Goal → of Statistics is to infer f

$Y = f(X) + \epsilon$ → error term / bias.

dependent variable.

$X = (X_1, X_2, X_3, \ldots, X_p)$

↳ features

→ Linear Models good for inference, but often may not be the best predictors.

→ On contrast, some highly non-linear approaches will predict extremely well but are hard to interpret

Parametric Way → 1) Select a Model

eg → Linear Regression

2) Train the Model to estimate the parameters.

Non Parametric → 1) Do not assume a f^n.

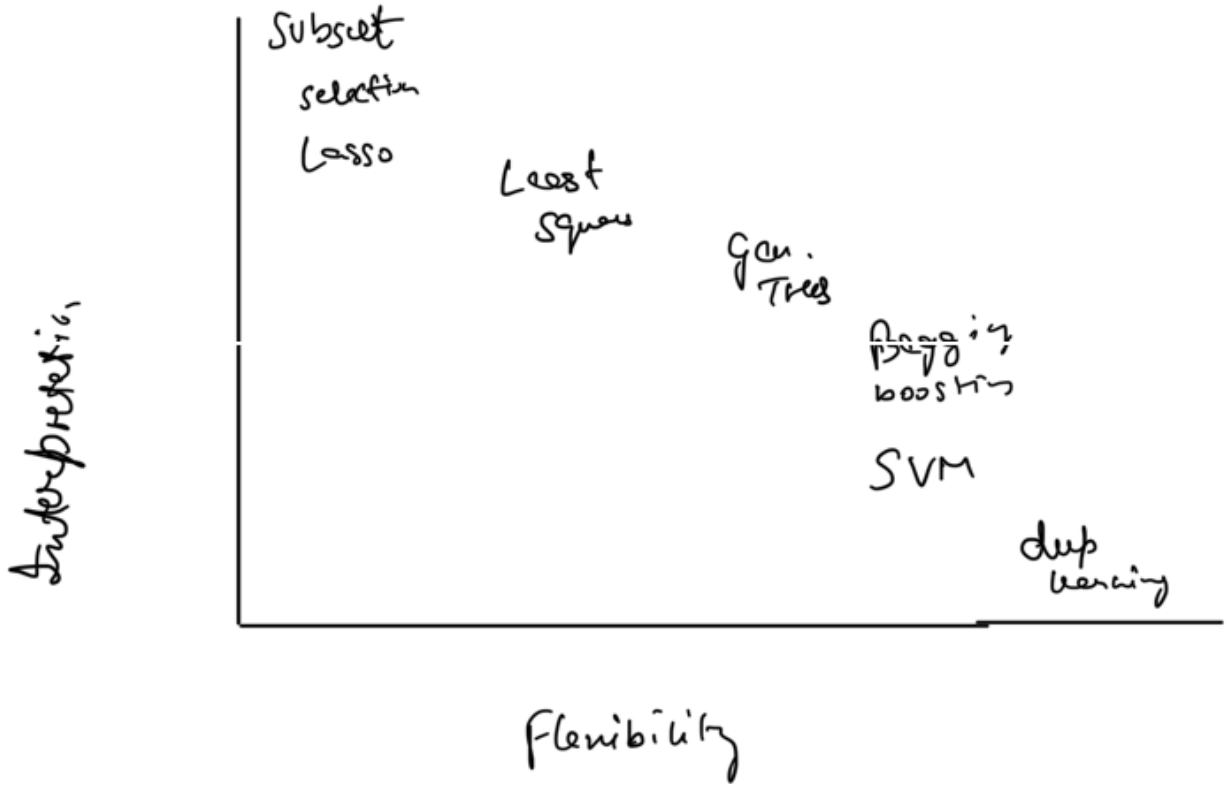
→ Observations

2) But lot more ...

eg → Thin-plate splines are needed to estimate.

but have to be cautions of overfitting.

⇒ are more flexible on the brighter side.



restricted models are good for inference

Unsupervised Learning → Categorising data without a response / output variable.

Clustering problems.

Regression v/s classification problems

variables → Quantitative
→ Qualitative (categorical)

Linear regression suitable for quantitative.

