lec#6

$$\hat{y} = g(x) = \underbrace{\bar{y}_{red}}_{b_0} + \underbrace{(\bar{y}_{green} - \bar{y}_{red})}_{b_1} x, \quad \text{let } n_g = \sum x_i, \quad P_g = \bar{x} = \frac{n_g}{n}$$

$n_r = n - n_g$

$$\bar{y} = \frac{1}{n}\left(\sum y_i\right) = \frac{1}{n}\left(\sum_{i=green} y_i + \sum_{i=red} y_i\right) = \frac{\sum_{i=green} y_i}{\cancel{n}} \cdot \frac{\cancel{n}_g}{n_g} + \frac{\sum_R y_i}{n} \cdot \frac{n_r}{n_r}$$

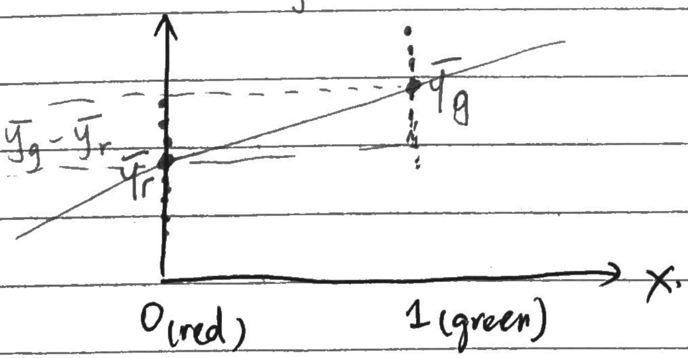
$$= P_g \frac{\sum_G y_i}{n_g} + (1-P_g)\frac{\sum_R y_i}{n_r} = P_g \bar{y}_g + (1-P_g)\bar{y}_r$$

$n_g \bar{y}_g = \sum_G y_i$

$$b_1 = \frac{\sum x_i y_i - n\overset{P_g}{\bar{x}}\bar{y}}{\underbrace{\sum x_i^2}_{n_g} - n\underbrace{\bar{x}^2}_{P_g}} = \frac{n_g\bar{y}_g - nP_g\bar{y}}{n_g - nP_g^2} \cdot \frac{\frac{1}{n}}{\frac{1}{n}} = \frac{\cancel{P_g}\bar{y}_g - \cancel{P_g}\bar{y}}{\cancel{P_g} - P_g^{\cancel{2}}} = \frac{\bar{y}_g - \bar{y}}{1-P_g}$$

$$= \frac{\bar{y}_g - P_g\bar{y}_g - (1-P_g)\bar{y}_r}{1-P_g} = \frac{\cancel{(1-P_g)}\bar{y}_g - \cancel{(1-P_g)}\bar{y}_r}{\cancel{1-P_g}} = \bar{y}_g - \bar{y}_r$$

$$b_0 = \bar{y} - b_1\underbrace{\bar{x}}_{P_g} = \cancel{P_g\bar{y}_g} + (1-\cancel{P_g})\bar{y}_r - (\cancel{\bar{y}_g} - \bar{y}_r)P_g = \bar{y}_r$$



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

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

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 square root of MSE called root mean squared error (RMSE)

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

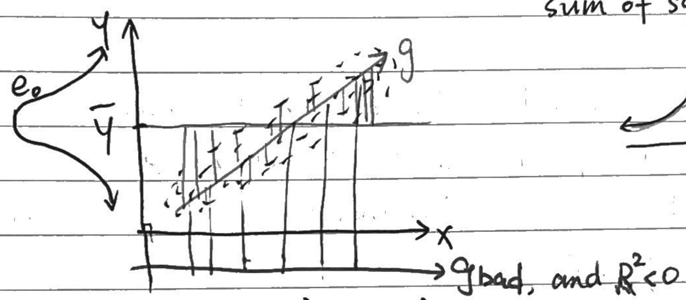
RMSE is in the same unit as y. (it is akin the s.d of the residuals s_e Also, from the CLT,

