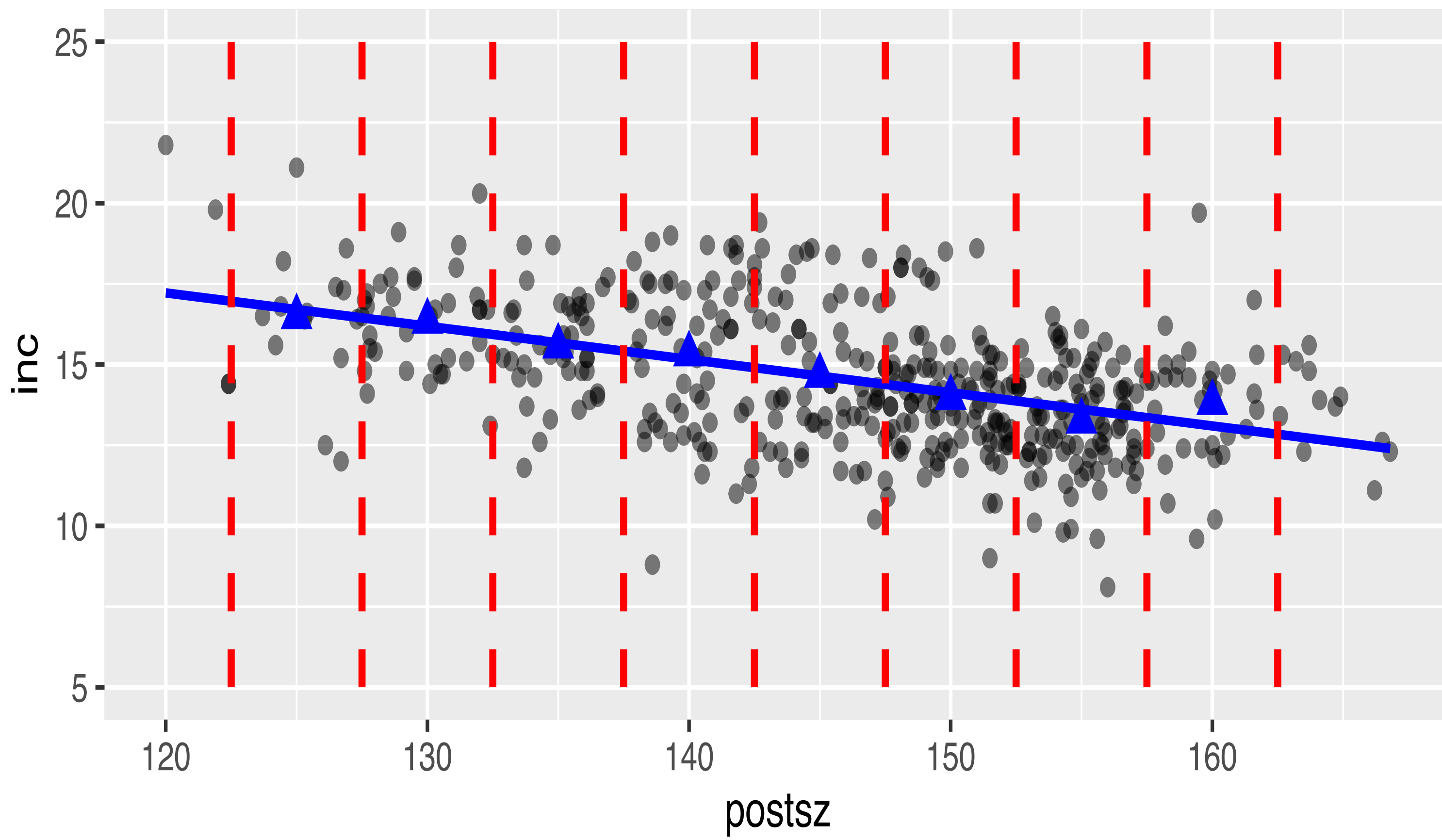
Multiple Linear Regression – A Geometric View

Topics

- Switch from an observation perspective to a variable perspective
- Review Linear Algebra
- Build intuition with a toy example
- Guest Lecture from the DS Education Program

Observation Perspective

Scatter plot examines (x_i, y_i)



Extend Empirical Risk

Minimize empirical risk to estimate y by a linear function of x

$$\min_{a,b} \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

Sum over the observations (x_i, y_i)

Minimizing values $\hat{a} = \bar{y} - \hat{b}\bar{x}$ $\hat{b} = r\frac{SD_y}{SD_x}$

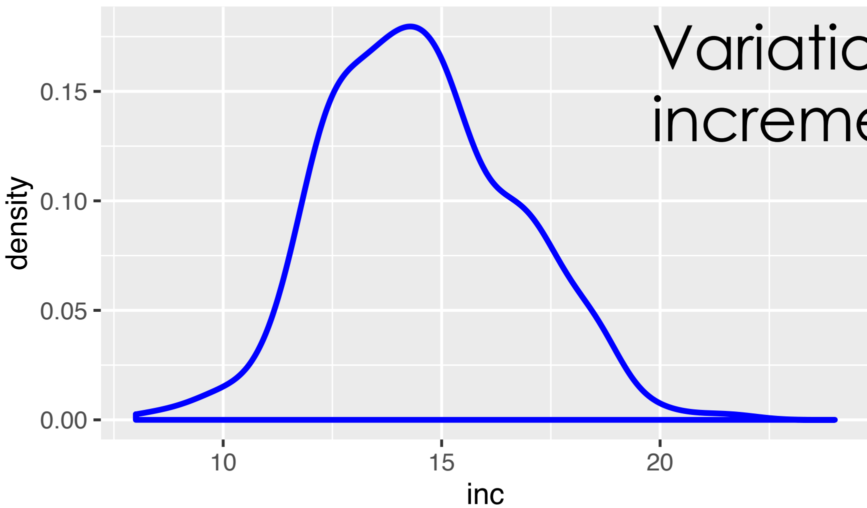
Predictor $\hat{y} = \hat{a} + \hat{b}x$

Observation perspective (x_i, y_i)

