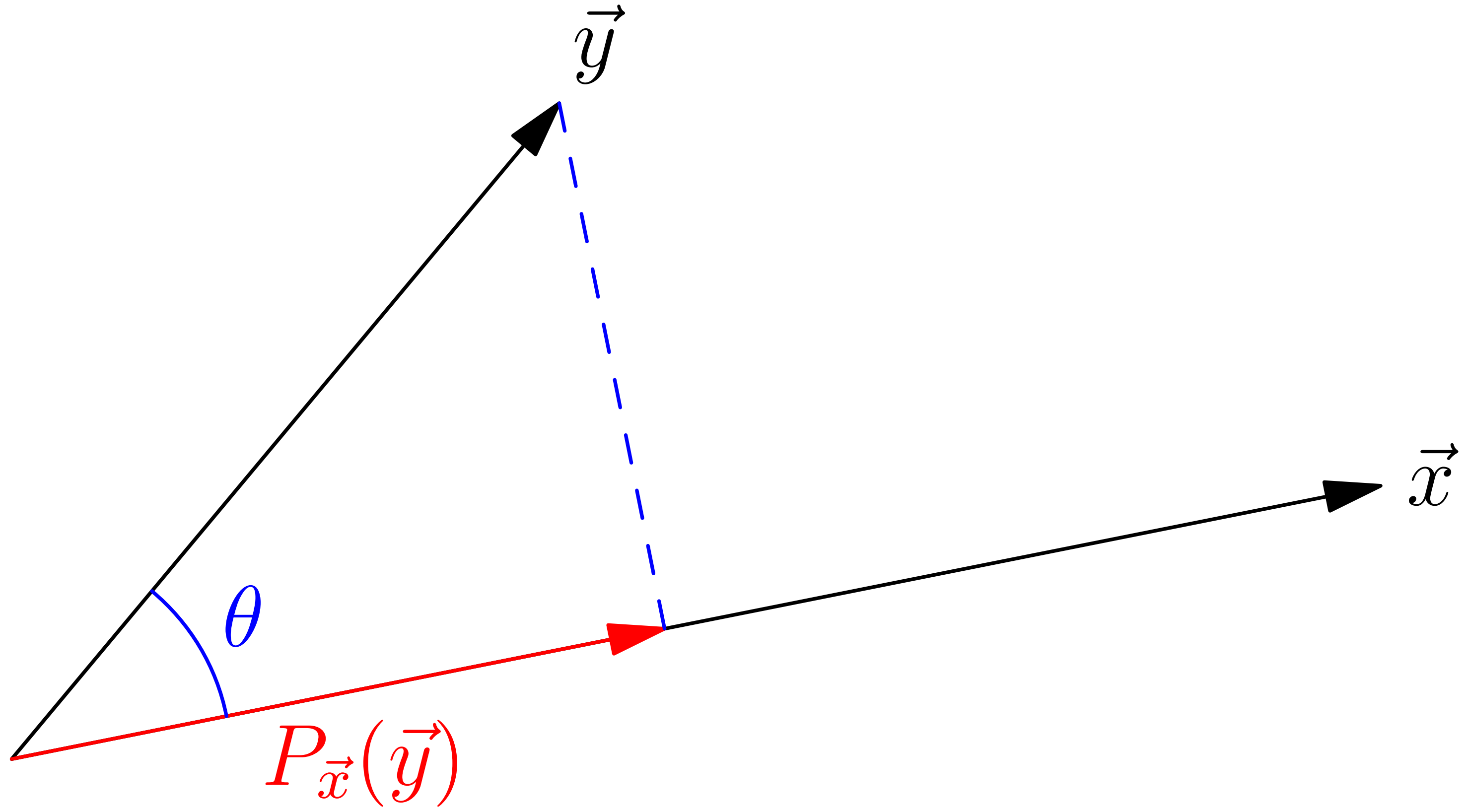
Linear Regression and ANOVA

Paul M. Magwene

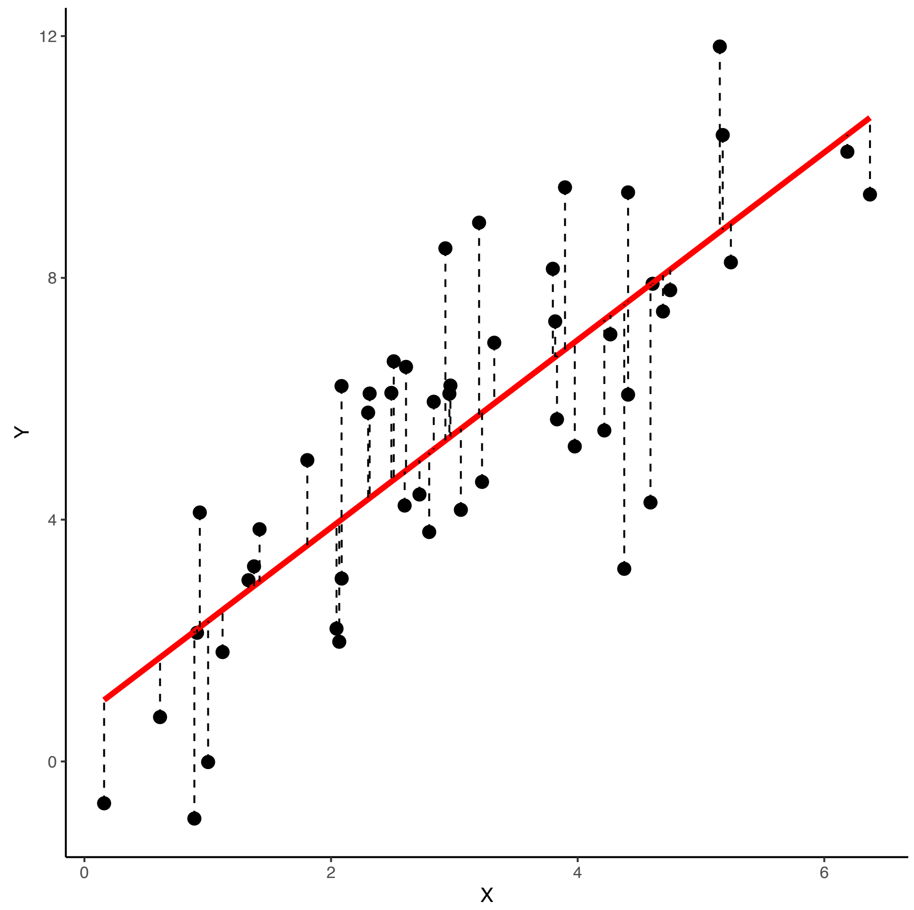
A bit of review



$$\cos\theta = ?$$

$$P_{\vec{x}}(\vec{y}) = ? \qquad C_{\vec{x}}(\vec{y}) = ?$$

Bivariate Regression: Variable Space Representation



The standard bivariate regression equation relating one observed variable X (the predictor) to another observed variable of interest, Y (the outcome) is usually written as:

$$\widehat{Y} = a + bX.$$

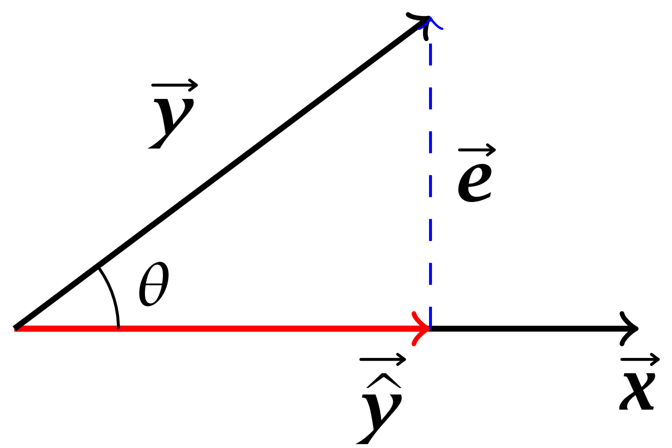
where $\widehat{Y}$ is the predicted value of Y and a (intercept) and b (slope) are chosen to minimize $\sum(Y_i - \hat{Y}_i)^2$.

Geometry of Bivariate Regression

In vector geometric terms, *regression is projection*! Consider the regression formula for mean-centered vectors:

$$\hat{Y} = bX$$

In vector terms we can view this as:



$\hat{y}$ is the closest vector to $\vec{y}$ in the subspace defined by $\vec{x}$, i.e. it is the scalar multiple of $\vec{x}$ that minimizes $|\vec{e}|$

