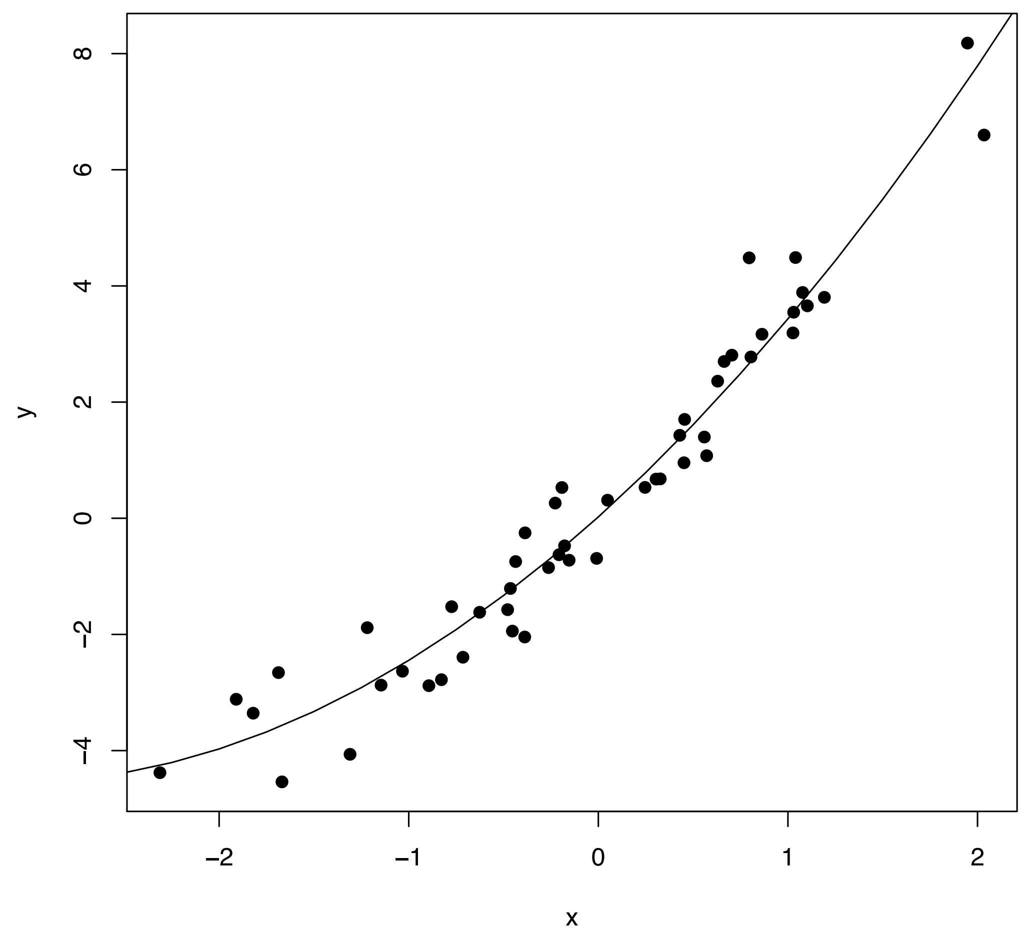
Curvilinear regression using **polynomial models** is simply multiple regression with the x_i replace by powers of x.

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

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 ***
```

ANOVA as regression

Two-group ANOVA as Regression

We can also use a geometric perspective to test whether the mean of a variable differs between two groups of subjects.

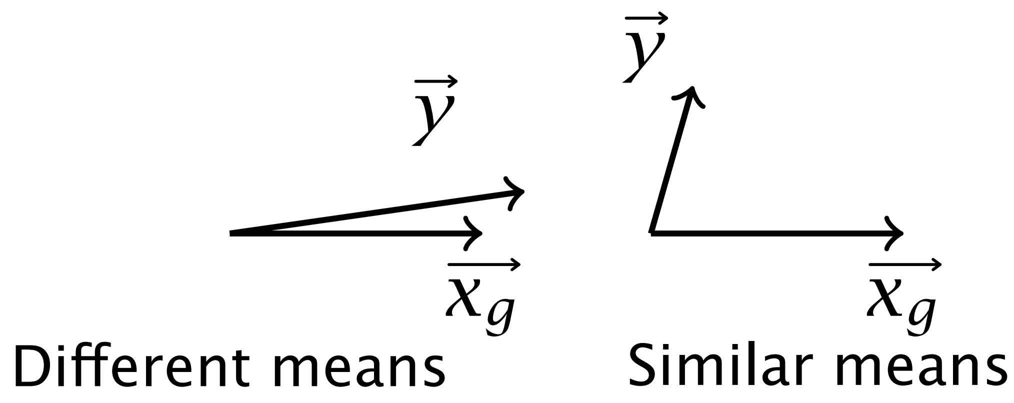
- Setup a 'dummy variable' as the predictor X_g. We assign all subjects in group 1 the value 1 and all subjects in group 2 the value -1 on the dummy variable. We then regress the variable of interest, Y, on X_g.