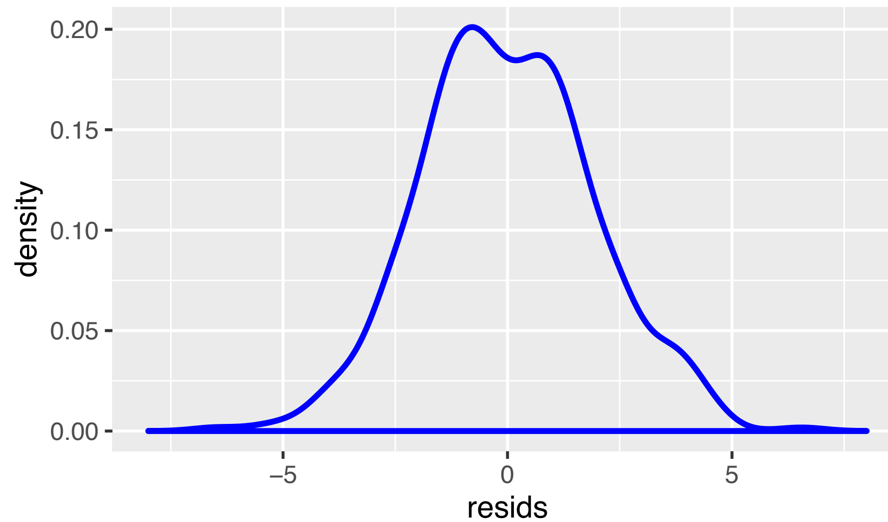
Observe $(x_i, y_i), \; i = 1, \ldots, n$

Predict $(x_i, \hat{y}_i), \; i = 1, \ldots, n$

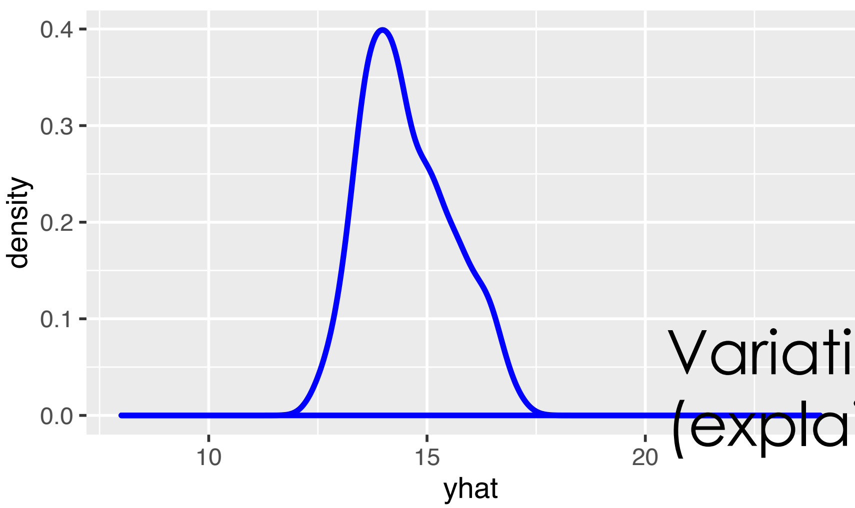
Error in prediction $e_i = y_i - \hat{y}_i, \; i = 1, \ldots, n$



Variation in growth increment



Variation in the errors (unexplained)



Variation along the line (explained)

Switch to Variable Perspective

Variable perspective

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\vec{\hat{y}} = \hat{a}\vec{1} - \hat{b}\vec{x}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \hat{a} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \hat{b} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Variable perspective

$$\vec{y},\ \vec{x},\ \vec{\hat{y}},\ \vec{e} \in \mathbb{R}^n$$

$$\vec{\hat{y}} = \hat{a}\vec{1} + \hat{b}\vec{x}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \hat{a} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \hat{b} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

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

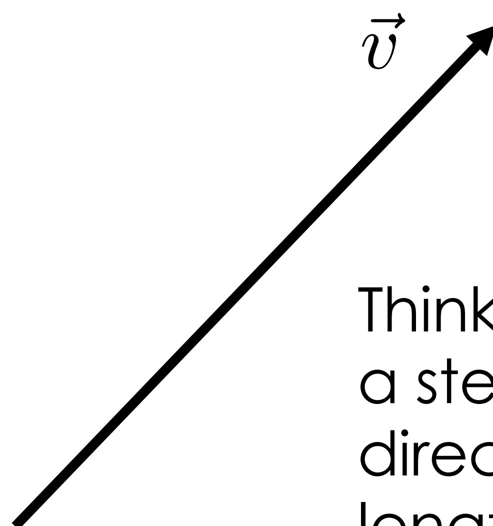
$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

Review
Vectors

Vector

Vector – consists of a length and direction $\vec{v}$

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

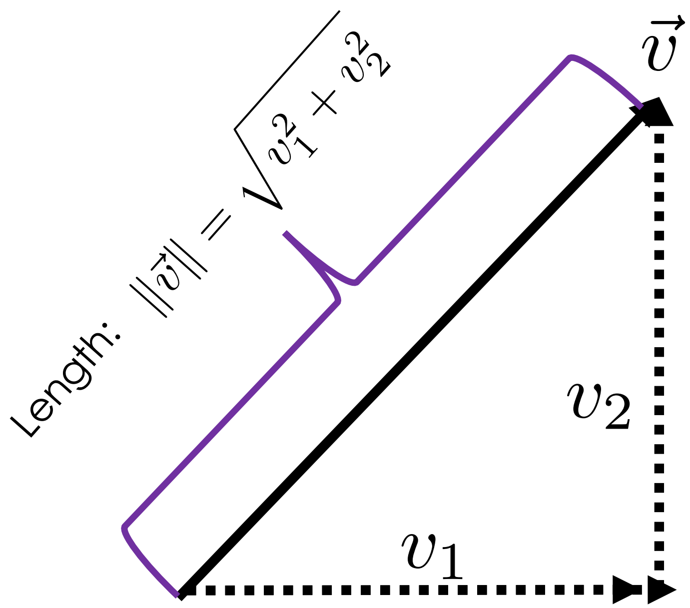


Think of the vector as a step in a particular direction for a certain length

Length of a Vector

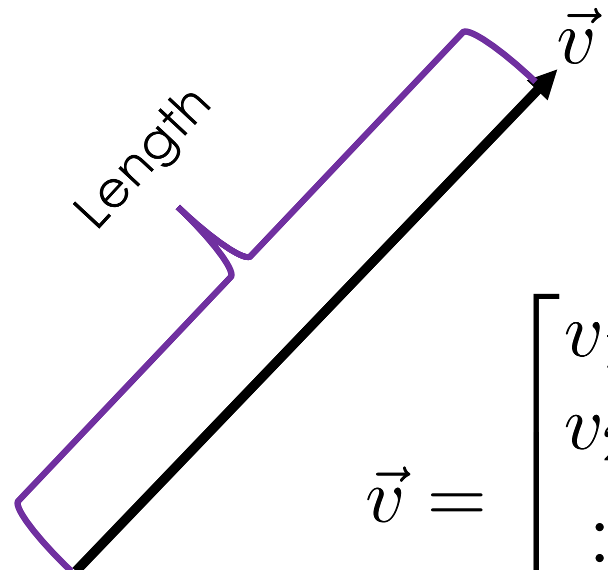
Pythagorean's theorem

$$\|\vec{v}\| = \sqrt{v_1^2 + v_2^2}$$



Length of a Vector

Pythagorean's theorem carries over to vectors in *n*-dimensions, i.e., $\vec{v} \in \mathbb{R}^n$



