$$SD(x) = \sqrt{\mathrm{Var}(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{x}_i - \bar{\mathrm{x}}\right)^2}.$$

$$\mathrm{Corr}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

Multi-Regression

$$\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\\ \vdots\\ \hat{y}_n\end{bmatrix} = \begin{bmatrix}1 & x_{11} & x_{12} & \dots & x_{1p}\\ 1 & x_{21} & x_{22} & \dots & x_{2p}\\ 1 & x_{31} & x_{32} & \dots & x_{3p}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{n1} & x_{n2} & \dots & x_{np}\end{bmatrix}\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_p\end{bmatrix}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$R^2 = 1 - \frac{\sum(y_i-\hat{y})^2}{\sum(y_i-\bar{y})^2}$$

MAD = median(|preds - true|)

$$Correlation = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

Population Covariance Formula

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N-1}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y})^2}$$

Correlation Coefficient Formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

L2

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2.$$

L1

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - (b_0 + b_1 x_i)|.$$

Closed Form

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

$$b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

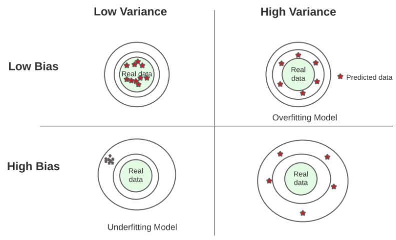
Logistic

$$p = \frac{1}{1+e^{-(b_0+b_1x)}}$$

Logit

$$\log\left(\frac{p}{1-p}\right)$$

odds ^ in parenth



Lasso

$$\sum_i(\text{observed price}_i - (b_0 + b_1\text{area}_i))^2 + \lambda(|b_0| + |b_1|)$$

Ridge

$$\sum_i(\text{observed price}_i - (b_0 + b_1\text{area}_i))^2 + \lambda(b_0^2 + b_1^2)$$

$$SSE_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{P}\beta_j^2$$

$$\frac{\hat{y}-\bar{y}}{\sigma_y} = r \times \frac{x-\bar{x}}{\sigma_x}$$

$$\hat{y} = \sigma_y \times r \times \frac{x-\bar{x}}{\sigma_x} + \bar{y}$$

$$\hat{y} = \boxed{(\frac{r\sigma_y}{\sigma_x})} \times x + \boxed{(\bar{y} - \frac{r\sigma_y}{\sigma_x}\bar{x})}$$

Recall regression line equation is defined as:

$$\hat{y} = \boxed{\hat{a}} + \boxed{\hat{b}}x$$

slope: $r\frac{SD\ of\ y}{SD\ of\ x} = r\frac{\sigma_y}{\sigma_x}$

intercept: $\bar{y} - slope \times \bar{x}$

Error for the i-th data point: $e_i = y_i - \hat{y}_i$

$$\begin{aligned} precision &= \frac{TP}{TP+FP}\\ recall &= \frac{TP}{TP+FN}\\ F1 &= \frac{2 \times precision \times recall}{precision + recall}\\ accuracy &= \frac{TP+TN}{TP+FN+TN+FP}\\ specificity &= \frac{TN}{TN+FP}\end{aligned}$$

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$$

CrossEntropy

$$H = -\sum p(x)\log p(x)$$

t-val from coeff

$$t_j = \frac{b_j}{SD(b_j)}$$

Some frequently used distance functions.	
Camberra: $d(x,y) = \sum_{i=1}^{m}\frac{\lvert x_i - y_i\rvert}{\lvert x_i + y_i\rvert}$ (2)	Euclidean: $d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$ (5)
Minkowsky: $d(x,y) = \left(\sum_{i=1}^{m}\lvert x_i - y_i\rvert^r\right)^{1/r}$ (3)	Manhattan / city - block: $d(x,y) = \sum_{i=1}^{m}\lvert x_i - y_i\rvert$ (6)
Chebychev: $d(x,y) = \max_{i=1}^{m}\lvert x_i - y_i\rvert$ (4)	

Bootstrap = sample w/ replacement to create more data

Classification

Prediction accuracy

$$\text{prediction accuracy} = \frac{(\text{number of correct predictions})}{n}$$

Prediction error

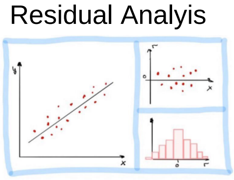
$$\text{prediction error} = \frac{(\text{number of incorrect predictions})}{n}$$

Decision Trees

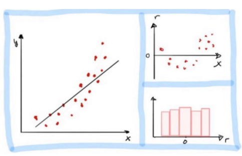
$$\text{Variance measure for split} = \frac{n_{\text{left}}}{n_{\text{parent}}}\text{Var}_{\text{left}} + \frac{n_{\text{right}}}{n_{\text{parent}}}\text{Var}_{\text{right}}$$

$$\text{Gini impurity for split} = \frac{n_{\text{left}}}{n_{\text{parent}}}\text{Gini}_{\text{left}} + \frac{n_{\text{right}}}{n_{\text{parent}}}\text{Gini}_{\text{right}}$$

Residual Analyis



Linear assumption is correct. There is no obvious relationship between residuals and x. Histogram of residuals is symmetric and normally distributed.



Linear assumption is incorrect. There is an obvious relationship between residuals and x. Histogram of residuals is symmetric but not normally distributed.