Math 342W Lecture 12

$$\begin{tikzcd} \boxed{X} \arrow[r, bend left, "G.S", "R^{-1}"'] & \boxed{Q} \arrow[l, bend left, "R"] \end{tikzcd}$$

Not a function of X

$$\overset{①}{SST} = \overset{②}{SSR} + \overset{③}{SSE}$$

① $\sum (y_i - \bar{y})^2$

② $\sum (\hat{y}_i - \bar{y})^2$

③ $\sum e_i^2$

SSR ↑, SSE ↓

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

$$= \sum 2\hat{y}_i - 2\hat{y}_0 \bar{y} + \bar{y}^2$$

$$= \|\vec{\hat{y}}\|^2 - 2\bar{y} \sum \hat{y}_i + n\bar{y}^2$$

$$= \|\vec{\hat{y}}\|^2 - 2\bar{y}\, \vec{\hat{y}}^T \vec{1}_n + n\bar{y}^2$$

$$= \|\vec{\hat{y}}\|^2 - 2\bar{y} (H\vec{y})^T \vec{1}_n + n\bar{y}^2$$

$$= \|\vec{\hat{y}}\|^2 - 2\bar{y}\, \vec{y}^T H \vec{1}_n + n\bar{y}^2$$

$$= \|\vec{\hat{y}}\|^2 - 2\bar{y} (n\bar{y}) + n\bar{y}^2$$

$$= \|\vec{\hat{y}}\|^2 - n\bar{y}^2$$

$$= \|\text{proj}_{\text{col}[X]}(\vec{y})\|^2 - n\bar{y}^2$$

$$= \|\text{proj}_{\text{col}[Q]}(\vec{y})\|^2 - n\bar{y}^2$$

$$= \left\|\sum_{j=0}^{p} \text{proj}_{(\vec{q}_j)}(\vec{y})\right\|^2 - n\bar{y}^2$$

$$= \sum_{j=0}^{p} \|\text{proj}_{(\vec{q}_j)}(\vec{y})\|^2 - n\bar{y}^2$$

* $\vec{q}_0$ is special because it is the projection onto $\vec{1}$

$$= \|\text{proj}_{\vec{q}_0}(\vec{y})\|^2 + \sum_{j=1}^{p} \|\text{proj}_{(\vec{q}_j)}(\vec{y})\|^2 - n\bar{y}^2$$

$$= \|\text{proj}_{\vec{1}_n}(\vec{y})\|^2 + \sum_{j=1}^{p} \|\text{proj}_{(\vec{q}_j)}(\vec{y})\|^2 - n\bar{y}^2$$

$$= \|\bar{y}\vec{1}_n\|^2 + \sum_{j=1}^{p} \|\text{proj}_{(\vec{q}_j)}(\vec{y})\|^2 - n\bar{y}^2$$

↳ pull out $\bar{y}$

$$= \cancel{n\bar{y}^2} + \sum_{j=1}^{p} \|proj_{(\vec{q}_j)}(\vec{y})\|^2 - \cancel{n\bar{y}^2}$$

$$\boxed{= \sum_{j=1}^{p} \|proj_{(\vec{q}_j)}(\vec{y})\|^2 = SSR}$$

$$X_{new} = \begin{bmatrix} X & \vec{x}_{new} \end{bmatrix} \Rightarrow$$ X_{new} is full rank which means no columns are linearly dependent

Q_{new} is now also full rank.

Now, $SSR_{new} = \sum_{j=1}^{p} \|proj_{(\vec{q}_j)}(\vec{y})\|^2 + \|proj_{(\vec{q}_{new})}(\vec{y})\|^2 = SSR + \|proj_{(\vec{q}_{new})}(\vec{y})\|^2$

$\Rightarrow SSR_{new} > SSR, \quad R^2_{new} > R^2, \quad RMSE_{new} < RMSE$

This is horrible b/c now we can add <u>random</u> garbage and our model becomes more accurate

$$X_{full} = \begin{array}{c|c} & \begin{array}{cc} p+1 & (n-(p+1)) \end{array} \\ n & \begin{bmatrix} X & \vec{x} & \vec{x} & \cdots & \vec{x} \end{bmatrix} \end{array} \Rightarrow \in \mathbb{R}^{n \times n}, \text{ full rank}$$

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

$\vec{\hat{y}} = H\vec{y} = \vec{y} \approx$ PERFECT FIT TO DATA

This is not good at all because we can't extend to future data

$$\hat{y} = \text{proj}_{X_{new}}(\vec{y}) = \sum_{j=0}^{p} \text{proj}_{\vec{x}_j}(\vec{y}) + \sum_{n=0}^{p_{new}} \text{proj}_{\vec{x}_n}(\vec{y})$$

$\Rightarrow$ leads to overfitting

Our performance metrics R^2, SSE, RMSE, etc are now considered in sample metrics and not to be trusted

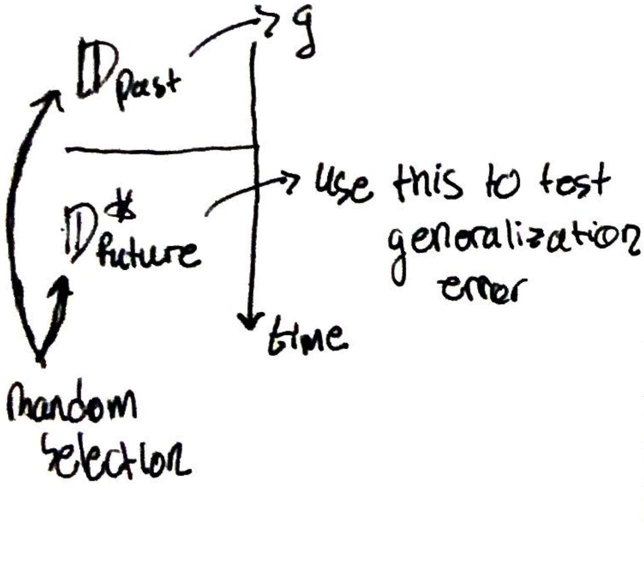
Sample $= \mathbb{D}$, n observations of $\langle \vec{x}_i, y_i \rangle$

We need honest performance metrics for "generalization error" or "generalization accuracy"

Generalization: using g (your model) on $\vec{x}^*$'s

We don't know what future y^*'s are....

$\triangleright$ we assume "stationarity" which means $y = t(\vec{z})$ is constant and the relationship between the x's and the z's is constant.



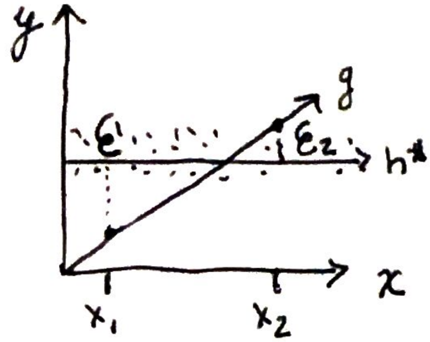
$e^* = y^* - \hat{y}^*$ for all $\vec{x}^*$ in $\mathbb{D}_*$

$\hookrightarrow$ This is an honest out of sample residual with n^* observations (oos residual)

oos $R^2 = SSR_* / SST_*$

oos RMSE $= \sqrt{\sum_{i=1}^{n^*} (y_{*i} - \hat{y}_{*i})^2 / n_*}$

Example: $p=1$, $n=2$



$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix} = I_2 \Rightarrow$ Perfect Fit "Overfit"

$y = h^*(x) + \varepsilon$ → New data is near h^*

Overfitting no effect on h^*, can only make g worse

Canonical Diagrams