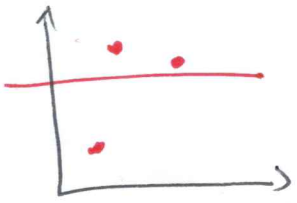
If $n = p+1 \Rightarrow X$ is square.

(no linear dpdt.)

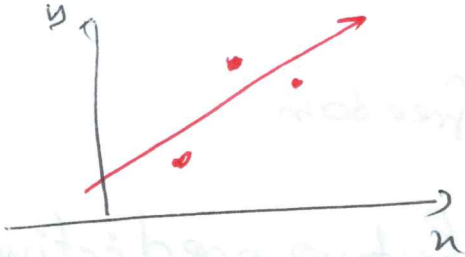
$\Rightarrow H = X(X^TX)^{-1}X^T = X X^{-1}(X^T)^{-1}X^T = I$

$\vec{\hat{y}} = H\vec{y} = I\vec{y} = \vec{y} \Rightarrow \vec{e} = \vec{y} - \vec{\hat{y}} = \vec{0} \Rightarrow SSE = 0 \Rightarrow R^2 = 100\%$



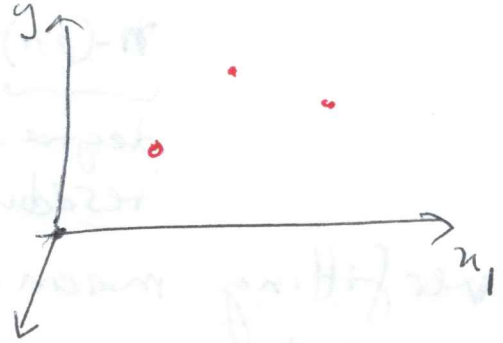
d.o.f = 1

"underfit"



d.o.f = 2

"fit well"



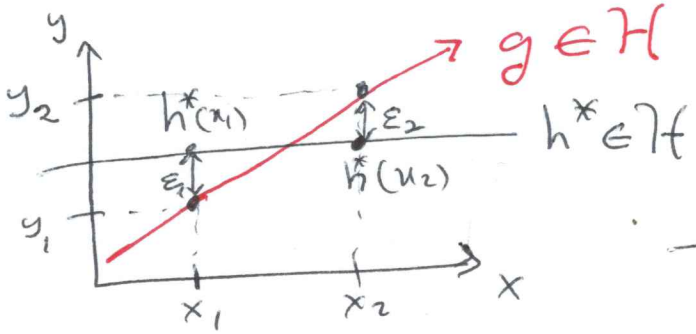
"overfit"

$y = h^*(\vec{x}) + \varepsilon$

$h^* \in \mathcal{H}$ (best model in ~~d~~ candidate set given

predict:

y. has two components $h^*(\vec{x})$, ε (sigma, noise)

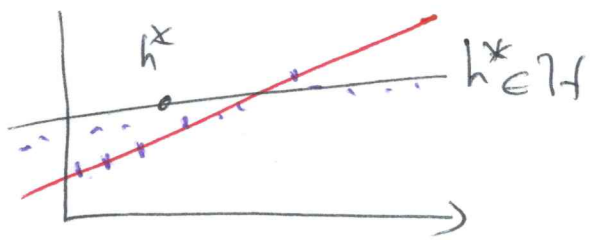


$g \in \mathcal{H}$

$h^* \in \mathcal{H}$

overfit

The overfix model g doesn't generalize (doesn't fit random well)



$h^* \in \mathcal{H}$

$$y = g(\vec{x}) + e$$

$$= g(\vec{x}) + \underbrace{(h^* - g)}_{\text{estimate error (this includes originary error}} + \varepsilon$$

$n \to \infty$ ∃" estimation error $\longrightarrow 0$

and p constant.

Note: $MSE = \frac{1}{\underbrace{n-(p+1)}_{\text{degree of freedom residual}}} SSE$

Overfitting means future predictive performance

You only see overfitting iff you never see the data before

*g and h^* are the same based on two observation