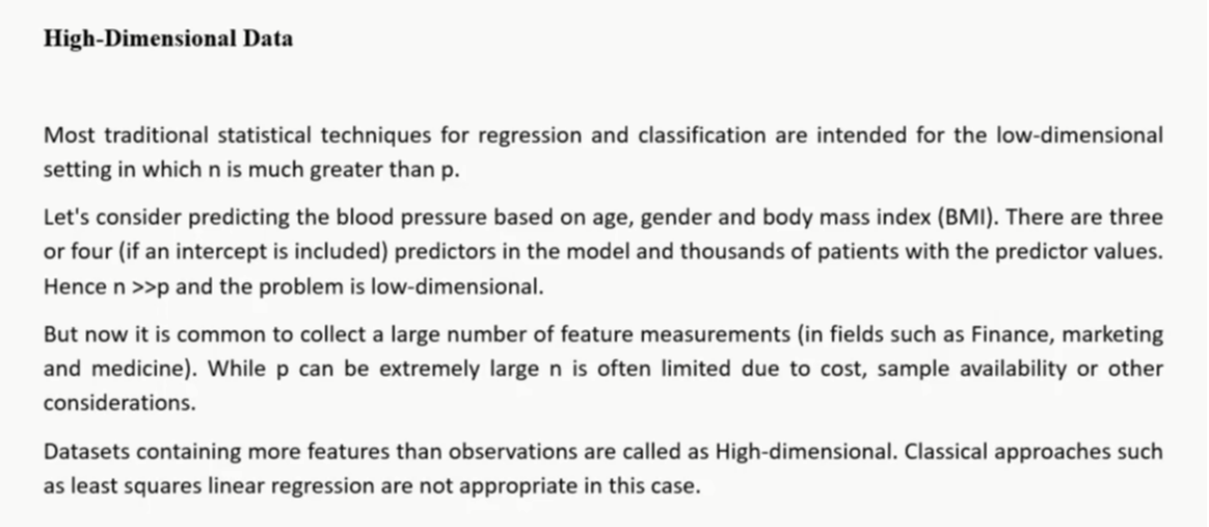
ML day 15

High Dimensional Data

- n - number of rows and p - number of parameters that is mean column or variables
- Most traditional n >> p , it is low dimensional and it is ok
- but if there is n<p ,n is too much greater than p there will be high dimensional data .
- Gap between the n and p must be less for high Dimensional Data .



High-Dimensional Data

Most traditional statistical techniques for regression and classification are intended for the low-dimensional setting in which n is much greater than p.

Let's consider predicting the blood pressure based on age, gender and body mass index (BMI). There are three or four (if an intercept is included) predictors in the model and thousands of patients with the predictor values. Hence n >>p and the problem is low-dimensional.

But now it is common to collect a large number of feature measurements (in fields such as Finance, marketing and medicine). While p can be extremely large n is often limited due to cost, sample availability or other considerations.

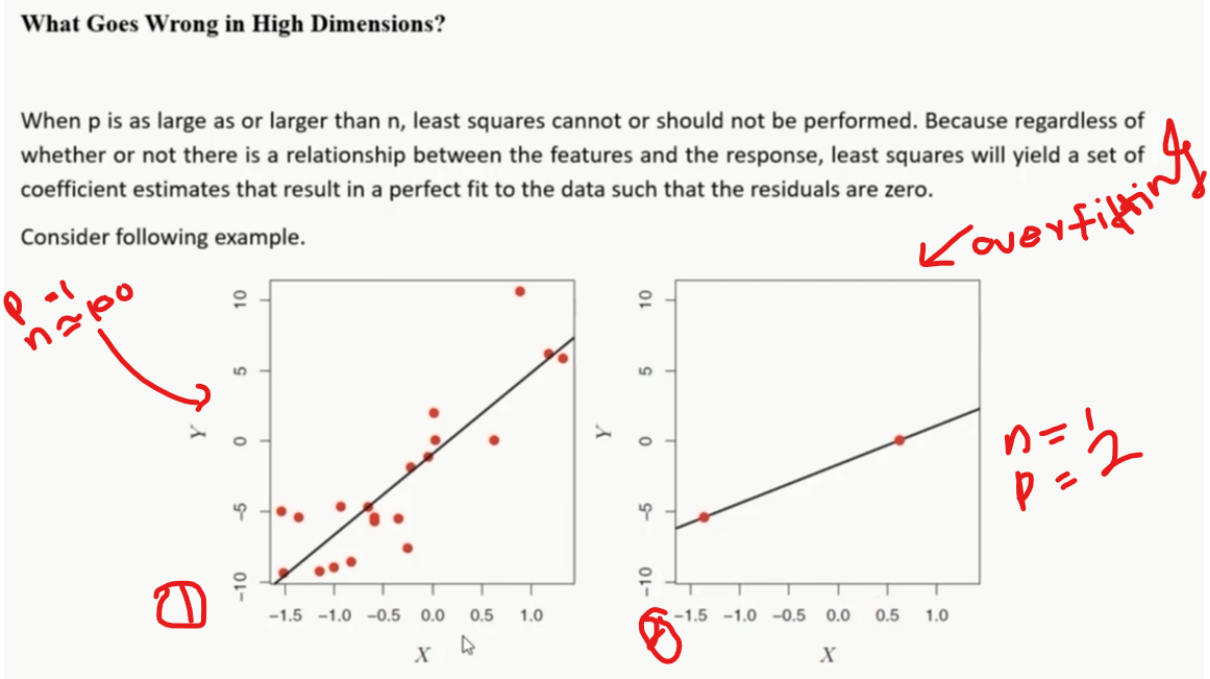
Datasets containing more features than observations are called as High-dimensional. Classical approaches such as least squares linear regression are not appropriate in this case.