$$\mathsf{y} = \mathsf{X}_g\mathsf{b} + \mathsf{e}$$

Group	Raw Y_i	Raw X_i	Centered y_i	Centered x_i
1	2	-1	-3	$-\frac{4}{3}$
	3	-1	-1	$-\frac{4}{3}$
2	5	1	0	$\frac{2}{3}$
	6	1	1	$\frac{2}{3}$
	6	1	1	$\frac{2}{3}$
	7	1	2	$\frac{2}{3}$
Mean	5	$\frac{1}{3}$	0	0

Two-group ANOVA as Regression, cont

- When the means are different in the two groups, X_g will be a good predictor of the variable of interest, hence $\vec{y}$ and $\overrightarrow{x_g}$ will have a small angle between them.
- When the means in the two groups are similar, the dummy variable will not be a good predictor. Hence the angle between $\vec{y}$ and $\overrightarrow{x_g}$ will be large.



Multi-group One-way ANOVA as Regression

- Exactly the same idea applies to g groups, except now instead of one grouping variable, we define $g - 1$ grouping variables, $\dim(X_g) = g - 1$.
- Then we calculate the multiple regression as we did before:

$$y = Xb + e$$

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

Estimate b as:

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

How Do We Construct the Grouping Matrix, X_g?

Two common methods are:

1. Dummy coding – define a set of g grouping variables, where values take either 0 or 1, depending on group membership, but *use only the first* $g-1$ columns:

$$U_j = \begin{cases} 1, & \text{for every subject in group } j, \\ 0, & \text{for all other subjects.} \end{cases}$$

   and

$$X_g = [U_1, U_2, \cdots, U_{g-1}]$$

2. Effect (deviation) coding – define the U_j as above, and set:

$$X_g = [U_1 - U_g, U_2 - U_g, \cdots, U_{g-1} - U_g]$$

