Regression-I

Linear Regression

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \beta_n x_n$

↳ for a given dataset (X_i, Y_i)

↳ closed form solution using Ordinary Least Squares (OLS) → $\sum_i (\hat{y}_i - y_i)^2$

↳ at minimum RSS (residual sum of squares)

Susceptible to Outliers

$\frac{\partial RSS}{\partial \beta_j} = 0 \rightarrow \hat{\beta} = (X^T X^{-1}) X^T Y = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \therefore Y = X\hat{\beta}$

$\{ y_i = \sum_j X_{ij} \beta_j$

Factor Variables {categorical}

↳ One hot Encoding: category with p values ↳ individual col^n of p-1 # {due to multicollinearity}

FV with multiple levels: subgroup similar categories {may acc to another var}

Ordered FV: can use single column with #s

Evaluations

→ RMSE $= \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{n}}$

Residual Standard Error (RSE) → $\sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{n-p-1}}$

{p = # of x_i dof}

→ R^2 statistics: how much of variation explained by model (0–1) ↳ $1 - \frac{RSS}{TSS}$ {$TSS = \sum (y - \bar{y})^2$}

↳ adjusted R^2 (incorporating dof) → $1 - \frac{(1-R^2)(n-1)}{n-p-1}$

→ t-statistics of coefficients: → $t_{b_i} = \frac{\hat{b}_i}{SE(\hat{b})}$ ↳ standard error $\left(\frac{\sigma}{\sqrt{n}}\right)$

↳ and p-value

↳ higher t_{b_i} ⇒ higher significant x_i

↳ can be used in feature selection

To penalize adding more x_i

$AIC = 2P + n \log\left(\frac{RSS}{n}\right)$

AIC ↓ Akaike's Information Criteria; P → # of predictors

{see also, BIC, AICc, Mallow C_p}

Feature Selection

↳ Forward Selection: ↻ Add x_i which has highest bump to RSS ↓ Stop: the ↑ in RSS is no more stat sig

Backward Selection: start: all features ↓ Next: drop the x_i with least t_{bi} ↓ Stop: all t_{b_i} (p-val) are stat significant

Penalized Regression: Ridge and Lasso regression

Multicollinearity

:→ having correlated features {no linear independence} {X_m dependent on $X_{i,i-n}$}

↳ makes β_i difficult to interpret

↳ inflates standard error

Confounding Variables

:→ Omission of some imp var from LR equation

↳ adverse eff on β_i and SE

Main effect

:→ relationship b/w y and x_j independent of other x_j

assumption that main effects are independent of each other → IF NOT

Interaction Effect

↳ Inter-dependence of y, x_i and x_j

↳ add interaction term $x_i x_j$ to LR eq^n

Confidence Interval

Confidence Interval: in the estimation of β_i

Prediction Interval: for the predicted $\hat{y}$

prediction interval usually much wider than confidence interval

Bootstrap method: N times
① get a bootstrap sample
② fit a LM
③ record β_i
④ record residue of some random sample $(\hat{y}_i - \hat{y})$