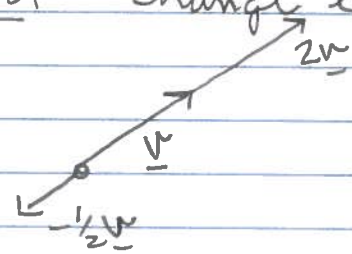
Linear Models – The Geometric Perspective

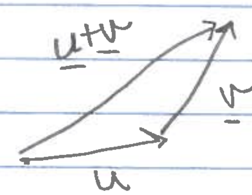
Review Vectors, Vector Space, and the innerproduct

<u>vector</u> – consists of a length and direction. Think of it as a step in a particular direction for a certain distance

<u>scale a vector</u> – change length



<u>add two vectors</u> $\underline{u}+\underline{v}$



Take a step according to length and direction of $\underline{u}$ and then from that point take a step in the direction of $\underline{v}$ for the length of $\underline{v}$

<u>Vector space</u> $V = \text{span}\{v_1, v_2, \ldots, v_p\}$

all vectors in V can be expressed as a linear combination of $v_1, \ldots, v_p$

That is $w \in V$ means $\underline{w} = c_1\underline{v}_1 + c_2\underline{v}_2 + \ldots + c_p\underline{v}_p$

<u>Inner Product</u> $\underline{u} \cdot \underline{v} = u_1v_1 + u_2v_2 + \ldots + u_nv_n$

vectors in $\mathbb{R}^n$

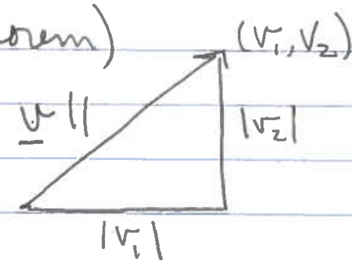
$\underline{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$ $\underline{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$

<u>Length</u> (Pythagorean's theorem)

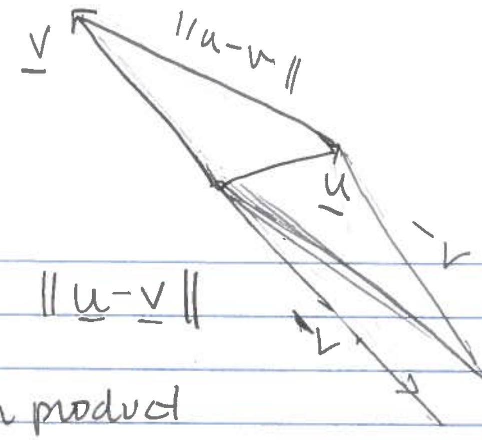
$\sqrt{v_1^2+v_2^2} := \|\underline{v}\|$

$\downarrow$

$\underline{v} \cdot \underline{v}$



Carries over to $v \in \mathbb{R}^3$



Distance between $\underline{u}$ and $\underline{v}$ $\quad \|\underline{u}-\underline{v}\|$

Geometric definition of inner product

$$\underline{u}\cdot\underline{v} = \|\underline{u}\|\,\|\underline{v}\|\cos\theta$$

angle between vectors $\underline{u}$ and $\underline{v}$

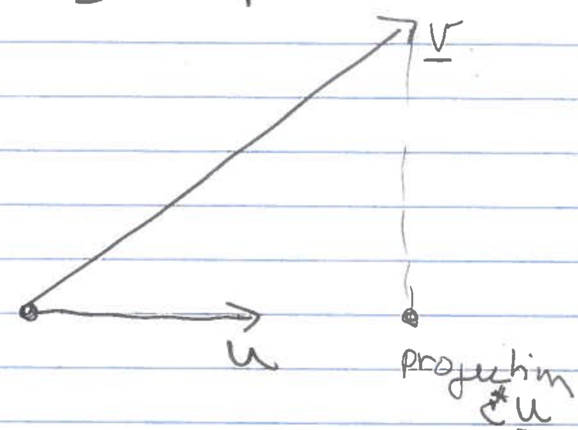
$$\|\underline{u}+v\|^2 = (\underline{u}+\underline{v})\cdot(\underline{u}+\underline{v})$$

$$= \underline{\|u\|^2+\|v\|^2} + \underbrace{2(u\cdot v)}_{\text{must be } 0}$$

Pythagoras says

Projections

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



$$c^* = \frac{\underline{v}\cdot\underline{u}}{\underline{u}\cdot\underline{u}}$$

Proof

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

$$= \|\underline{v}-c^*u\|^2 + \|c^*u-cu\|^2 + \underbrace{2(v-c^*u)\cdot(c^*u-cu)}$$

inner product is 0 $\quad 2(c^*-c)(v-c^*\underline{u})\cdot u \qquad c^* = \frac{v\cdot u}{u\cdot u}$

$$= 2(c^*-c)(v\cdot u - v\cdot u)$$

$= 0$ so minimized for c^*

What does this have to do with fitting a linear model?

