M342W 2/22

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

$$n_r = n - n_g$$

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

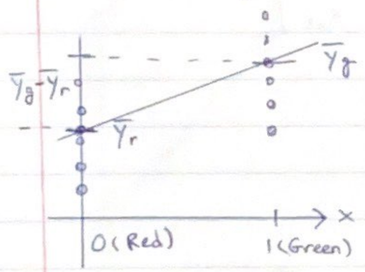
$$= p_g \frac{\sum_G y_i}{n_g} + (1-p_g)\frac{\sum_R y_i}{n_r} = p_g \bar{y}_g + (1-p_g)\bar{y}_r$$

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

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

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

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



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

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

How well does g predict? We need a "model performance metric". In the SVM this was accuracy or misclassification error. Here, it ~~will~~ can ~~be~~ 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?

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):

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

RMSE is in the same unit as y (just like the standard deviation is in the same unit as the random variable). Also, from the CLT,
$[g(x) \pm 1.96 \cdot RMSE]$
is approx a 95% confidence interval for the true y at that x?

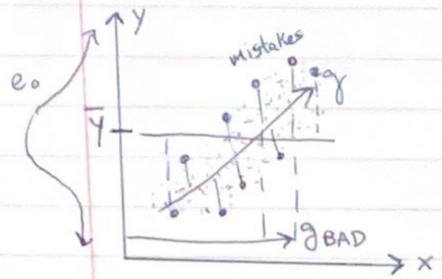
"standard deviation of the residuals" S_e

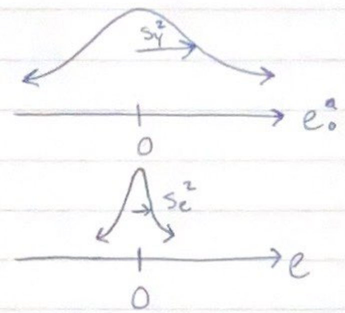
RMSE is a very important metric in regression models.

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

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

$$SSE_0 = \sum_{i=1}^{n} e_{0,i}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \underbrace{SST}_{\text{Sum of squares total}} = (n-1)s_y^2$$



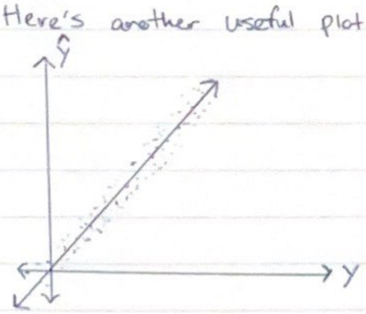


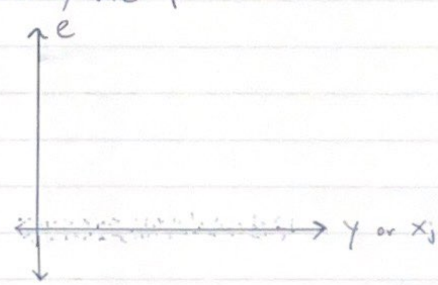
$$\frac{SSE}{SST} = \frac{(n-1)s_e^2}{(n-1)s_y^2} = \frac{s_e^2}{s_y^2}$$

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

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's another useful plot especially when $p > 1$:





$R^2 = 1$ implies $RMSE = 0$

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

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 $p = 2$.

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

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

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

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

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

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

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

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

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

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

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