$$\hat{y} = P_{\vec{x}}(\vec{y}) = \left(\frac{\vec{x} \cdot \vec{y}}{\vec{x} \cdot \vec{x}}\right)\vec{x}$$

Bivariate regression in vector terms

For mean centered vectors the regression equation is:

$$\vec{\hat{\boldsymbol{y}}} = b\vec{\boldsymbol{x}}$$

From the previous slide we see that we can solve b as:

$$b = \frac{\vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{y}}}{\vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{x}}}$$

Bivariate regression: Alternate formulas for slope I

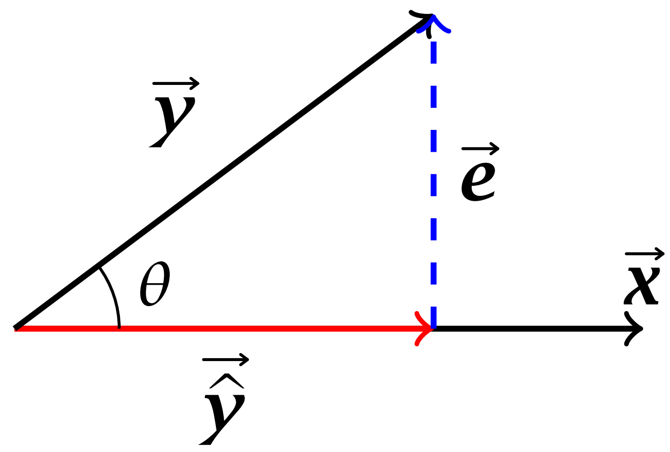
Regression equation for mean-centered vectors: $\vec{\hat{\boldsymbol{y}}} = b\vec{\boldsymbol{x}}$

There are multiple, equivalent ways to write the solution for b:

$$
\begin{aligned}
b &= \frac{\vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{y}}}{(\vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{x}})} \\
&= \frac{\vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{y}}}{|\vec{\boldsymbol{x}}|^2} \\
&= \frac{|x||y| \cos\theta}{|x|^2} \\
&= \cos\theta \frac{|y|}{|x|} \\
&= r_{XY} \frac{|y|}{|x|}
\end{aligned}
$$

