Fitting Linear models by least squares

Given TS $(x_1, y_1) \dots (x_n, y_n)$

Suppose association looks linear

$\Rightarrow$ fit straight line $\ell(x) = b_0 + b_1 x$

How to find b_0, b_1 ?

Minimize resubstitution error:

$$(\hat{b}_0, \hat{b}_1) = \underset{b_0\ b_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \quad (*)$$

Choose b_0 and b_1 that give the best predictive performance for training sample.

Differentiate $(*)$ wrt b_0 and b_1

$$\Downarrow$$

$$\hat{b}_1 = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum (x_i - \bar{x})^2$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

Generalize to more than 1 predictor variable

$$\ell(\underline{x}) = b_0 + b_1 x_1 + \dots b_p x_p$$

Fitting a linear model as a linear algebra problem

Notation

$$\underline{a} = (a_1, \dots a_p)^T$$

For X $n \times p$ matrix

- $\underline{x}_i$ is i-th row of X
- X^j is j-th column of X

Inner product of $\underline{a}$ and $\underline{b}$ $\langle \underline{a}, \underline{b} \rangle =$

$= \sum_i a_i b_i$

$\|\underline{a}\|^2 = \langle \underline{a}, \underline{a} \rangle = \sum a_i^2$ squared norm of $\underline{a}$

$\langle \underline{a}, \underline{b} \rangle = \|a\| \, \|b\| \cdot \cos \sphericalangle (\underline{a}, \underline{b})$

$\underline{a}$ and $\underline{b}$ are called orthogonal $\underline{a} \perp \underline{b}$ if $\langle \underline{a}, \underline{b} \rangle = 0$

First consider simplest regression problem: Fitting a line through the origin

Find b such that $\sum_i (y_i - b x_i)^2 = \min !$

Differentiate wrt b:

$$\sum_i (y_i - b x_i) x_i = 0 \iff \langle \underline{y - b\underline{x}}, \underline{x} \rangle = 0$$

$\underline{r}$ residual vector

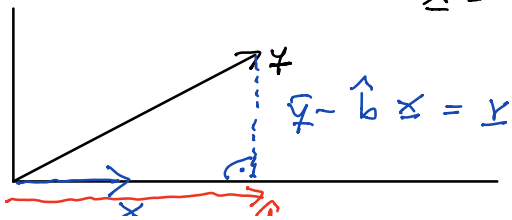
$$\Downarrow$$

$$\langle y, \underline{x} \rangle - b \langle \underline{x}, \underline{x} \rangle = 0$$

$$\hat{b} = \frac{\langle y, \underline{x} \rangle}{\|\underline{x}\|^2}$$

$y = (y_1, \ldots y_n)$

$\underline{x} = (x_1, \ldots x_n)$



$b\underline{x}$

Pythagoras $\|y\|^2 = \|b\underline{x}\|^2 + \|\underline{r}\|^2$

$$\|\underline{r}\|^2 = \|y\|^2 - \|b\underline{x}\|^2$$

$$= \|y\|^2 - \frac{\langle y, \underline{x}\rangle^2}{\|\underline{x}\|^4}\|\underline{x}\|^2$$

$$= \|y\|^2 - \frac{\langle \underline{x}, y\rangle^2}{\|\underline{x}\|^2}$$

Moving on: multiple predictor variables

X: design matrix

X has n rows

p colums $\quad p$ = number of predictor variables + 1

$TS: (\underline{x}_1, y_1) \ldots (\underline{x}_n, y_n)$

$$X: n\begin{bmatrix} \vdots & \underline{x}_1 \\ \vdots & \underline{x}_2 \\ \vdots & \vdots \\ \vdots & \underline{x}_n \end{bmatrix}$$

$\|\underline{r}\|^2 = \|y - X\underline{b}\|^2$ want to minimize

$\hat{\underline{b}} = \underset{\underline{b}}{\operatorname{argmin}} \|y - X\underline{b}\|^2$

Clearly $(y - X\hat{\underline{b}})$ has to be orthogonal to $X^1 \ldots X^p$

Suppose X^i was not orthogonal to

$y - X\underline{b}$ Then

we could take $\underline{y}$ and regress it on X^i and reduce the residual sum of squares

In matrix form

$\therefore X^T(y - X\hat{b}) = \underline{0}$ → residual vector $\perp$ to columns of X.