Assessing Model Accuracy →

* very important to select which method is suitable for the given problem.

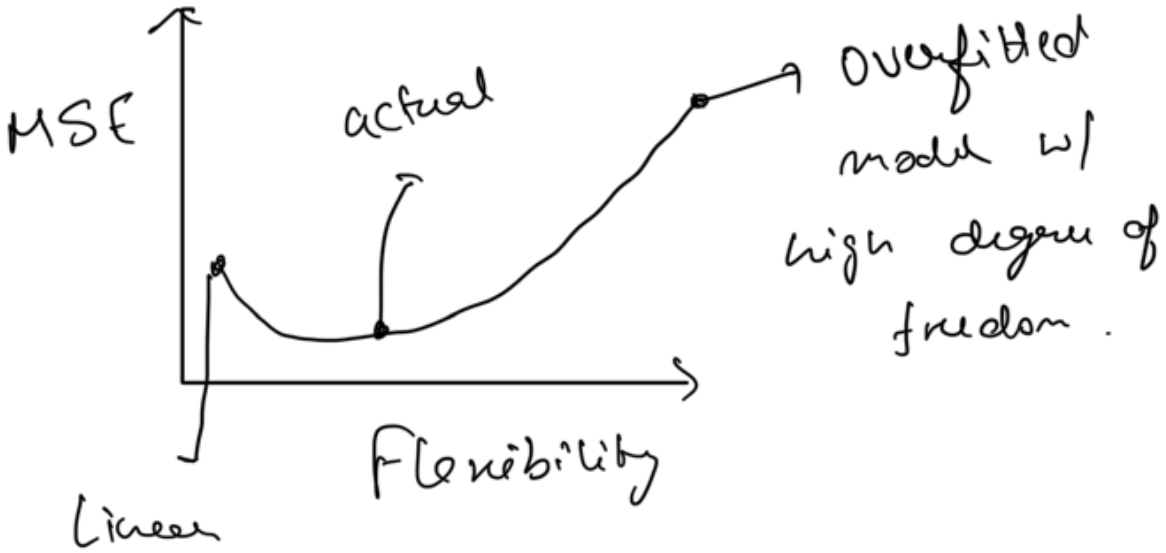
For Regression →

We use MSE (Mean squared error) to quantify the accuracy

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

MSE calculated on test data >>> ↑ IMP MSE on training data.

Some overfitted curves may give very less training MSE, but when tested with unseen data will give Larger MSE.



Overfitting → is the case when a linear model would have given less test MSE than high degree of freedom.

The Bias - Variance Trade off

The U-shaped curve of MSE is

the result of this trade off

MSE on test data

$$E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\varepsilon)$$

variance of $\hat{f}(x_0)$ | squared of Bias of $f(x_0)$ | variance of the error term

3 fundamental properties

Our aim to find something that reduces the variance and the bias.

{both terms $\text{var}(\hat{y})$ & $(\text{Bias}(y))^2$ are non -ve so the MSE can never be less than irreducible error}

for a more flexible f. the variance will be high as for different testing

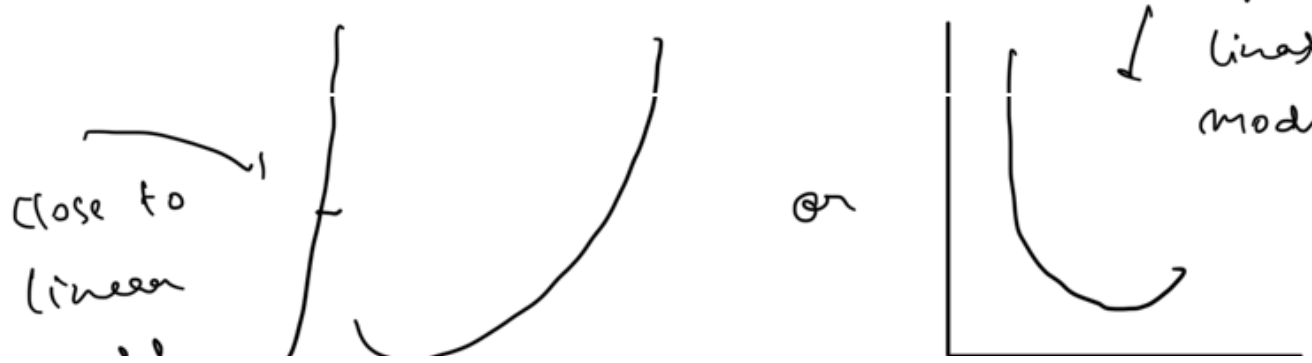
data, the output is highly varied.

