3/26/18

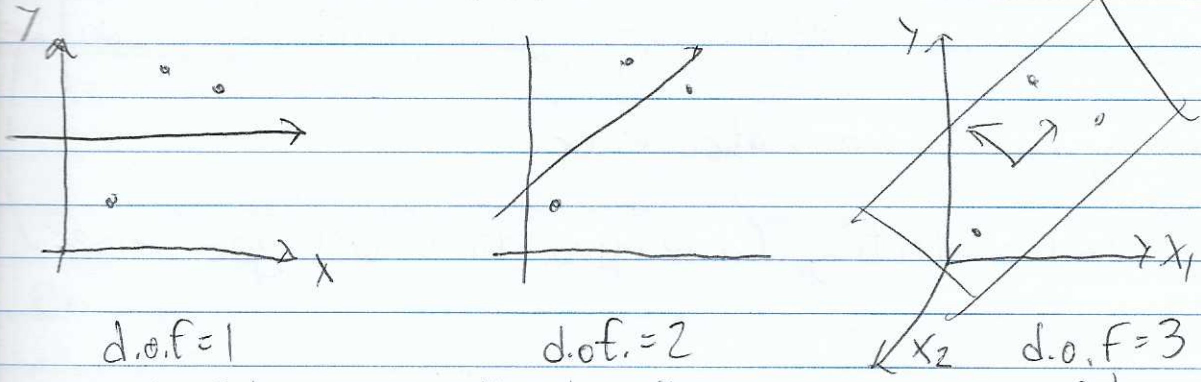
Lecture 14

If $n = p+1 \Rightarrow X$ is square (and no linear dependence)

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

$\hat{\vec{y}} = H\vec{y} = I\vec{y} = \vec{y} \Rightarrow \vec{e} = \vec{y} - \hat{\vec{y}} = \vec{0}$

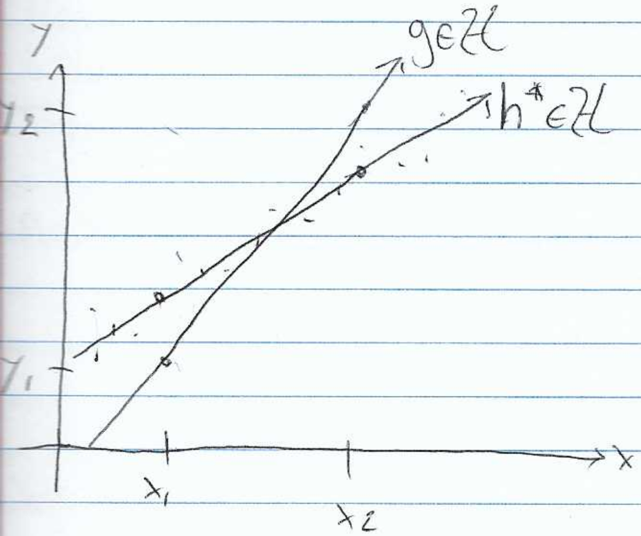
$\Rightarrow SSE = 0 \Rightarrow R^2 = 100\%$



d.o.f = 1 underfit | d.o.f. = 2 "fit well" | d.o.f = 3 overfit

$y = h^*(\vec{x}) + \varepsilon$ $h^* \in \mathcal{H}$ best model in candidate set given predictions at hand.

y has two components $h^*(\vec{x}), \varepsilon$ (signal, noise)



overfit

The overfit model g doesn't "generalize" ("doesn't fit new data well")

$y = g(\vec{x}) + e = g(\vec{x}) + \underbrace{(h^* - g)} + \varepsilon$

estimation error

this indicates overfitting error

If $n \to \infty$ and $p = \text{constant} \Rightarrow \text{estimation error} \to 0$

Overfitting $\Rightarrow$ future predictive performance suffers

DRY (Don't Repeat Yourself)

Overfitting is fitting too much ε and not enough h^*

$$RMSE = \sqrt{\frac{1}{n-(p+1)} SSE}$$

You can only see overfitting on data you've never seen before.

Overfitting error is estimation error

If I start overfitting, future predictions will get worse.