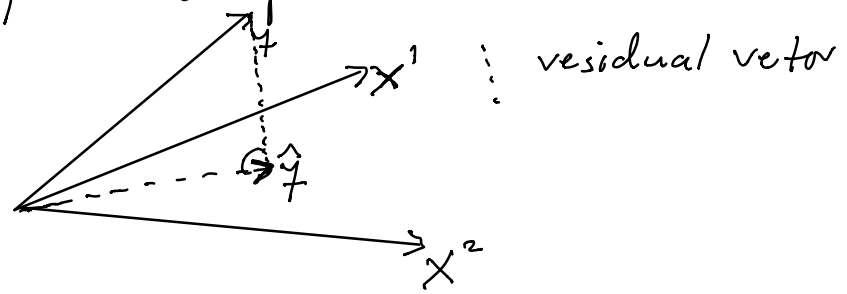
$X^T y - X^T X \hat{b} = 0$ or

$\hat{b} = (X^T X)^{-1} X^T y$ "normal equations"

Define $\hat{y} = X\hat{b}$ vector of predicted values

$r = y - \hat{y} \perp [X^1 \dots X^p]$ space spanned by $X^1 \dots X^p$

Picture for $n = 3$ $p = 2$



$\Rightarrow \hat{y}$ = projection of y on $[X^1 \dots X^p]$

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

$= H$. hat matrix

H is a projection: H is symmetric $H^T = H$ and idempotent $H^2 = H$

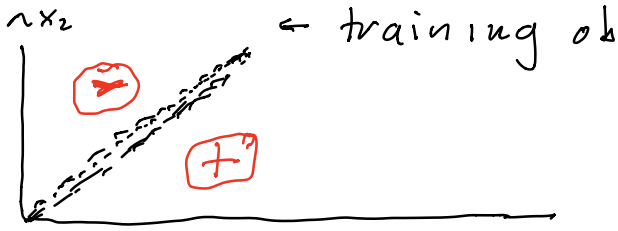
Verify idempotence:

$$H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$$

$$= X(X^TX)^{-1}X^T = H$$

Note: If rank $X < p \Rightarrow$ problem

Why: rank (X^TX) = rank $(X) \Rightarrow (X^TX)^{-1}$ does not exist $\Rightarrow \hat{b}$ is not unique.



X^1 X^2 linearly dependent

Coefficient vector $\hat{b}$ not unique but $\hat{y}$ = projection of y on $X^1 \ldots X^p$ always unique.

If rank $(X) < p$ then $\hat{y}$ is unique but there are infinitely many ways to represent $\hat{y}$ as a linear combination of $X^1 \ldots X^p$

Handling categorical predictors

X is predictor variable with values $a_1 \ldots a_k$ categorical.

Convert X into $(K-1)$ indicator variables $Z^1 \ldots Z^{K-1}$

$z_i^j = 1$ if $x_i = a_j$

z_i^j: value of z^j for observation i

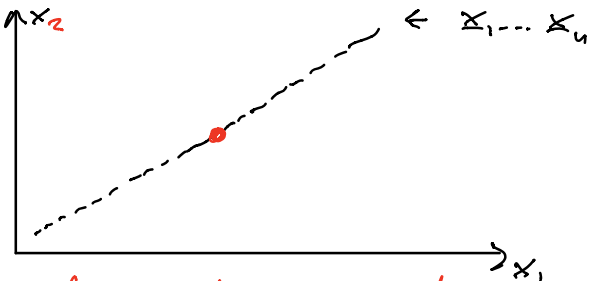
Why $K-1$? If we had K indicator variables, then $\sum(z^1 + z^2 \ldots + z^K) = \underline{1}_n$

But design matrix already has column of $\underline{1} \Rightarrow X$ would be rank deficient

Interpreting regression coefficients

$$y(\underline{x}) = \langle \hat{\underline{b}}, \underline{x} \rangle = \hat{b}_1 x_1 + \ldots \hat{b}_p x_p$$

Interpretation of b_i = rate of change in $y(\underline{x})$ if x_i is changed but the values of all other predictors are held fixed.



No information in data about what would happen to y if I changed x_1 but x_2 was held fixed.