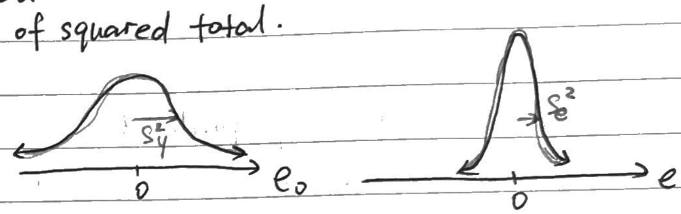
$[g(x) \pm 1.96 \cdot RMSE]$ is approx a 95% confidence interval for the true y at the x. RMSE is a very important metric in regression models.

Another important error / performance metric is "R-squared" which is the "proportion of variance explained".
Consider the null model $g_0 = \bar{y}$. What is the SSE of this model?
Let's call it SSE_0.

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

sum of squared total.



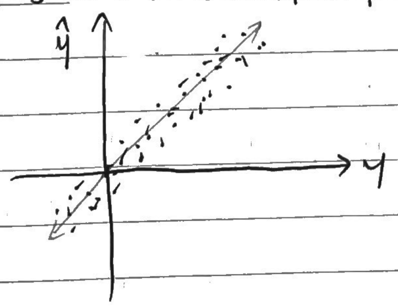


$$\frac{SSE}{SST} = \frac{(n-1) s_e^2}{(n-1) s_y^2} = \frac{s_e^2}{s_y^2}$$

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

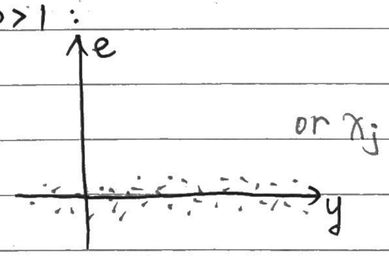
R^2 can never be more than 100%. But R^2 can be negative. This occurs when $s_e^2 > s_y^2$ meaning the model is predicting worse than $g_0 = \bar{y}$.

Here is some other useful plot especially when $p > 1$:



$R^2 = 1 \Leftrightarrow RMSE = 0$

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



Q: If R^2 is 99%, does this mean the model is for sure "good"?
No. Because if the initial variance was so large, even a 99% reduction wouldn't result 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 $p=2$.

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

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

$$b_0 = \underset{w_0 \in \mathbb{R}}{\text{argmin}} \{SSE\}, \quad b_1 = \underset{w_1 \in \mathbb{R}}{\text{argmin}} \{SSE\}, \quad b_2 = \underset{w_2 \in \mathbb{R}}{\text{argmin}} \{SSE\}$$

This problem can be solved more simply with matrix algebra and a metrix equation:

$$\mathbb{D} = \langle X, \vec{y} \rangle, \text{ Let } X = [\vec{1}_n \ \vec{x}_{\cdot 1} \ \vec{x}_{\cdot 2}] = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$$

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

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

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

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

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

$$= \vec{y}^T \vec{y} - 2 (X\vec{w})^T \vec{y} + (X\vec{w})^T X \vec{w} = \underbrace{\vec{y}^T}_{1 \times n} \underbrace{\vec{y}}_{n \times 1} - 2 \underbrace{\vec{w}^T}_{1 \times (P+1)} X^T \vec{y} + \vec{w}^T \underbrace{X^T}_{(P+1) \times n} X \vec{w}$$

$$\frac{\partial SSE}{\partial \vec{w}} := \begin{bmatrix} \frac{\partial SSE}{\partial w_0} \\ \frac{\partial SSE}{\partial w_1} \\ \vdots \\ \frac{\partial SSE}{\partial w_p} \end{bmatrix} \xrightarrow{\text{set}} \vec{0}_{p+1} \text{ and solve for } b_0, b_1, \ldots, b_p.$$