In general, effect coding is more similar to standard ANOVA contrasts. See this ☞ web-page for a more in depth discussion of different coding schemes.

ANOVA: Example Data Set

	g_1	g_2	g_3	g_4	
	20	21	17	8	
	17	16	16	11	
	17	14	15	8	
$M_{g.}$	18	17	16	9	$M_{..} = 15$

$$y = \begin{bmatrix} 20 \\ 17 \\ 17 \\ 21 \\ 16 \\ 14 \\ 17 \\ 16 \\ 15 \\ 8 \\ 11 \\ 8 \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

ANOVA: Example Data Set, cont

Solving for b we find:

$$\mathrm{b} = \begin{bmatrix} 15 \\ 3 \\ 2 \\ 1 \end{bmatrix}, \qquad |\hat{y}|^2 = 150, \ |e|^2 = 40$$

Since, $\dim(\mathcal{V}_x) = 3$, and $\dim(\mathcal{V}_e) = 8$, we get:

$$F = \frac{\dim(\mathcal{V}_e)|\vec{\hat{y}}|^2}{\dim(\mathcal{V}_x)|\vec{e}|^2} = 10$$

Here's the more conventional ANOVA table for the same data:

Source	df	SS	MS	F	$\Pr(F)$
Experimental	3	150	50	10	.0044
Error	8	40	5		
Total	11	190			

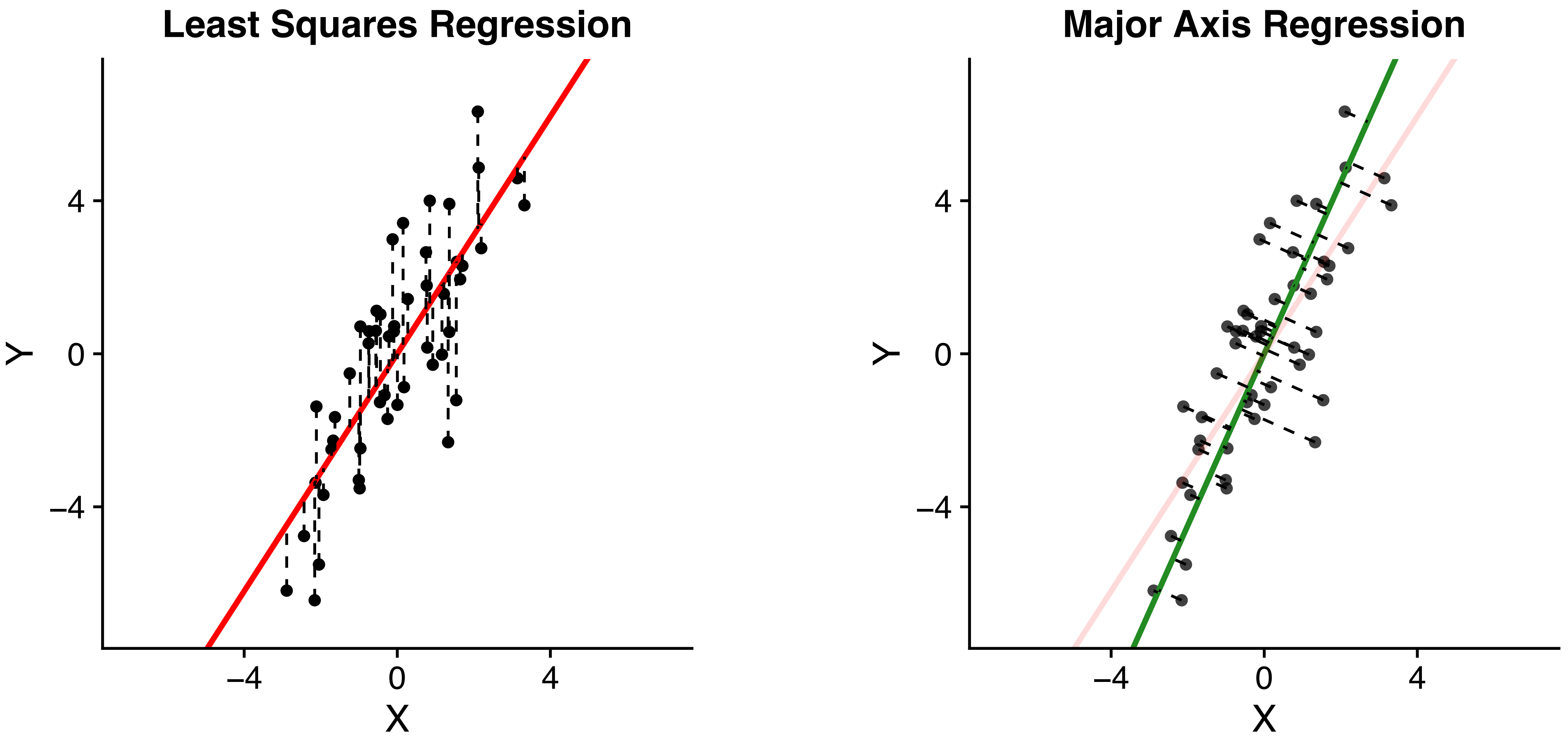
More Complex ANOVA models

- Multi-way ANOVA – used when samples are classified with respect to two or more factors (grouping variables). Allow for exploring interactions between facdtors.
- Nested ANOVA – used when there is more than one grouping variable, and the grouping variables form a nested hieararchy (groups, subgroups, subsubgroups)

As before, all of these can be treated as regression problems with appropriate design matrices!

Looking ahead

What if we changed the optimality criterion for linear regression?



Vector geometry of least squares regression and major axis regression

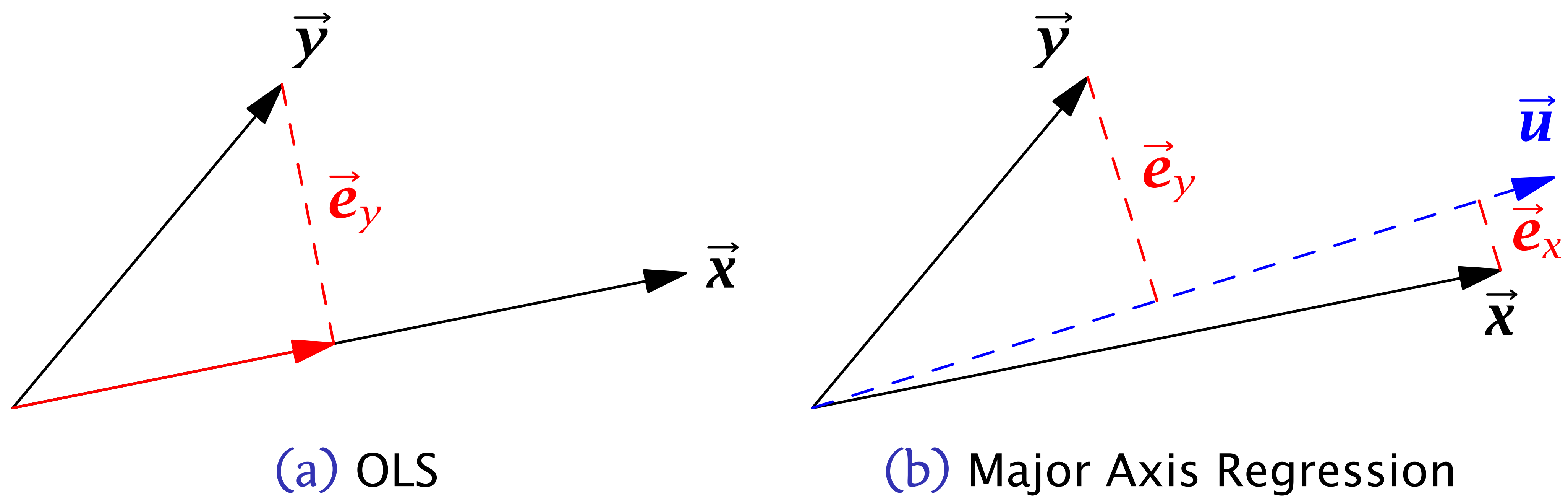


Figure: Vector geometry of ordinary least-squares and major axis regression.