Geometry of Goodness of Fit

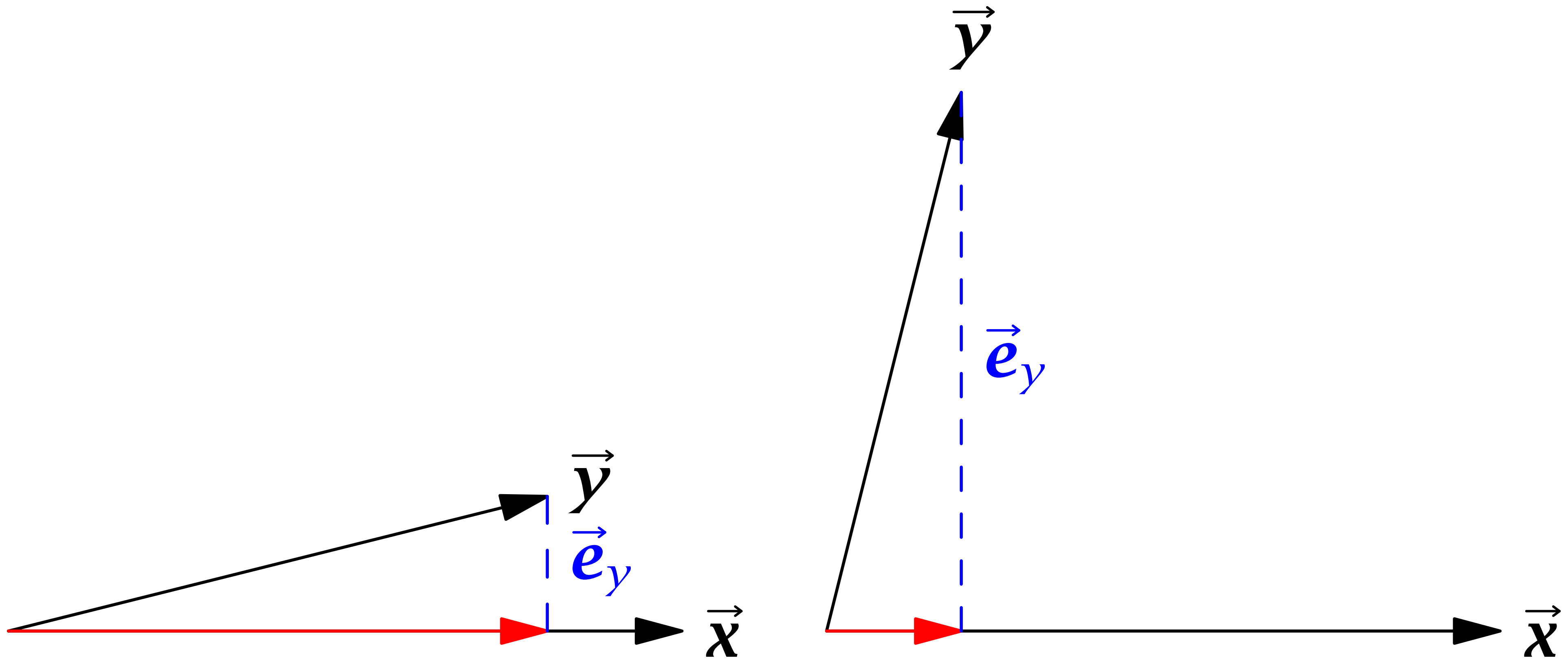
Geometric interpretation of regression goodness-of-fit:



$$|\vec{\hat{y}}|^2 + |\vec{e}|^2 = |\vec{y}|^2$$

The better the goodness-of-fit, the smaller the angle, $\cos\theta$, and the shorter residual vector, $\vec{e}$.

Geometry of Goodness of Fit



(a) Good fit (b) Bad fit

Bivariate Regression, Goodness of Fit

How well does our prediction agree with our outcome?

- Measure the angle between $\vec{\hat{y}}$ and $\vec{y}$:

$$R = \cos\theta_{\vec{y},\vec{\hat{y}}} = \frac{|\vec{\hat{y}}|}{|\vec{y}|}$$

- In the single-predictor case $R = r_{XY}$, but this is not generally true when we have multiple predictors.
- Note that $|\vec{y}|$ can be expressed as follows:

$$\begin{aligned} |\vec{\hat{y}}|^2 + |\vec{e}|^2 &= |\vec{y}|^2 \\ SS_{\text{regression}} + SS_{\text{residual}} &= SS_{\text{total}} \end{aligned}$$

- With simple substitution we can show that:

$$\begin{aligned} SS_{\text{regression}} &= R^2 SS_{\text{total}} \\ SS_{\text{residual}} &= (1 - R^2) SS_{\text{total}} \end{aligned}$$