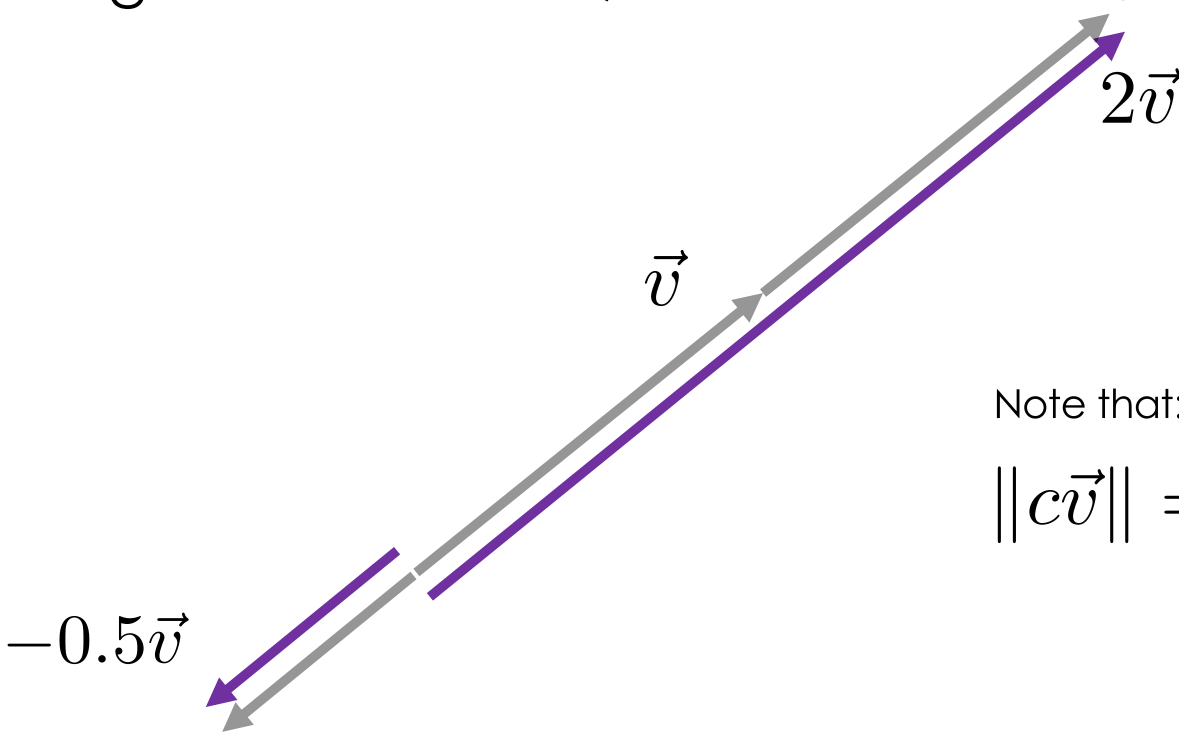
$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

$$\|\vec{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

Scale a Vector

Change the length of the vector, for some scalar c

$$c\vec{v} = \begin{bmatrix} cv_1 \\ cv_2 \\ \vdots \\ cv_n \end{bmatrix}$$



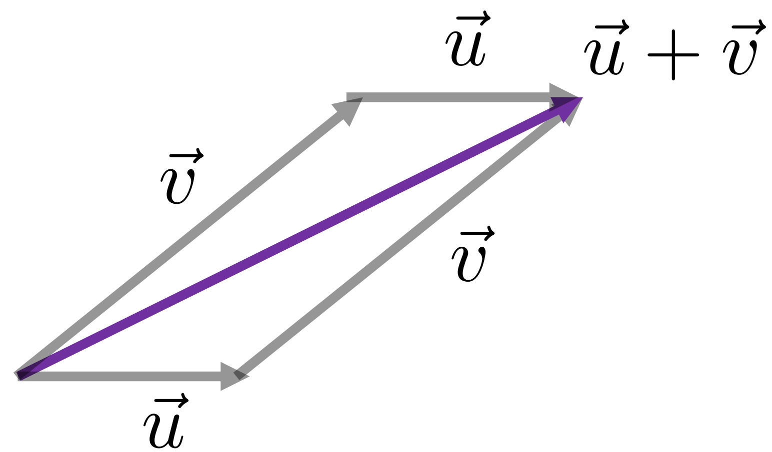
Note that:

$$\|c\vec{v}\| = |c|\|\vec{v}\|$$

Add Two Vectors

Take a step according to the length and direction of u and from that point take a step in the direction of v for the length of v

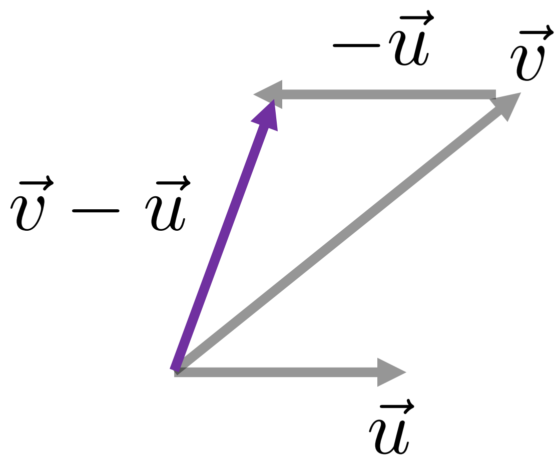
$$\vec{u} + \vec{v} = \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_n + v_n \end{bmatrix}$$



Subtract Two Vectors

Take a step according to the length and direction of v and from that point take a step in the direction of -u for the length of u

