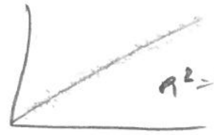
Lec 9 Math 390 2/20/18

RMSE vs R^2 — what is more important for assessing predictive ability? RMSE (usually) Why?

RMSE answers the question: how good are my predictions? what is the std err of my predictions?

$R^2 \uparrow \Rightarrow$ RMSE $\downarrow$

$R^2 \downarrow \Rightarrow$ RMSE $\uparrow$



$R^2 = 99\%$ How can this be bad?

$\rightarrow$ BAD!!

If $R^2 = 99\%$, could RMSE still be big? Yes. Maybe there was a ton of variance in y, you explained most of it, but there is still a lot left...

RMSE $\approx s_e$ (a corrected std dev. in residuals)

Empirical Rule

$\hat{y} \pm 2 \cdot s_e \approx 95\%$ of all predictions $\left(\text{if } \mathcal{E} \sim N(0, \sigma^2)\right)$

$\hat{y} \pm 3 \cdot s_e \approx 99.7\%$ " " " " " " (" " " ")

What if you're doing regression i.e. $y \in \mathbb{R}$ and $p = 1$ but the feature is a factor with two levels.

$x \in \mathcal{X} = \{\text{Red}, \text{Green}\}$ How to model? Try linear model

let red $\rightarrow 0$, green $\rightarrow 1$ and create a binary variable $\tilde{x} \in \{0, 1\}$.

What would the hyp. set look like?

$$\mathcal{H} = \{ w_0 + w_1 \tilde{x} : w_0 \in \mathbb{R}, w_1 \in \mathbb{R} \} = \{ w_0 + w_1 \mathbb{1}_{x = \text{Green}} : w_0 \in \mathbb{R}, w_1 \in \mathbb{R} \}$$

so the final model $\hat{y} = b_0 + b_1 \tilde{x} = b_0 + b_1 \mathbb{1}_{x = \text{green}}$

You can fit this with least squares. What do you think the answer will be?

$$\hat{y} = \begin{cases} \bar{y}_{red} & \text{if } x = red \\ \bar{y}_{green} & \text{if } x = green \end{cases} = \underbrace{\bar{y}_{red}}_{b_0} + \underbrace{\overbrace{(\bar{y}_{green} - \bar{y}_{red})}^{\Delta \bar{y}}}_{b_1} \mathbb{1}_{x = green}$$

red is called the ref. level/categ.

b_1 represents the added effect of green over red.

Let's see if we can prove this. Let $p = \frac{\sum \mathbb{1}_{x_i = green}}{n}$ i.e. prop of green

$(1-p) =$ prop of red

$b_0 = \bar{y} - b_1 \bar{x}$

Assume $b_1 = \bar{y}_g - \bar{y}_r$

$$\bar{y} = \frac{y_1 + \dots + y_n}{n} = \frac{\overbrace{y_1 \dots y_{n_G}}^{green}}{n} + \frac{\overbrace{y_1 \dots y_{n_R}}^{red}}{n} = \frac{\bar{y}_g n_g}{n} + \frac{\bar{y}_r n_r}{n} = p\bar{y}_g + (1-p)\bar{y}_r$$

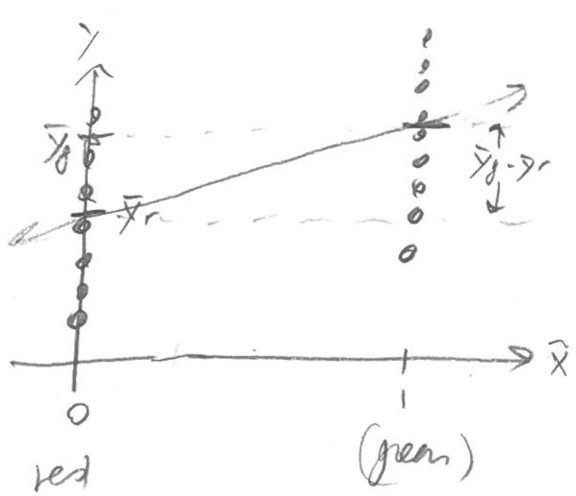
$\bar{x} = p$

$$\Rightarrow b_0 = p\bar{y}_g + (1-p)\bar{y}_r - p(\bar{y}_g - \bar{y}_r) = (1-p)\bar{y}_r + p\bar{y}_r = \bar{y}_r \checkmark$$

$$b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum_g (1) y_i + \cancel{\sum_r (0) y_i}^{0} - np\bar{y}}{\sum_g 1^2 + \cancel{\sum_r 0^2}^{0} - np^2} = \frac{\overset{pn}{\overset{\parallel}{n_g}}\bar{y}_g - np\bar{y}}{\underset{\underset{np}{\parallel}}{n_g} - np^2}$$

$$= \frac{\bar{y}_g - \bar{y}}{1-p} = \frac{\bar{y}_g - (p\bar{y}_g + (1-p)\bar{y}_r)}{1-p} = \frac{\cancel{(1-p)}\bar{y}_g - \cancel{(1-p)}\bar{y}_r}{\cancel{1-p}} = \bar{y}_g - \bar{y}_r \checkmark$$

What does this look like?



What if there were more than 2 levels in the factor?

$x \in \{red, green, blue\}$

Recall our idea from early in the class:

$X \Rightarrow \{\tilde{X}_1, \tilde{X}_2, \tilde{X}_3\}$ one variable becomes 3 dummy variables ("i.e. dummification")

$X_1 = \mathbb{1}_{x = red}$

$X_2 = \mathbb{1}_{x = green}$

$X_3 = \mathbb{1}_{x = blue}$

Can we fit a model on $y \sim x$. No, not without dummifying.

This leads us into "multivariate linear regression".

Let's remind ourselves of the modeling setup:

let's say we have $\mathbb{D}$ with p features and normally we used $\mathcal{H}$ to be

$$\mathcal{H} = \{w_0 + w_1 x_1 + \ldots + w_p x_p : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}, \ldots, w_p \in \mathbb{R}\}$$

$$= \{\vec{w} \cdot \vec{x} : \vec{w} \in \mathbb{R}^{p+1}\} \quad \text{where } \vec{x} = [1 \; x_1 \ldots x_p]$$

Let's choose some $\mathcal{A}$, ordinary least squares:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum \left(y_i - (w_0 + w_1 x_{i1} + \ldots w_p x_{ip})\right)^2$$

$\underset{w_0 \in \mathbb{R}, w_1 \in \mathbb{R}, \ldots, w_p \in \mathbb{R}}{\text{argmin}} \; SSE \Rightarrow$ nightmare

Need better way

MIDTERM 1

MIDTERM 2