Variable space view of multiple regression

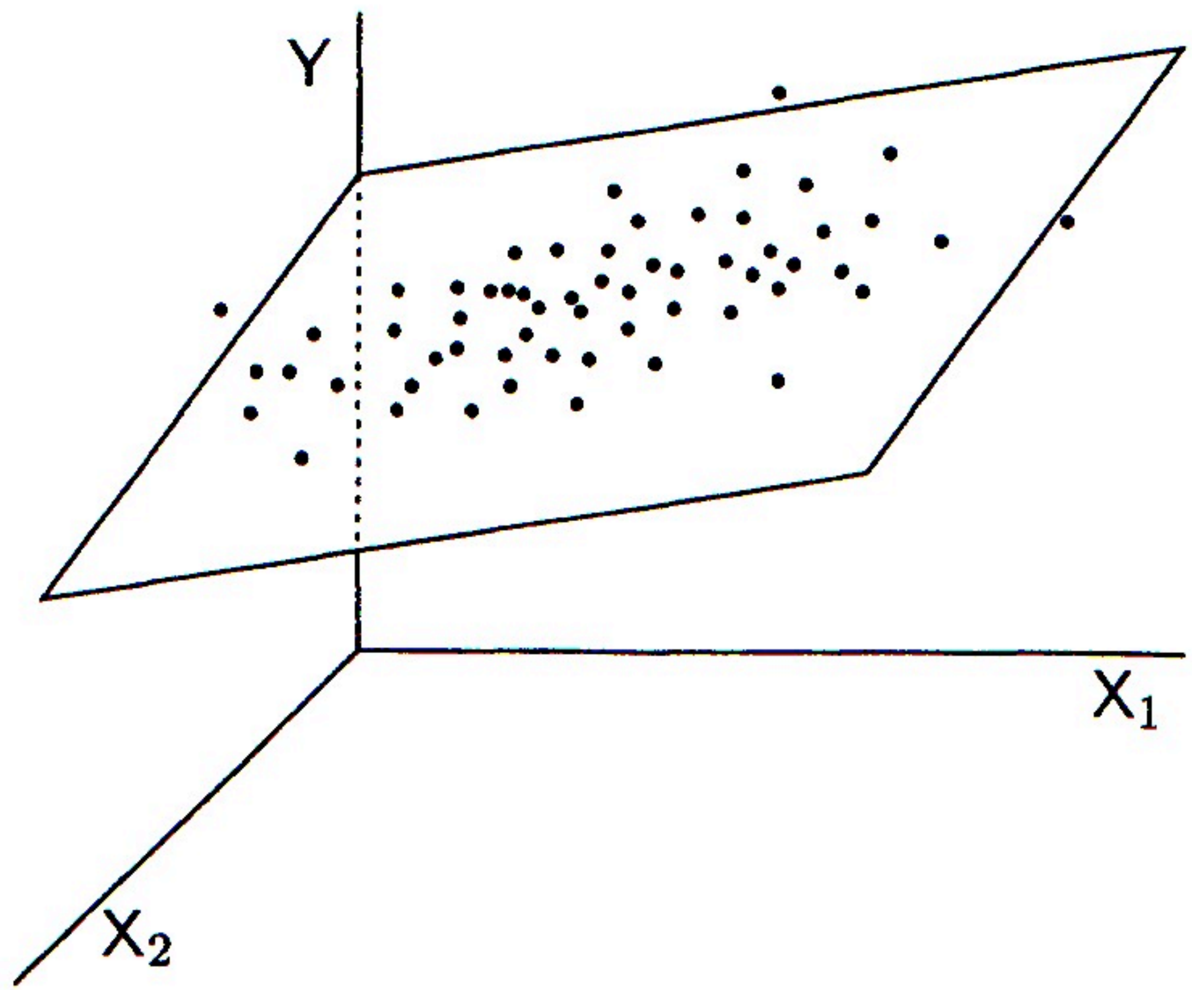
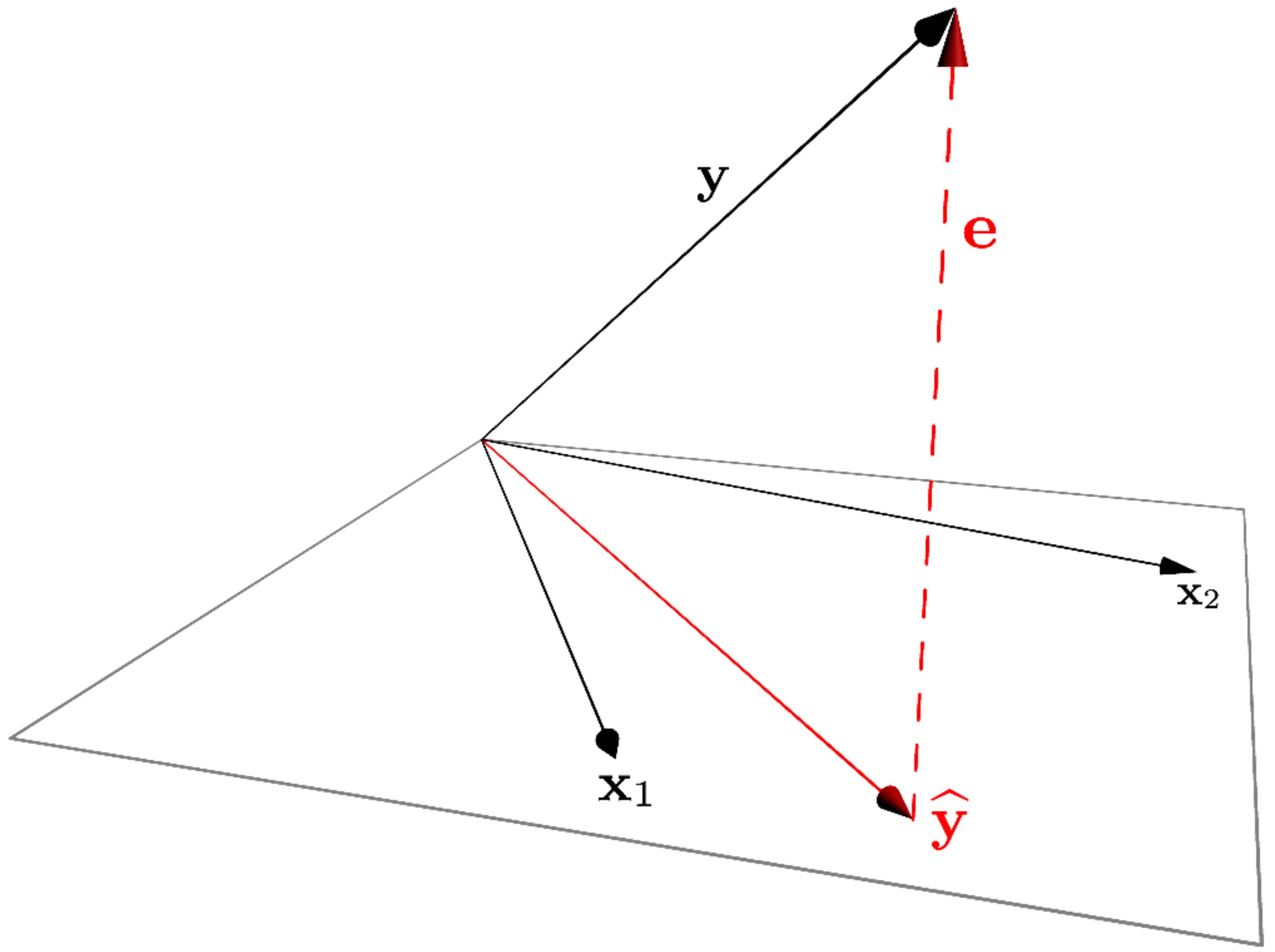