Bias → error introduced by approximating a real life problem (with high complexity) by a simple model like linear regression

Simple Linear Regression = generally higher Bias.

if the approximated model is closer to the actual underlying model then the Bias is Low

More Flexible Model = Variance ↑ Bias ↓



model

we try and find the

Minima of this curve to have

low variance and Bias.

for qualitative data → (categorical variable)

MSE: $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$

↓

training error rate

Bayes Classifier

$P(Y = j / X = x_0)$

for a problem w/ 2 possible outcomes

$P_r(Y = 1 / X = x_0) > 0.5$

Bayes error rate $= 1 - E\left(\max_j Pr(Y = j / X)\right)$

↳ analogous to the irreducible error.

↳ this is basically the overlap.

K- nearest neighbours -

In reality we don't know the conditional relation $P_r(Y = j / X = x_0)$

So Bayes classifier is the Gold Standard.

given a +ve integer K we see

K nearest neighbours and assign

the class which most of the

neighbours are a part of

$$P_r(Y = j / X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

↳ class is assigned with the largest probability.

less K = more flexible

{ high variance
low Bias }

High K = less flexible (tends to linearity)

Choosing the correct level of flexibility is very imp.

Chapter - 3 Linear Regression

Statement/ Problem-1 → expenditure on advertisement v/s sales.

3 mods → news, TV, radio } out of 3 which is the best use

Interaction / Synergy effect

Spending 50K on TV and Radio instead of 100K on radio or TV.

We can solve this by Linear Regression

LR → single predictor variable X

$$Y = \beta_0 + \beta_1 X$$

eg → sales = $\beta_0 + \beta_1 \times TV$

β_0 ↓ intercept β_1 ↓ slope

Residual → $e_i = y_i - \hat{y}_i$

RSS → Residual sum of squares

$$RSS = e_1^2 + e_2^2 + \dots e_n^2$$

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

taking derivatives w.r.t. β_1 & β_0

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\dots}$$

$$\sum_{i=1} (x_i - \bar{x})$$

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$m = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

In general the sample mean $\bar{y}$ is a good estimator of the population mean μ.

over a large set of samples and taking mean of the mean it will converge to μ. Law of large numbers.

$$\text{var}(\hat{\mu}) = \frac{\sigma^2}{n}$$

as n increases the deviation from the mean decreases

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sigma^2 = \text{var}(\varepsilon)$$

estimate of σ is RSE

Residual Standard error

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

Confidence interval

CI of 95% for β_1

$$= \hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

$$\Rightarrow \beta_1 \in \left[\hat{\beta}_1 - 2\,SE(\hat{\beta}_1),\ \hat{\beta}_1 + 2\,SE(\hat{\beta}_1)\right]$$

Hypothesis Testing

most common : {null hypothesis}

$H_0 \rightarrow$ There is no relationship b/t X and Y

{alternate hypothesis}

H_a : There is some relationship b/t X and Y.

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq a$$

So β_1 must be sufficiently distant from 0.

how do we define | sufficiently distant |

We compute the t-statistic

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

this measures how many S.Ds away is our result from mean.

Small p value = There is no association

Hence we conclude $\beta_1 \neq 0$

and reject null hypothesis

Assessing the Accuracy of the model

1) Residual Standard Error
2) R^2 Statistic

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

R^2 statistic

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

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

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

if it is close to 0 that means that our prediction model is garbage

R^2 is the measure of linear relationship b/t X & Y

$$Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{SD(X)}\ \sqrt{SD(Y)}}$$

Multiple Linear Regression

in the case of Advertising.csv, we need a relation which accounts for all 3 factors.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

$$\text{sales} = \beta_0 + \beta_1 \times TV + \beta_2 \times news + \beta_3 radio$$

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

$$= \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip}\right)^2$$

doing multiple regression here reveals that newspaper advertisement

is garbage for brain

Some important questions.