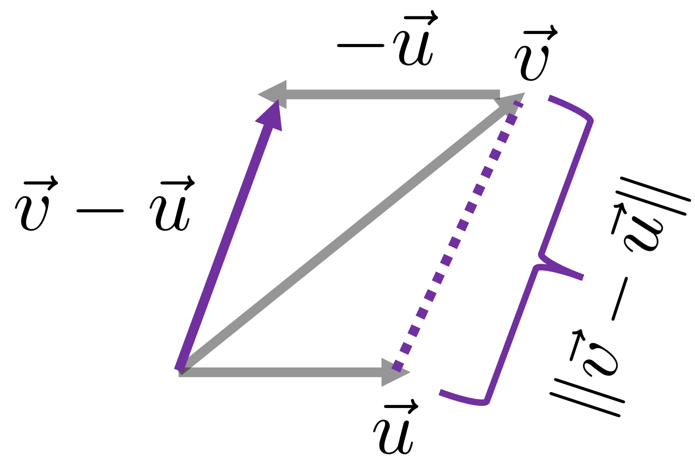
$$\vec{v} - \vec{u} = \begin{bmatrix} v_1 - u_1 \\ v_2 - u_2 \\ \vdots \\ v_n - u_n \end{bmatrix}$$



Distance Between Two Vectors

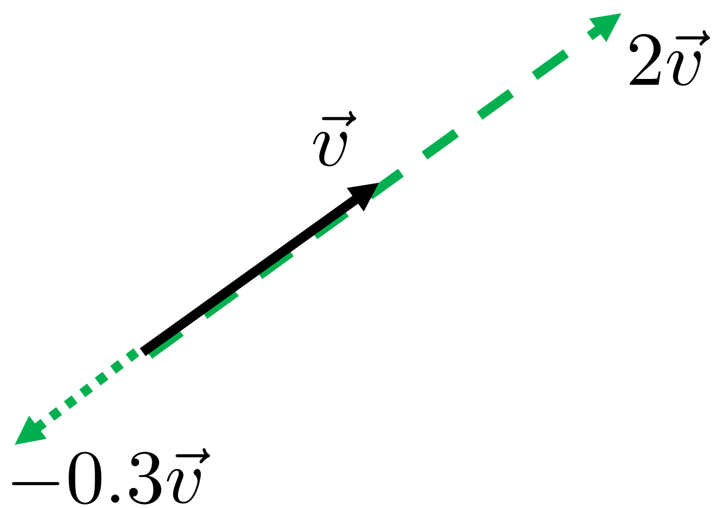
The distance between u and v is the length of v – u.

$$\vec{v} - \vec{u} = \begin{bmatrix} v_1 - u_1 \\ v_2 - u_2 \\ \vdots \\ v_n - u_n \end{bmatrix}$$



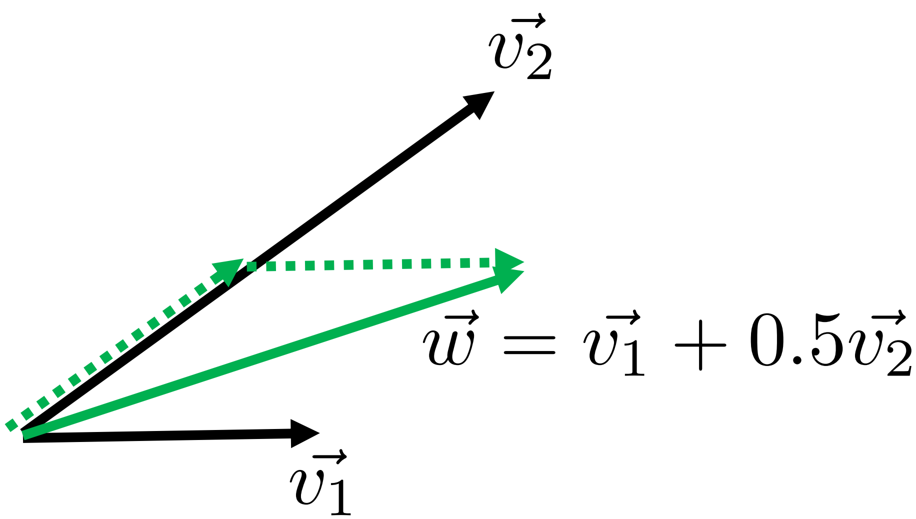
Vector space – span $\{\vec{v}\}$

All vectors in the vector space can be expressed as a scalar multiple of the spanning vector $\vec{v}$



Vector space – span $\{\vec{v}_1, \vec{v}_2\}$

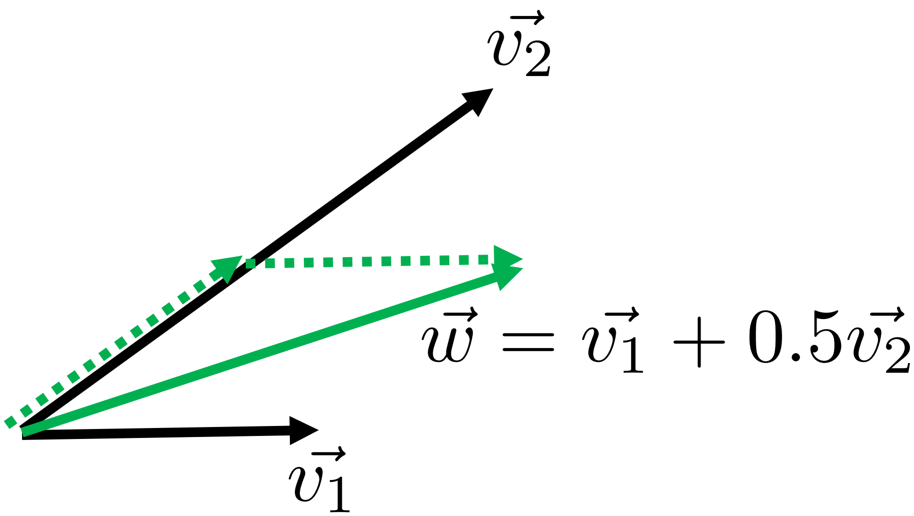
All vectors in the vector space can be expressed as a linear combination of the spanning vectors



Vector space – span$\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_p\}$

All vectors in the vector space can be expressed as a linear combination of the spanning vectors

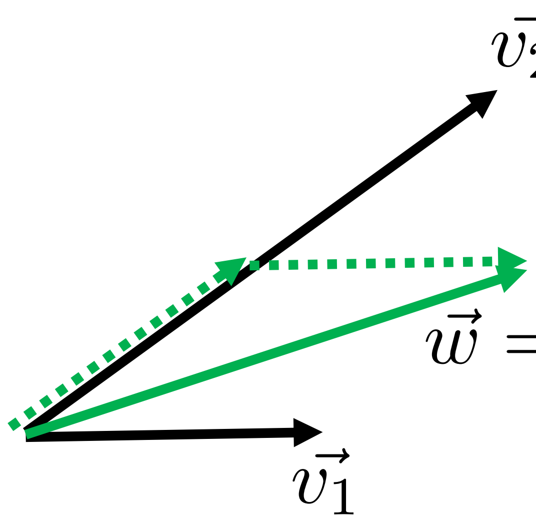
$$\vec{w} = c_1\vec{v}_1 + c_2\vec{v}_2 + \ldots + c_p\vec{v}_p$$



$$\vec{w} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_p \\ \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix}$$