Figure 4.1: *The regression of* Y *onto* X_1 *and* X_2 *as a scatterplot in variable space.*

Subject Space Geometry of Multiple Regression



Multiple Regression

Let Y be a vector of values for the outcome variable. Let X_i be explanatory variables and let x_i be the mean-centered explanatory variables.

$$Y = \hat{Y} + e$$

where –

Uncentered version:

$$\hat{Y} = a\mathbf{1} + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p$$

Centered version:

$$\hat{y} = b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

Statistical Model for Multiple Regression

In matrix form:

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

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

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

Estimating the Coefficients for Multiple Regression

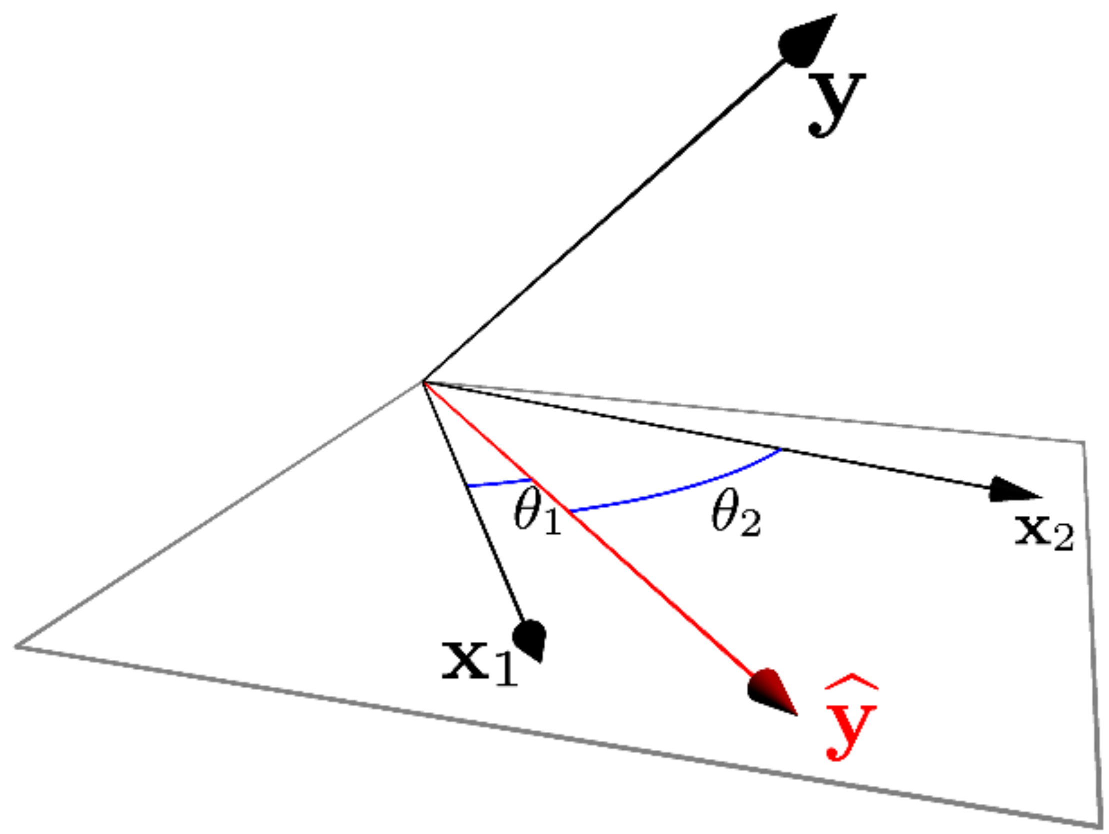
$$y = Xb + e$$

Estimate b as:

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

Multiple Regression Loadings

The regression **loadings** should be examined as well as the regression coefficients.



