Some useful quantities

$$SSE = RSS = \sum_{i=1}^{n} (y_i - \hat{f}(x))^2 \qquad R^2 = 1 - \frac{SSE}{S_{yy}} = 1 - \frac{RSS}{TSS}$$

$$S_{yy} = TSS = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

$$S_{xx} = \sum_{i=1}^{n} (x - \hat{x})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x - \hat{x})(y - \hat{y})$$

$$SSR = SSE - S_{yy} \Rightarrow S_{yy} = \underbrace{SSE}_{\text{irreducible error}} + \underbrace{SSR}_{\text{reducible error}}$$

Training and validation

In ML we don't really care about the past (statistically).

We split out data in two (or even three) sets. Eg: data = train | validation

Now we need comparison of means between samples (training set and validation set).

E_{MSE} : Expectation for the mean of SSE

n.b $MSE = \frac{1}{n} SSE$

$E_{MSE}(Y - \hat{Y})$: Expectation for difference of mean between Y and our approximation.

$$E_{MSE}(Y - \hat{Y})^2 = \underbrace{Var(\hat{Y}) + Bias(\hat{Y})^2}_{\text{reducible error}} + \underbrace{Var(\epsilon)}_{\text{irreducible error}}$$

RSE: $\sqrt{\frac{1}{n-2} SSE}$, unbiased estimator for unknown data!

RSE is run with the model we trained on <u>new</u> data (not even the validation).

Ie

train, validation, test

during training

(R^2, MSE)

↑

RSE for quality-checks

Feature-engineering

$$L(\lambda u + \mu v) = \lambda L(u) + \mu L(v)$$

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

i, Test significance of predictors, remove those that fail the test.

ii, Add new features (possibly from other data sources)

iii) Break the rules! Add non-linearity.

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

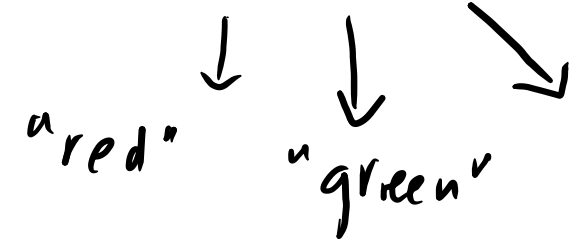
$$f(x_1, x_2, x_1 \cdot x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

x_3

Linearisation!

$\vec{f}(x) \Rightarrow$ numbers (real) regression

$\vec{f}(x) =$ labels (0, 1, 2) classification



labels are qualitative, categorical data