Vector space – span$\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_p\}$

All vectors in the vector space can be expressed as a linear combination of the spanning vectors

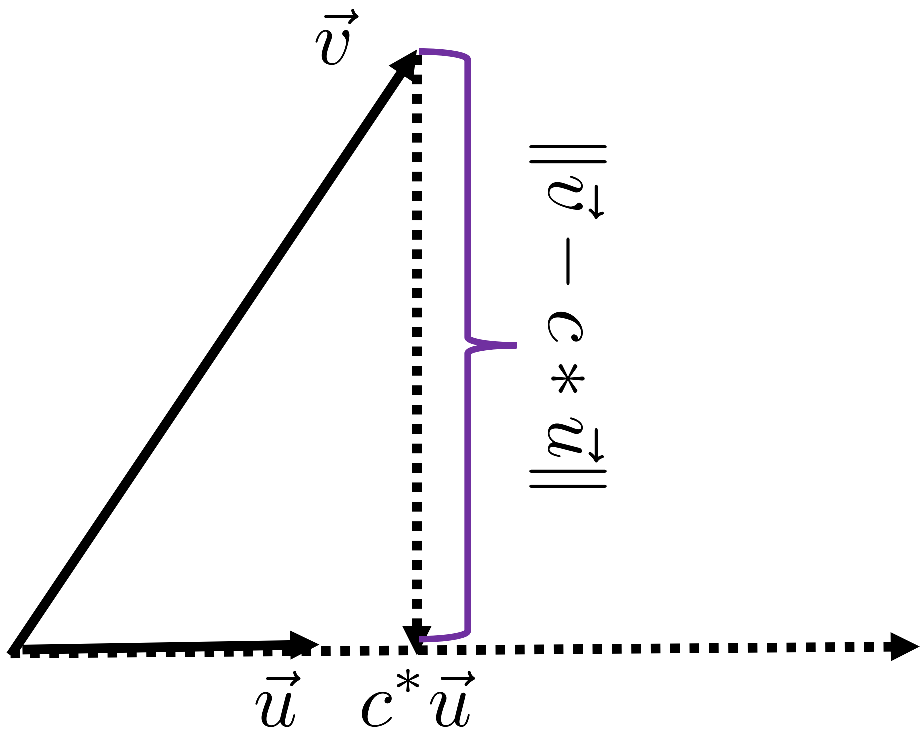


$$\vec{w} = \vec{v}_1 + 0.5\vec{v}_2$$

$$\vec{w} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_p \\ \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix}$$

$$w_i = c_1 v_{1i} + c_2 v_{2i} + \cdots + c_p v_{pi}$$

Projection



Projecting v onto u means we find the point in the subspace span{u} that is as close to v as possible

Inner product $\vec{u} \cdot \vec{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n$

- Length $\|\vec{v}\| = \sqrt{\vec{v} \cdot \vec{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$
- Pythagorean's theorem in n-dimensions
- Distance between two vectors $\|\vec{u} - \vec{v}\|$
- Inner product for orthogonal vectors is 0

AKA Dot Product

Projection: $c^* \vec{u}$

$\vec{v} - c^* \vec{u}$ Orthogonal to $\mathrm{span}\{\vec{u}\}$

Find the vector in the span{u} that is closest to v

$$\|\vec{v} - c\vec{u}\|^2 = \|\vec{v} - c^*\vec{u} + c^*\vec{u} - c\vec{u}\|^2$$

$$= \|\vec{v} - c^*\vec{u}\|^2 + \|c^*\vec{u} - c\vec{u}\|^2$$

$$+2(\vec{v} - c^*\vec{u}) \cdot (c^*\vec{u} - c\vec{u})$$

Inner Product is 0

$$= \|\vec{v} - c^*\vec{u}\|^2 + \|c^*\vec{u} - c\vec{u}\|^2$$

Minimized for c = c*

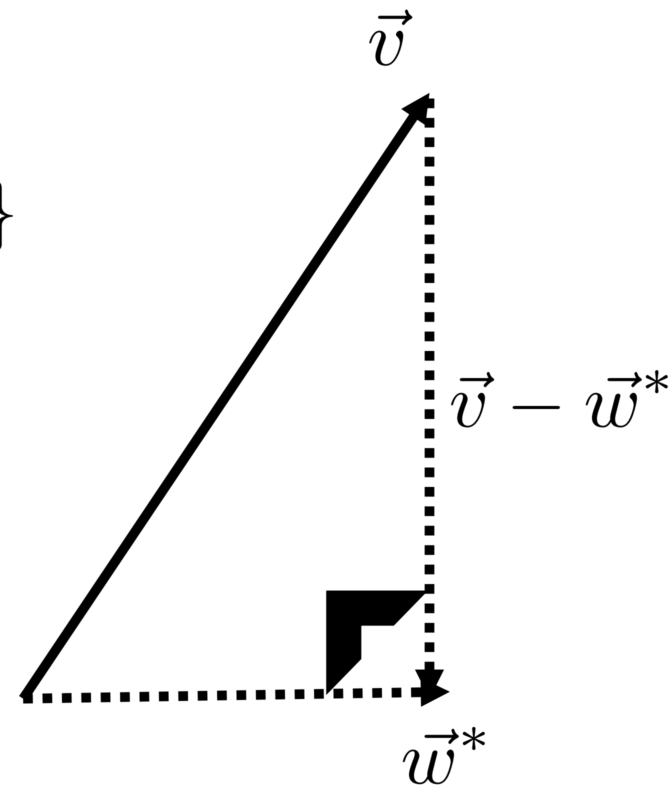
General:

Consider the span: $\mathrm{span}\{\vec{u}_1, \ldots, \vec{u}_p\}$

Find the closest vector in this span to $\vec{v}$

$$\min_{\vec{w} \in \mathrm{span}\{\vec{u}_1, \ldots, \vec{u}_p\}} \|\vec{v} - \vec{w}\|^2$$

The minimizing vector will be
the projection onto the span



$$(\vec{v} - \vec{w}^*) \cdot \vec{w}^* = 0$$

Bring in the Data

Variable perspective

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\vec{\hat{y}} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

The fitted values are in the span $\vec{\hat{y}} \in \operatorname{span}\{\vec{1}, \vec{x}\}$