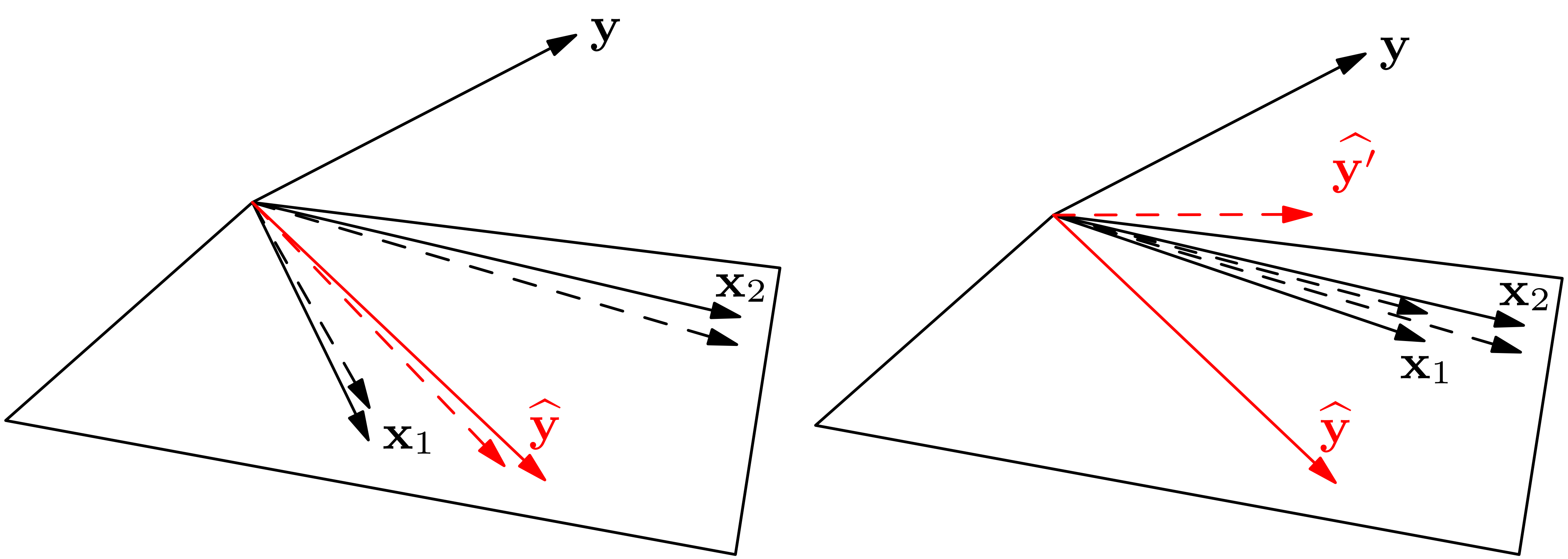
Loadings are given by:

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

Multiple regression: Cautions and Tips

- Comparing the size of regression coefficients only makes sense if all the predictor variables have the same scale
- The predictor variables (columns of X) must be linearly independent; when they're not the variables are **multicollinear**
- Predictor variables that are **nearly multicollinear** are, perhaps, even more difficult to deal with

Why is near multicollinearity of the predictors a problem?



(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).

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.

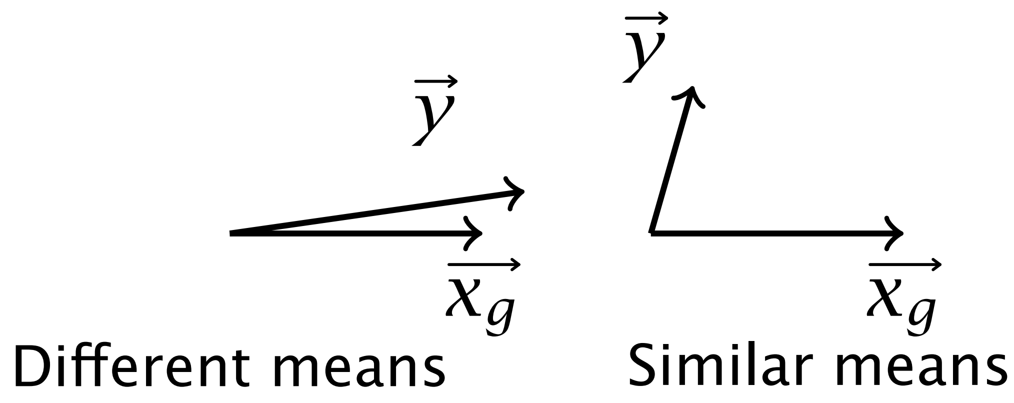
- Setup a ‘dummy variable’ as the predictor X_g. We assign all subjects in group 1 the value 1 and all subjects in group 2 the value -1 on the dummy variable. We then regress the variable of interest, Y, on X_g.

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

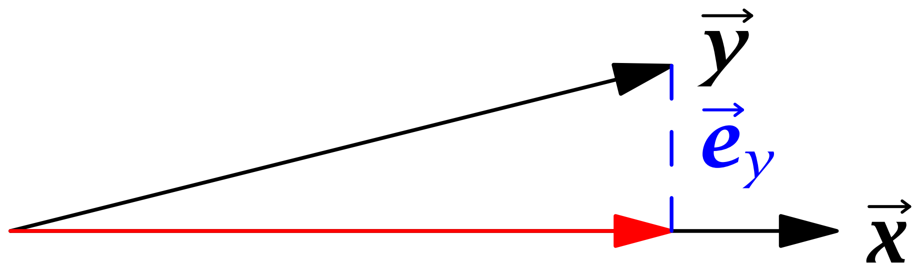
	Raw		Centered	
Group	Y_i	X_i	y_i	x_i
1	2	-1	-3	$-\frac{4}{3}$
	3	-1	-1	$-\frac{4}{3}$
2	5	1	0	$\frac{2}{3}$
	6	1	1	$\frac{2}{3}$
	6	1	1	$\frac{2}{3}$
	7	1	2	$\frac{2}{3}$
Mean	5	$\frac{1}{3}$	0	0