Related to L_2 loss

Consider a variable with values $y_1, y_2, \ldots, y_n$

Write it as an $n \times 1$ column vector $\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

For L_2 loss what is the constant that 'best predicts' $\underline{y}$

Recall

$$\min_c \sum_{i=1}^{n} (y_i - c)^2 \qquad \text{and minimizing } c \text{ is } \hat{c} = \bar{y}$$

Reexpress this as length

$$\min_c \|\underline{y} - c\underline{1}\|^2 = \min_c (\underline{y} - c\underline{1}) \cdot (\underline{y} - c\underline{1})$$

where $\underline{1}$ is the $n \times 1$ vector of 1s

We are looking for the projection of $\underline{y}$ into span$\{\underline{1}\}$

We have just seen that $\dfrac{\underline{y} \cdot \underline{1}}{\underline{1} \cdot \underline{1}} = \hat{c}$

This is none other than $\bar{y}$

Note also that

$$\|\underline{y} - \hat{c}\underline{1}\| = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \sqrt{n}\, SD(\underline{y})$$

length = spread/variation in $\underline{y}$ that is not explained by $\bar{y}$

Geometric approach to simple linear regression

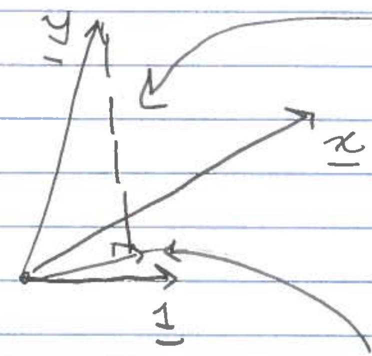
$$\min_{a,b} \sum_{i=1}^{n} \left(y_i - (a + b x_i)\right)^2 \qquad L^2 \text{ loss}$$

re write as length of a vector (squared)

$$\min_{a,b} \left\| \underline{y} - (a\underline{1} + \beta \underline{x}) \right\|^2$$

Looking for the projection of $\underline{y}$ into span$\{\underline{1}, \underline{x}\}$

Call this projection $\hat{\underline{y}}$



error: $\underline{y} - \hat{\underline{y}} := \underline{e}$ orthogonal to $\hat{\underline{y}}$

Residual: that part of $\underline{y}$ orthogonal to span$\{\underline{1}, \underline{x}\}$

$\hat{\underline{y}}$ lives in the span of $\underline{1}$ & $\underline{x}$

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

Note

$\bar{e} = 0$ because $\underline{e}$ is orthogonal to $\underline{1}$

so $\underline{e} \cdot \underline{1} = 0$

$\underline{x} \cdot \underline{e} = 0$ again because $\underline{e}$ is orthogonal to $\underline{x}$

and

$\hat{\underline{y}} \cdot \underline{e} = 0$ residuals are orthogonal to the fitted values

Total Sum of squares

$$\|\underline{y} - \bar{y}\underline{1}\|^2 = \|\underline{y} - \hat{\underline{y}} + \hat{\underline{y}} - \bar{y}\underline{1}\|^2$$

$$= \|\underline{y} - \hat{\underline{y}}\|^2 + \|\hat{\underline{y}} - \bar{y}\underline{1}\|^2$$

cross product is 0 Why?

Error Sum of Squares

Regression Sum of Squares

Geometric Approach to Multiple Linear Regression

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) \right)^2$$

Re express

$$\underset{n\times 1}{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}} - \underset{n\times(p+1)}{\begin{pmatrix} 1 & x_{11} & x_{21} & & x_{p1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & & x_{pn} \end{pmatrix}} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$(p+1)\times 1$

$$\min_{\underline{\beta}} \left\| \underline{y} - \mathbb{X} \underline{\beta} \right\|^2$$

design matrix $[\underline{1}, \underline{x}_1, \ldots \underline{x}_p]$

$n \times 1$ column vectors

The minimizing value is $\hat{\underline{\beta}}$ which we get by projecting $\underline{y}$ onto the span of $\mathbb{X}$

So we know

$$\underline{y} = \underbrace{X\hat{\underline{\beta}}}_{\hat{\underline{y}}} + \underline{e}$$

error that is orthogonal to span$(\mathbb{X})$

pre-multiply by X^t

$$X^t y = X^t X \hat{\underline{\beta}} + \underbrace{X^t \underline{e}}$$

This must be 0 because $\underline{e}$ is orthogonal to X, $\underline{1}\cdot \underline{e} = 0$, $\underline{x}_1 \cdot \underline{e} = 0$ etc

$$\underset{(p+1)\times 1}{X^t \underline{y}} = \underset{(p+1)\times(p+1)}{X^t X}\, \underset{(p+1)\times 1}{\hat{\underline{\beta}}}$$

These are called normal equations

Solve for $\hat{\underline{\beta}}$ by multiplying by $(X^t X)^{-1}$

$$\hat{\underline{\beta}} = (X^t X)^{-1} X^t \underline{y}$$

Observations

(A) We can only solve for $\hat{\beta}$ if X^tX has an inverse

The columns of X cannot be collinear

That is we cannot have variables that are linear combinations of others

(B) If $p+1 > n$ then we have a problem - too many variables so we will not be able to solve for $\hat{\beta}$

(C) The $\hat{\beta}$ coefficients depend on the presence of the other variables in the model.

Fit a simple linear model of y to x_1 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and a multiple linear model of y to $\underline{x}_1$ and $\underline{x}_2$

$$\hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1 \underline{x}_1 + \hat{\gamma}_2 \underline{x}_2$$

$\hat{\gamma}_1 \neq \hat{\beta}_1$ unless $\underline{x}_1$ & $\underline{x}_2$ are orthogonal

(D) Decomposition of Total SS still works

$$\text{Total SS} = \|\underline{y} - \bar{y}\underline{1}\|^2$$

$$= \underbrace{\|\underline{y} - \hat{\underline{y}}\|^2}_{\text{Error SS}} + \underbrace{\|\hat{\underline{y}} - \bar{y}\underline{1}\|^2}_{\text{Reg SS}}$$

multiple $R^2 := \text{RegSS}/\text{Tot SS}$

(E) If columns in X are highly collinear then the $\hat{\beta}$ can have a high standard error because other $X\hat{\gamma}$ can be close to $\hat{y}$ and $\hat{\beta} - \hat{\gamma}$ can be large

How does this relate to our scatter plots?