Recall

Data $(y_1, y_2, \ldots y_n)$

Find the summary statistic that minimizes the L_2 error

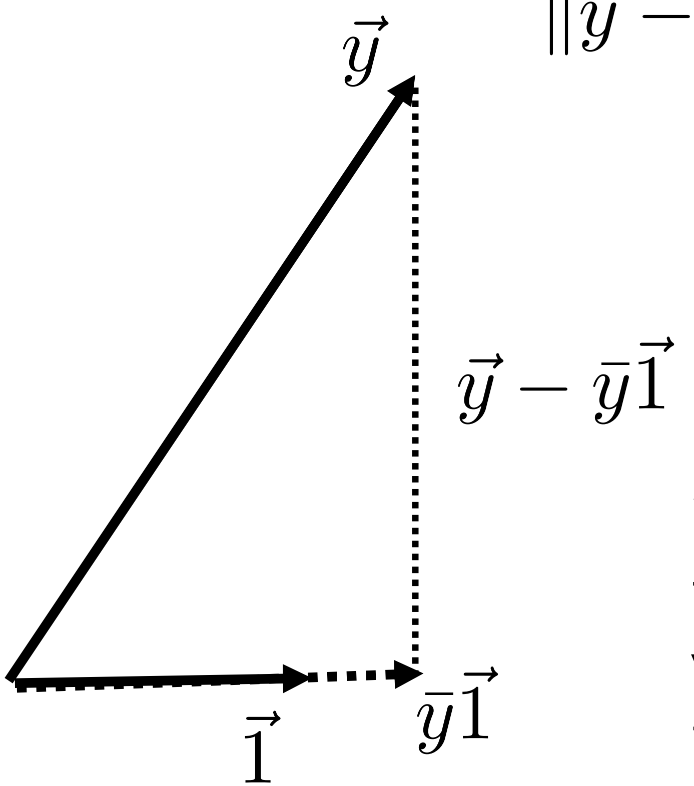
$$\min_c \sum_{i=1}^{n} (y_i - c)^2 \Rightarrow \bar{y}$$

Equivalent to finding the closest vector in the span of 1 to y

$$\min_c \|\vec{y} - c\vec{1}\|^2 \Rightarrow \bar{y}$$

Summarizing y by a constant

$$\begin{aligned}\|\vec{y} - c\vec{1}\|^2 &= \|\vec{y} - \bar{y}\vec{1} + \bar{y}\vec{1} - c\vec{1}\|^2 \\ &= \|\vec{y} - \bar{y}\vec{1}\|^2 + \|\bar{y}\vec{1} - c\vec{1}\|^2\end{aligned}$$



When we minimized the L_2 loss to find the best summary of y, we were in effect projecting y onto the span of 1

Least Squares

$$\min_{a,b} \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

$$= \min_{a,b} \|\vec{y} - (a\vec{1} + b\vec{x})\|^2$$

Minimize the squared distance between y and a vector in the span of 1 and x

$$\vec{\hat{y}} \in \mathrm{span}\{\vec{1}, \vec{x}\}$$

Pythagorean's theorem tells us that the projection will have the smallest distance

Least Squares

Data $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$

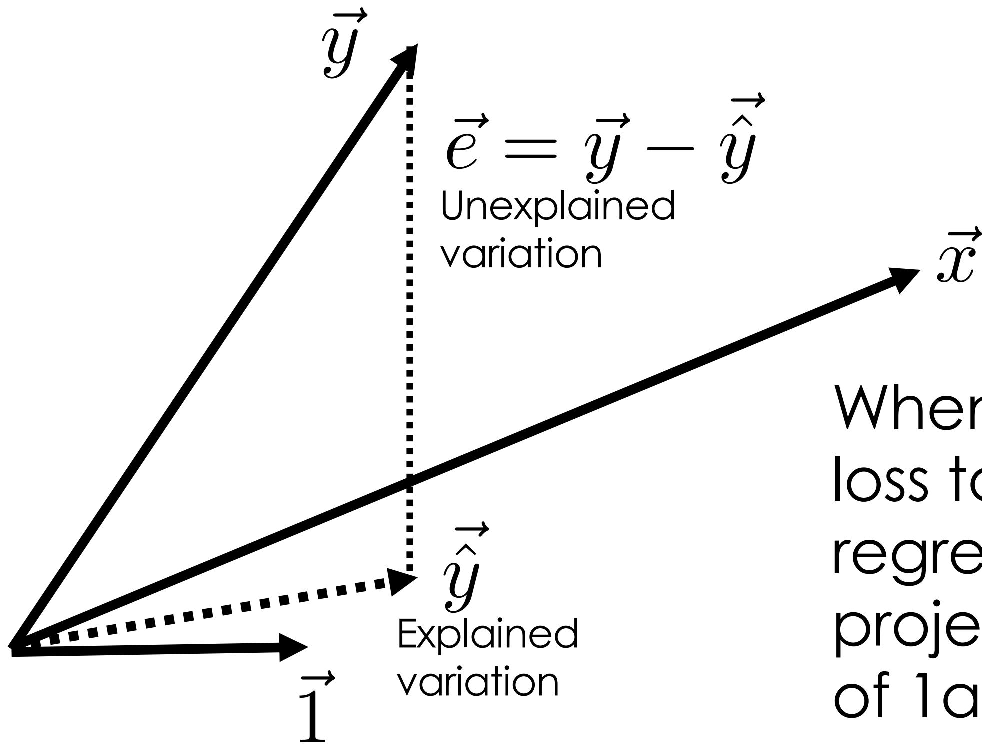
Find the coefficients to the line that minimizes the L_2 error

$$\min_{a,b} \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

Equivalent to finding the closest vector in the span of 1 and x to y

$$\min_{a,b} \|\vec{y} - (a\vec{1} + b\vec{x})\|^2$$

Regression from the Variable Perspective



When we minimized the L_2 loss to find the best regression line, we were projecting y onto the span of 1and x.