1) Is at least one of the p predictors a good estimate.

2) Do all predictors help predict y or only a subset is useful

Null hypothesis

$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$

H_a : at least one β_j is $\neq 0$

hypothesis test is provided by computing the F-statistic

$$F = \frac{(TSS - RSS)/p}{\frac{RSS}{(n-p-1)}}$$

$$TSS = \Sigma (y_i - \bar{y})^2 \; ; \; RSS = \Sigma (y_i - \hat{y}_i)^2$$

F is close to 1 is H_0 is true

Q) how large does the F-statistic need to be before we reject it.

it follows F-distribution so we can find the p-value

Null Hypothesis $H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots$

$\cdots = \beta_p = 0$

{q of the co-efficients are zero}

$$F = \frac{(RSS_0 - RSS)/q}{\left(\frac{RSS}{n-p-1}\right)}$$

---

variable selection

Selecting a subset of p that actually matters.

there are 2^p subsets and we cannot consider every subset unless p is very small.

Methods

1) Forward selection : begin with a null model
   → then add the variable with lowest RSS
   → then continue till a stopping rule.

2) Backward selection → start with p features then reduce

3) Mixed.

Regression with qualitative variables

only have 2 levels

$$x_i = \begin{cases} 1 & \text{if yes} \\ 0 & \text{else} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \{ 1 \\ \beta_0 + \varepsilon_i & \{ 0 \end{cases}$$

β_0 = avg. value when $x = 0$

β_1 = avg. diff where $x = 1$ & $x = 0$ }

for 3 eg East, west, south

$x_1 \rightarrow$ south

$x_2 \rightarrow$ west

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$y = \begin{cases} \beta_0 + \beta_1 + \varepsilon & \rightarrow \text{south} \\ \beta + \beta_2 + \varepsilon & \rightarrow \text{west} \\ \beta + \varepsilon & \rightarrow \text{east} \end{cases}$$

Regression assumes linearity and additive nature

{basically superimposing ki one thing doesn't effect the output of others}

But there is synergy effect in marketing in stats it is known as interactive effect

for the case of advertisement and capturing the relation b/t TV and radio

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 X_1 + \varepsilon$$

$$= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon$$

$$= \beta_0 + \beta_1' X_1 + \beta_2 X_2 + \varepsilon$$

If interaction b/t two variables is statistically significant then I should keep both of the main effects as well.

This interaction variable is specially very useful in the case of quantitative and qualitative variables.