Two-group ANOVA as Regression, cont

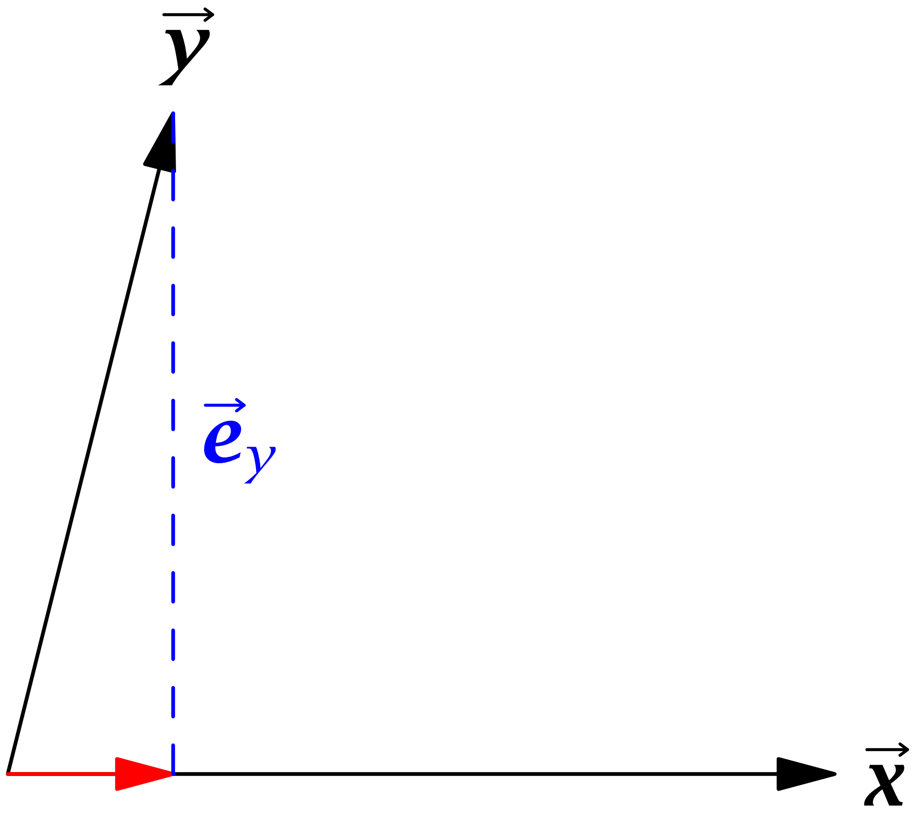
- When the means are different in the two groups, X_g will be a good predictor of the variable of interest, hence $\vec{y}$ and $\overrightarrow{x_g}$ will have a small angle between them.
- When the means in the two groups are similar, the dummy variable will not be a good predictor. Hence the angle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



Geometry of Goodness of Fit



(a) Good fit



(b) Bad fit

Comparing the Effect Space and the Error Space

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.

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

We compare these by defining a statistic, F:

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

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

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:

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

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:

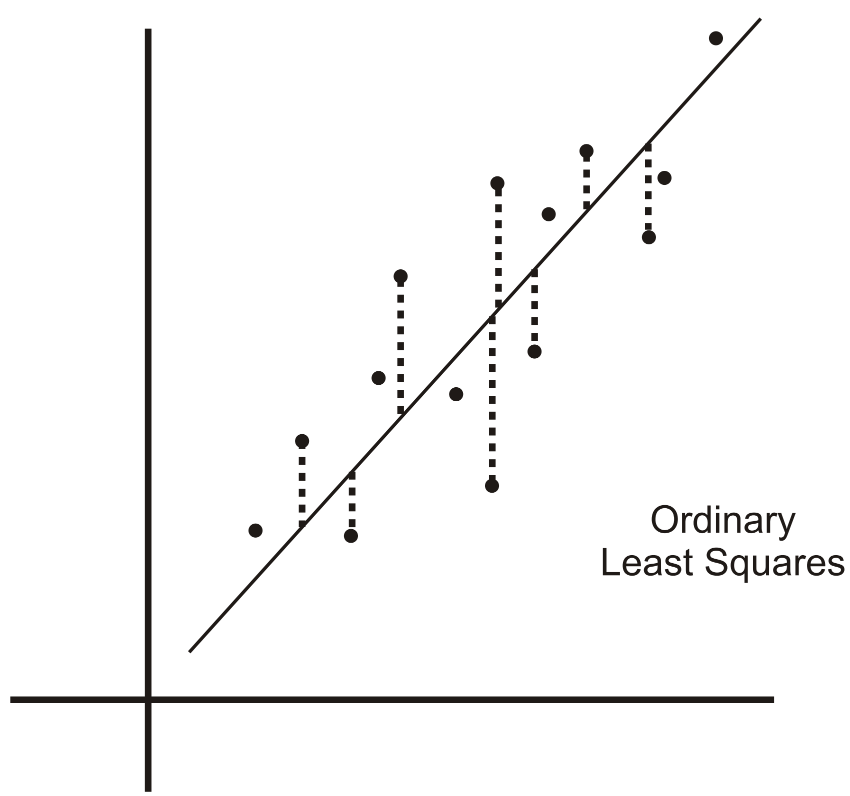
Source	df	SS	MS	F	$\Pr(F)$
Experimental	3	150	50	10	.0044
Error	8	40	5		
Total	11	190			