Useful Properties

- Average of the residuals is 0 $\vec{e} \cdot \vec{1} = 0$

- The inner product of the fitted values and residuals is 0

$$\vec{e} \cdot \vec{\hat{y}} = 0$$

- The inner product of the residuals and x is 0

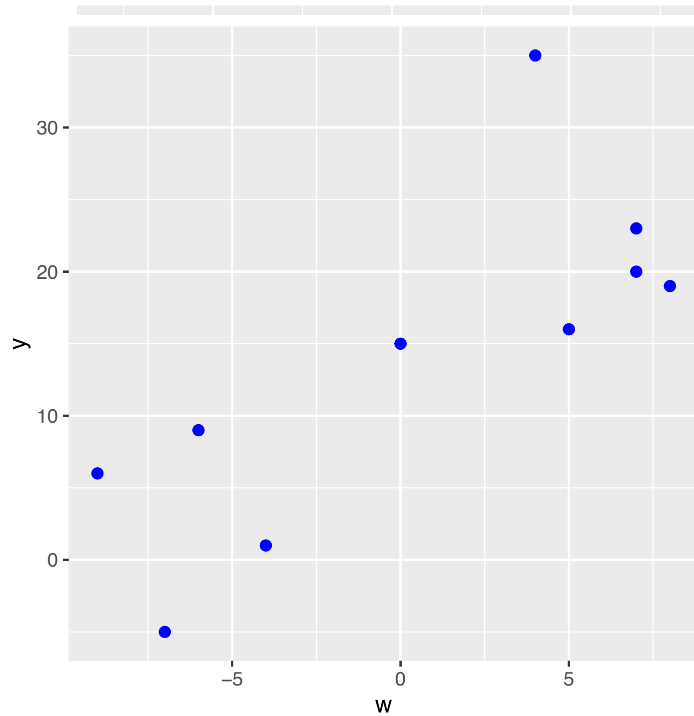
$$\vec{e} \cdot \vec{x} = 0$$

Toy Example

Toy Example

$$\begin{array}{c} \vec{y} \\ \begin{bmatrix} 20 \\ 19 \\ 35 \\ \vdots \\ 15 \\ 1 \end{bmatrix} \end{array} \qquad \begin{array}{cc} \vec{x} & \vec{w} \end{array} \begin{bmatrix} 2 & 7 \\ 1 & 8 \\ 9 & 4 \\ \vdots & \vdots \\ 5 & 0 \\ 3 & -4 \end{bmatrix}$$

Correlation(w,y) = 0.77



Correlation(x,y) = 0.17

Best simple linear regression – fit y to w

$$\vec{\hat{y}} = \hat{\beta}_0 \vec{1} + \hat{\beta}_1 \vec{w}$$

$$= 13\vec{1} + 1.4\vec{w}$$

$$\frac{\text{Explained SS}}{\text{Total SS}} = 0.60 \quad (= r^2)$$

But, what about a 2-variable model to predict y?

The correlation between x and y is weak so it seems like we will gain little if we add x to the equation.

Two variable regression: fit y to w and x

Consider the model we are fitting

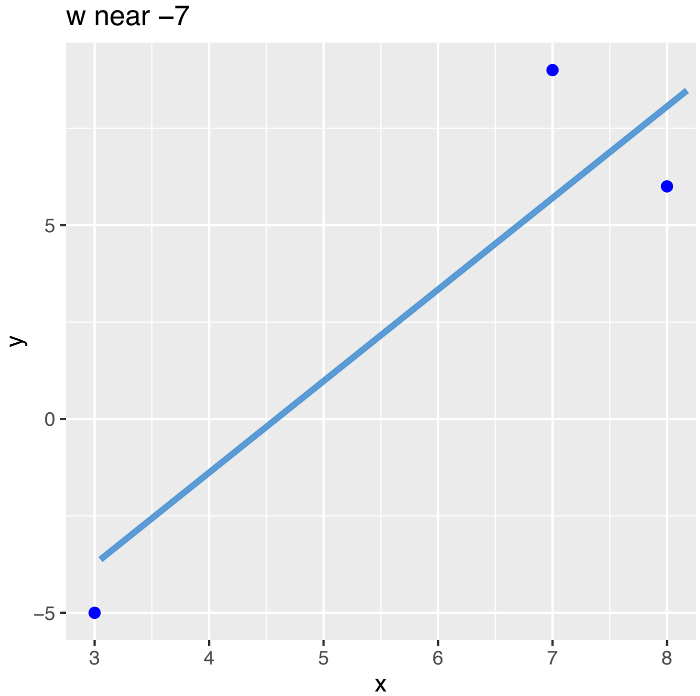
$$\vec{y} \approx \beta_0 \vec{1} + \beta_1 \vec{w} + \beta_2 \vec{x}$$

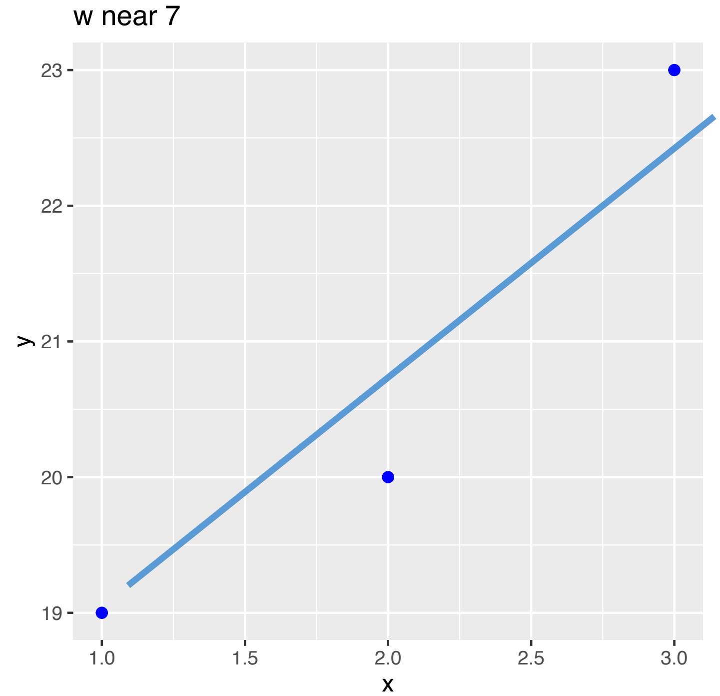
For a fixed w = 7, this model is: $\vec{y} \approx (\beta_0 + 7\beta_1)\vec{1} + \beta_2 \vec{x}$

For a fixed w = -7, this model is: $\vec{y} \approx (\beta_0 - 7\beta_1)\vec{1} + \beta_2 \vec{x}$

Notice that for a fixed value of w, we see that the relationship between x and y is linear with the same slope.

There aren't many points in these plots, but they show a linear relationship of slope about 2.





