lecture 07

$\hat{y} = g(x) = \underbrace{\bar{y}_{red}}_{b_0} + \underbrace{(\bar{y}_{green} - \bar{y}_{red})}_{b_1} x$, let $n_g =$

$\sum x_i$, $p_g = \bar{x} = \frac{n_g}{n}$ $\quad n_r = n - n_g$

$\bar{y} = \frac{1}{n}(\sum y_i) = \frac{1}{n}(\underset{i:green}{\sum y_i} + \underset{i:red}{\sum y_i}) = \frac{\sum_G y_i}{n} \frac{n_g}{n_g} + \frac{\sum_R y_i}{n} \frac{n_r}{n_r}$

$= p_g \frac{\sum_G y_i}{n_g} + (\underbrace{1 - p_g}_{p_r}) \frac{\sum_R y_i}{n_r} = p_g \bar{y}_g + (1 - p_g) \bar{y}_r$

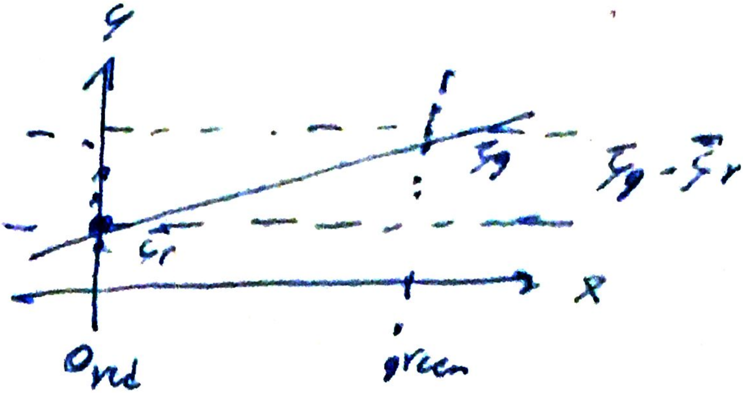
$b_1 = \frac{\overbrace{\sum x_i y_i}^{G} - n\bar{x}\bar{y}}{\underbrace{\sum x_i^2}_{n_g} - n \underbrace{\bar{x}^2}_{p_g}} = \frac{n_g \bar{y}_g - n p_g \bar{y}}{n_g - n p_g^2} \cdot \frac{\frac{1}{n}}{\frac{1}{n}} =$

$\frac{p_g \bar{y}_g - p_g \bar{y}}{p_g - p_g^2} = \frac{\bar{y}_g - \bar{y}}{1 - p}$

$= \frac{\bar{y}_g - (p_g \bar{y}_g + (1 - p_g)\bar{y}_r)}{1 - p_g} = \frac{(1 - p_g)\bar{y}_g - (1 - p_g)\bar{y}_r}{1 - p_g}$

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

$b_0 = \bar{y} - b_1 \bar{x} = p_g \bar{y}_g + (1 - p_g)\bar{y}_r - (\bar{y}_g - \bar{y}_r) p_g = \bar{y}_r$



What if $x \in \{red, green, blue\}$? This is then $p = 2$ and we need an OLS solution for $p \geq 1$. But intuitively.

$$g(x) = \begin{cases} \bar{y}_{red} \text{ if } x = red \\ \bar{y}_{green} \text{ if } x = green \\ \bar{y}_{blue} \text{ if } x = blue \end{cases} = \underbrace{\bar{y}_{red}}_{b_0} + \underbrace{(\bar{y}_{green} - \bar{y}_{red})}_{b_1} \overset{\mathbb{1}_{x=green}}{x_1} + \underbrace{(\bar{y}_{blue} - \bar{y}_{red})}_{b_2} \overset{\mathbb{1}_{x=blue}}{x_2}$$

How well does g predict? We need a "model performance metric". In the SVM this was accuracy or misclassification error. Here it will can also be what we use internally in the algorithm

$$SSE := \sum_{i=1}^{n} e_i^2 = \sum (y_i - g(x_i))^2$$

Is SSE interpretable? No, let's take the mean at least, call that mean squared error (MSE):

$$MSE = \frac{1}{n-2} SSE$$

But this is still in the squared unit of the phenomenon so it's still uninterpretable. We can take the squared root of MSE called root mean squared error (RMSE)

$s_e = RMSE = \sqrt{\frac{1}{n-1}\sum e_i^2} = \sqrt{MSE}$

RMSE is in the same unit as y (it's akin to the standard deviation of the residuals) & also from the CLT

$$[\hat{y}(x) \pm 1.96 \cdot RMSE]$$

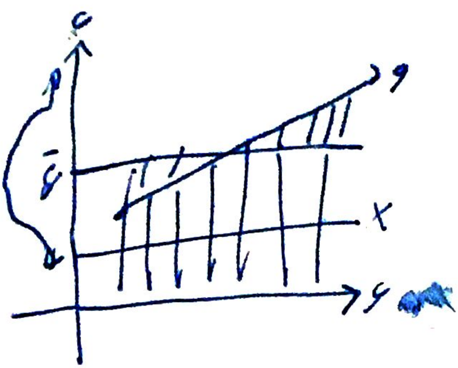
is approx a 95% confidence interval for the true y at that x. RMSE is a very important metric in regression models.