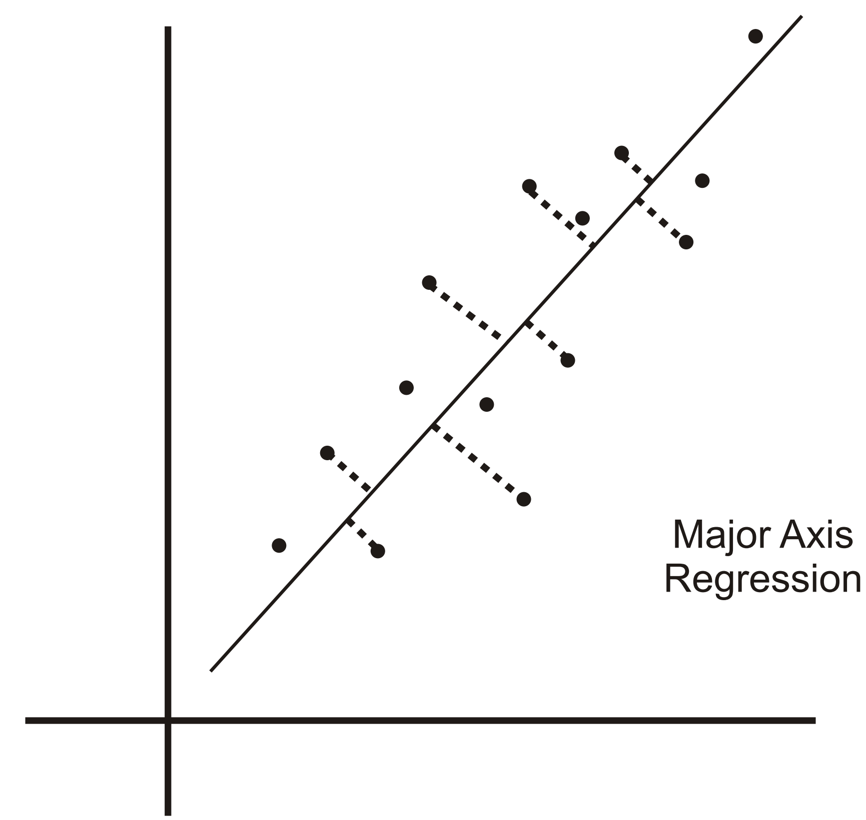
More Complex ANOVA models

- 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!

Least Squares Regression vs. Major Axis Regression





Vector Geometry of Major Axis Regression

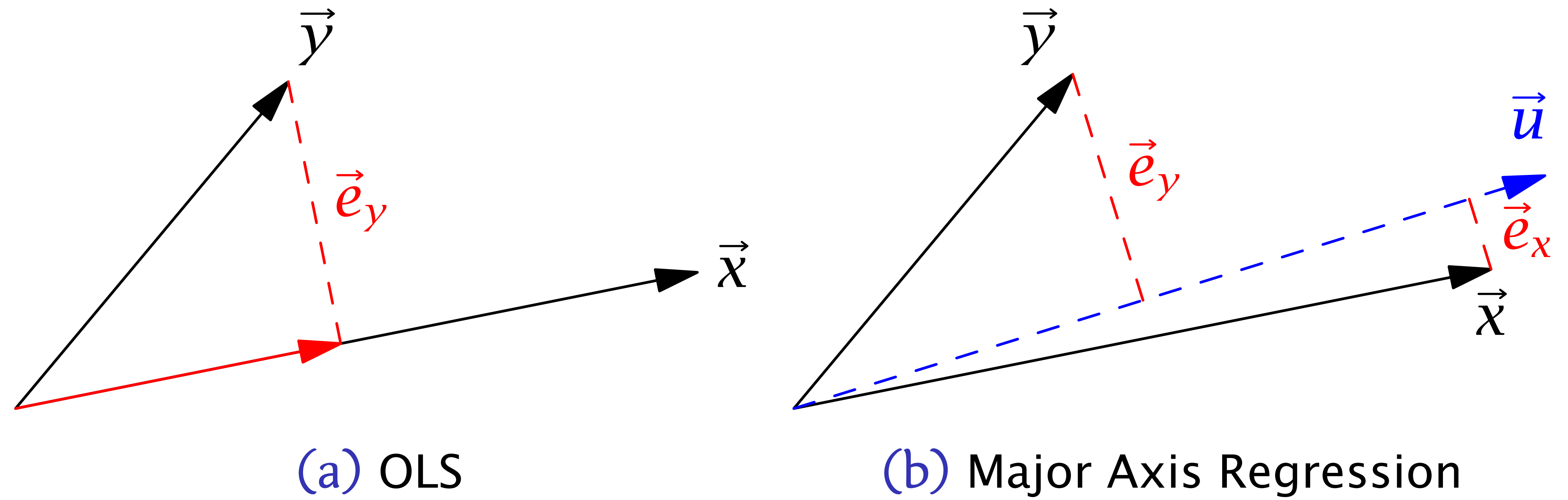


Figure: Vector geometry of ordinary least-squares and major axis regression.