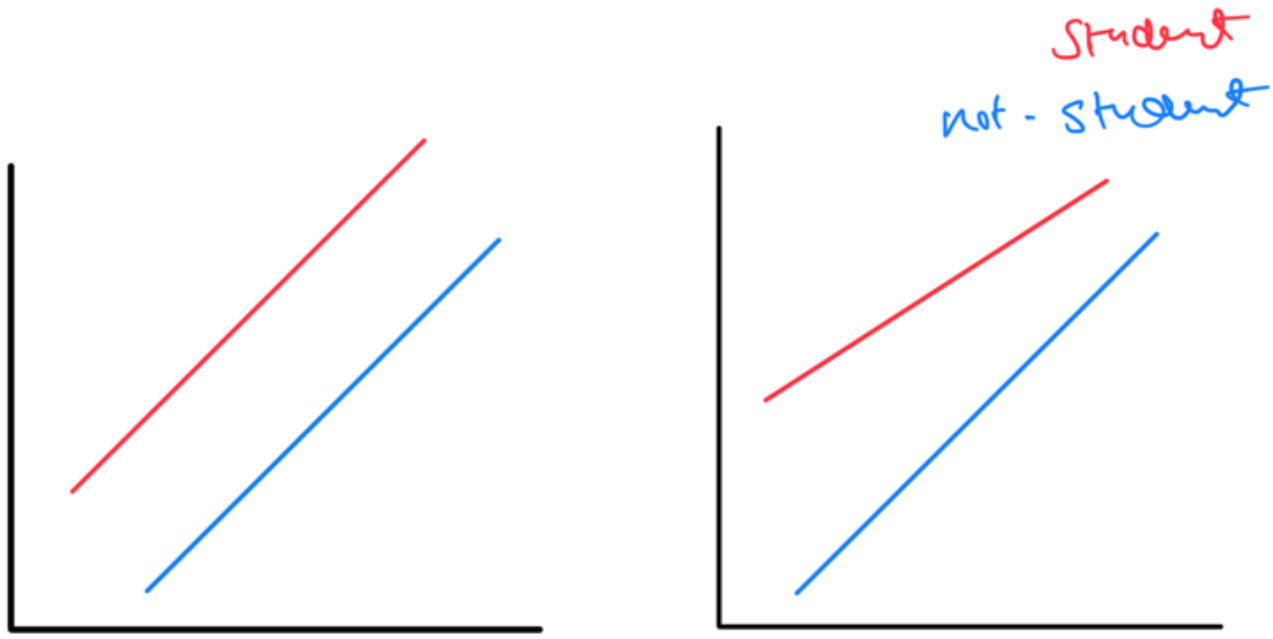
Eg → In the case of credit

with X_1 → Income (quant)

X_2 → Student (qualitative)

$$Y = \beta_0 + \beta_1 \times \text{income} + \begin{cases} \beta_2 + \beta_3 \times \text{income} & \text{if} \\ 0 & \end{cases} \begin{matrix} 1 \\ 0 \end{matrix}$$

$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income} \\ \beta_0 + \beta_1 \times \text{income}. \end{cases}$



W/o interaction

With interaction

Non Linear Relationships

extending linear model to accomodate non-linear relationships by using polynomial Regression

Eg → in Auto.CSV.

$$mpg = \beta_0 + \beta_1 * hp + \beta_2 * hp^2 + \varepsilon$$

to capture the quadratic nature of

mpg V/S bhp

To check for correlated error.

We plot residuals as a function of time and there should be no discernible pattern.

If error terms are positively correlated then we may see tracking in the residuals, i.e. adjacent residuals may have similar values

3) Non constant Variance of Error

terms that $\text{var}(\varepsilon_i) = \sigma^2$.

Identified by heteroscedasticity or the presence of a funnel shape in residual plot.

Soln → transform it to a concave f^n like $\log(Y)$ or $\sqrt{Y}$

4) Problem of outliers

Outlier can be seen in the residual plot easily.

⇒ Instead of plotting residuals, we can plot the studentized residuals computed by dividing each residual e_i by its estimated std. error.

v

$> 3 \rightarrow$ then outlier.

5) High Leverage observations

obs that have high residual & leverage → change the predictors a lot.

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \bar{x})^2}$$

this is a bigger problem to identify in multiple regression

6) Collinearity

If i have two predictors that are collinear then I test my null hypothesis then I

might get an incorrect answer

So I should always plot the scatter matrix

A better way to assess multi-collinearity is Variance Inflation Factor (VIF)

$$VIF \rightarrow \frac{\text{var of } \hat{\beta}_j \text{ w/ full model}}{\text{var of } \hat{\beta}_j \text{ fitted on its own}}$$

Smallest value of VIF is 1,

which means there is no collinearity

VIF > 5 or 10 is problematic

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{x_j / x_{-j}}}$$

KNN with small K will have high variance & less bias as it is flexible - whereas with high K it tends to regression which is not very flexible.

large K = smooth fit

In general a non parametric approach will predict better than ~~parametric~~ approach unless the parameters are close to true form.

When $p = 1$ or 2

KNN outperforms LR.

with $p = 3$ mixed results

but $p \geq 4$ {p: no. of predictors}

Linear regression outperforms KNN

as dimensions increase they pose a challenge for KNN

Curse of dimensionality

{there is no near neighbour}

eg → $n = 50$, $p = 20$