Slopes for these subsets are both about 2

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 w + \hat{\beta}_2 x$$

$$\hat{y} = 7 \times 10^{-15} + 2w + 3x$$

$$\frac{\text{Explained SS}}{\text{Total SS}} = 1.0$$

Multiple R^2

$$\begin{array}{c} \vec{y} \\ \begin{bmatrix} 20 \\ 19 \\ 35 \\ \vdots \\ 15 \\ 1 \end{bmatrix} \end{array} \quad \begin{array}{cc} \vec{x} & \vec{w} \end{array} \begin{array}{c} \\ \begin{bmatrix} 2 & 7 \\ 1 & 8 \\ 9 & 4 \\ \vdots & \vdots \\ 5 & 0 \\ 3 & -4 \end{bmatrix} \end{array} \quad \begin{bmatrix} \beta_0 \\ \beta_x \\ \beta_w \end{bmatrix}$$

Here, y is perfectly described by a linear function of x and w, even though the pairwise plots didn't reveal this relationship

Interpretation of the fitted coefficients

$$\vec{\hat{y}} = \hat{\beta}_0 \vec{1} + \hat{\beta}_1 \vec{w}$$

$$= 13\vec{1} + 1.4\vec{w}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 w + \hat{\beta}_2 x$$

$$\hat{y} = 2w + 3x$$

Notice that w has a coefficient of 1.4 in the simple linear model, and a coefficient of 2 in the two-variable model.

These two coefficients are not the same because the models are different. The 1.4w yields the best fit when w is alone in the model

The 2w yields the best fit when w is in a model with x. That is, the coefficient is dependent on the other variables in the model.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 w + \hat{\beta}_2 x$$

$$\hat{y} = 7 \times 10^{-15} + 2w + 3x$$

$$\frac{\text{Explained SS}}{\text{Total SS}} = 1.0$$

$$\begin{array}{c} \vec{y} \\ \begin{bmatrix} 20 \\ 19 \\ 35 \\ \vdots \\ 15 \\ 1 \end{bmatrix} \end{array} \qquad \begin{array}{cc} \vec{x} & \vec{w} \end{array} \begin{array}{c} \\ \begin{bmatrix} 2 & 7 \\ 1 & 8 \\ 9 & 4 \\ \vdots & \vdots \\ 5 & 0 \\ 3 & -4 \end{bmatrix} \end{array} \quad \begin{bmatrix} \beta_0 \\ \beta_x \\ \beta_w \end{bmatrix}$$

Here, y is perfectly described by a linear function of x and w, even though the pairwise plots didn't reveal this relationship

How well is y described by w and v?

$$\frac{\text{Explained SS}}{\text{Total SS}} = \begin{array}{c} 0.6 \\ 0.7 \\ 1.0 \end{array}$$

$$\begin{array}{ccc} \vec{y} & & \begin{array}{cc} \vec{w} & \vec{v} \end{array} \\ \begin{bmatrix} 20 \\ 19 \\ 35 \\ \vdots \\ 15 \\ 1 \end{bmatrix} & & \begin{bmatrix} 7 & 14 \\ 8 & 16 \\ 4 & 8 \\ \vdots & \vdots \\ 0 & 0 \\ -4 & -8 \end{bmatrix} \end{array}$$

How well is y described by w and v?

$$\frac{\text{Explained SS}}{\text{Total SS}} = \quad \mathbf{0.6}$$

$$\begin{array}{ccc} \vec{y} & \vec{w} & \vec{v} \\ \begin{bmatrix} 20 \\ 19 \\ 35 \\ \vdots \\ 15 \\ 1 \end{bmatrix} & \begin{bmatrix} 7 & 14 \\ 8 & 16 \\ 4 & 8 \\ \vdots & \vdots \\ 0 & 0 \\ -4 & -8 \end{bmatrix} & \end{array}$$

The fit is the same as for the simple linear regression with w because v is in the span of w.

We can still find y-hat, but we do not have a unique solution for the coefficients of v and w.

$$\begin{array}{c} \vec{y} \\ \begin{bmatrix} 19.4 \\ 18.3 \\ 34.8 \\ \vdots \\ 15.6 \\ 1.8 \end{bmatrix} \end{array} \qquad \begin{array}{cc} \vec{x} & \vec{w} \\ \end{array} \begin{bmatrix} 2 & 7 \\ 1 & 8 \\ 9 & 4 \\ \vdots & \vdots \\ 5 & 0 \\ 3 & -4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_x \\ \beta_w \end{bmatrix}$$

Add error to y

$$\hat{y} = -19.8 + 3.01x + 1.98w$$

$$R^2 = 0.98$$

$$
\begin{array}{c}
\vec{y} \\
\begin{bmatrix} 19.4 \\ 18.3 \\ 34.8 \\ \vdots \\ 15.6 \\ 1.8 \end{bmatrix}
\end{array}
\qquad
\begin{array}{c}
\begin{array}{cccc} & \vec{x} & \vec{w} & \vec{v} \end{array} \\
\begin{bmatrix} 1 & 2 & 7 & 14 \\ 1 & 1 & 8 & 16 \\ 1 & 9 & 4 & 8 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 5 & 0 & 0 \\ 1 & 3 & -4 & -8 \end{bmatrix}
\end{array}
\begin{bmatrix} \beta_0 \\ \beta_x \\ \beta_w \\ \beta_v \end{bmatrix}
$$

y has errors
v = 2w

What is the fit?

Y-hat is the same as the fit to x and w

But the coefficients are not uniquely determined

Suppose we have only 5 observations

$$
\begin{array}{c}
y \\
\begin{bmatrix} 19.4 \\ 18.3 \\ 34.8 \\ 15.6 \\ 1.8 \end{bmatrix}
\end{array}
\qquad
\begin{array}{c}
\begin{array}{ccccc} \mathbf{1} & x & w & x^2 & w^2 \end{array} \\
\begin{bmatrix} 1 & 2 & 7 & 4 & 49 \\ 1 & 1 & 8 & 1 & 64 \\ 1 & 9 & 4 & 81 & 16 \\ 1 & 5 & 0 & 25 & 0 \\ 1 & 3 & -4 & 9 & 16 \end{bmatrix}
\end{array}
\begin{array}{c}
\beta \\
\begin{bmatrix} \beta_0 \\ \beta_x \\ \beta_w \\ \beta_{x^2} \\ \beta_{w^2} \end{bmatrix}
\end{array}
$$

$$5 = n = 1 + p$$

$$\hat{y} = -19 + 3x + 2.3w - 1.2x^2 - 0.02w^2 \qquad R^2 = 1$$

Summary Properties of Multiple Linear Regression

- Multiple Linear Least Squares regression is equivalent to projecting y onto the span of the features.

- When $p = n$ the errors are 0 and the fit is perfect.

- When $\mathrm{rank}\ \mathbb{X} < p$ there is not a unique solution for the coefficients

- When $\vec{y} \in \mathrm{span}\{\mathbb{X}\}$ the features perfectly predict y.