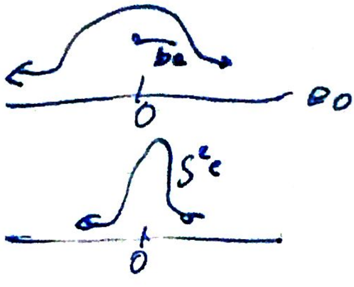
Another important error/performance metric is "R-squared", which is the "proportion of variance explained". We will now explore this definition

Consider the null model $g_0 = \bar{y}$. What is the SSE of this model? Let's call it SSE_0

$$SSE_0 = \sum_{i=1}^{n} e_{0,i}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SST = (n-1)s_y^2$$

Sum of Squares Total



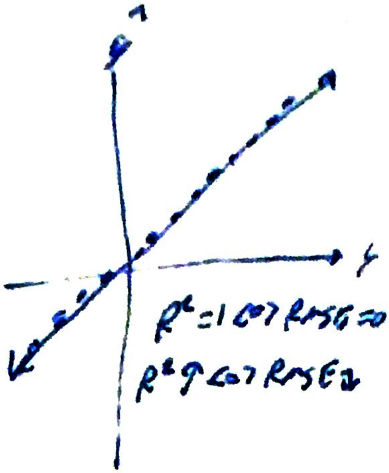


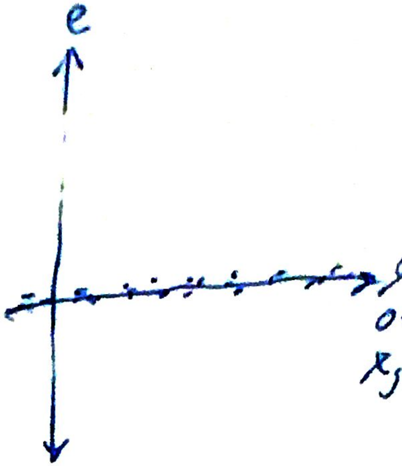
$$\frac{SSE}{SST} = \frac{(n-1)s_e^2}{(n-1)s_y^2} = \frac{s_e^2}{s_y^2}$$

$$R^2 = \frac{SST - SSE}{SST} = \frac{(n-1)s_y^2 - (n-1)s_e^2}{(n-1)s_y^2} = \frac{\overbrace{s_y^2 - s_e^2}^{\Delta s^2}}{s_y^2} = \frac{\Delta s^2}{s_y^2}$$

R-Squared can never be more than 100%. But R-Squared can be negative. This occurs when $S^2_e > S^2_y$ meaning the model is predicting worse than $g_0 = \bar{y}$

Here's some other useful plots especially when $p \geq 1$





If $R^2 = 99\%$, does this mean the model is for sure good? No because if the initial variance was SO very large, even a 99% reduction wouldn't result in a small residual variance i.e. RMSE still could be high after 99% variance reduction

We now would like to generalize the least squares estimation algorithm to cases where $p > 1$. Let's begin with $p = 2$.

$$H = \{ w_0 + w_1 x_1 + w_2 x_2 : \underbrace{w_0, w_1, w_2 \in \mathbb{R}}_{\vec{w} \in \mathbb{R}^3} \}$$

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - w_0 - w_1 x_{1,i} - w_2 x_{2,i})^2$$

$b_0 = \underset{w_0 \in R}{\arg\min}\{SSE\}$, $b_1 = \underset{w_1 \in R}{\arg\min}(SSE)$, $b_2 = \underset{w_2 \in R}{\arg\min}(SSE)$

The problem can be solved more easily with matrix algebra and a matrix arg min

$D = \langle X, \vec{y} \rangle$, let $X = [\vec{1}, \vec{x}_{\cdot 1}, \vec{x}_{\cdot 2}] =$

$$\begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ \vdots & \vdots & \vdots \\ & x_{n1} & x_{n2} \end{bmatrix}$$

e.g. $\hat{y}_1 = \vec{x}_1 \vec{w}$, $\hat{y}_2 = \vec{x}_2 \vec{w}$

$$\hat{\vec{y}} = X\vec{w} = \begin{bmatrix} w_0 + w_1 x_{11} + w_2 x_{12} \\ w_0 + w_1 x_{21} + w_2 x_{22} \\ \vdots \\ w_0 + w_1 x_{n1} + w_2 x_{n2} \end{bmatrix}$$

define $\vec{e} := \vec{y} - \hat{\vec{y}}$

$$SSE = \sum_{i=1}^{n} e_i^2 = \vec{e}^T \vec{e} = (\vec{y} - \hat{\vec{y}})^T (\vec{y} - \hat{\vec{y}}) = (\vec{y}^T - \hat{\vec{y}}^T)(\vec{y} - \hat{\vec{y}})$$

$$= \vec{y}^T \vec{y} - \hat{\vec{y}} + \vec{y} - \vec{y}^T \hat{\vec{y}} = \vec{y}^T \vec{y} - 2\hat{\vec{y}}^T \vec{y} + \hat{\vec{y}} + \hat{\vec{y}}$$

$$= \vec{y}^T \vec{y} - 2(X\vec{w})^T \vec{y} + (X\vec{w})^T X\vec{w} =$$

$$= \vec{y}^T \vec{y} - 2\vec{w}^T X^T \vec{y} + \vec{w}^T X^T X\vec{w}$$