What goes wrong in High Dimensions ?

- So consider the chart 1 , in here there is only one p and near there are 100 of rows , So think if we increase the number of parameters the best fit line

will get curved

- when we consider the image number 2 , there is n = 1 and p =2 , so n>p it is true but this will cause the overfitting becuase this data set will not suite for the unseen data



Curse of Dimensionality

- example :
  - If there is a child we will give cookies to test , first think if there is two flavors at least child must eat two biscuit to select best , if there is three shapes and two flavors child must eat 2 × 3 = 8 cookies
- When the number of features grows the number of data also getting high
- This one also not a good thing .
- What can be do in here ? So we have to reduce the dimensions .

Dimension Reduction

- There are two main ways to do dimensionality Reduction
  1. Feature Selection
  2. Dimensionality Reduction Algorithms

Feature Selection : Supervised Learning

- Best Subset Selection
- Forward Selection
- backward Selection

Best Subset Selection

- We have to create model from null model to Full Model
- After That we have to get the less RSS value (Residual Sum of Squares)
  -

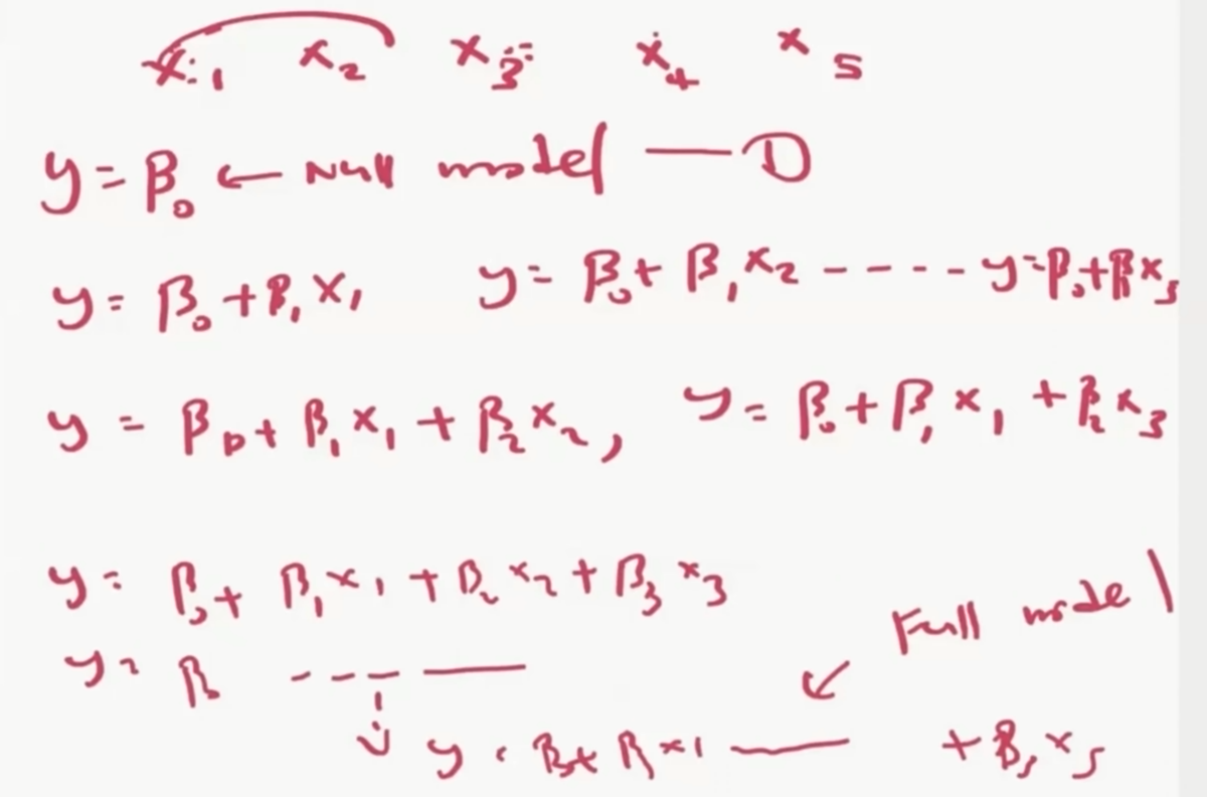
1. **Definition**:
   - **RSS (Residual Sum of Squares)**:
     - The sum of the squared differences (residuals) between the actual and predicted values.
     - Formula:

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

     Where $y^{(i)}$ is the actual value, $\hat{y}^{(i)}$ is the predicted value, and m is the number of data points.
   - **MSE (Mean Squared Error)**:
     - The average of the squared differences between the actual and predicted values. It is essentially the **RSS divided by the number of data points**.
     - Formula:

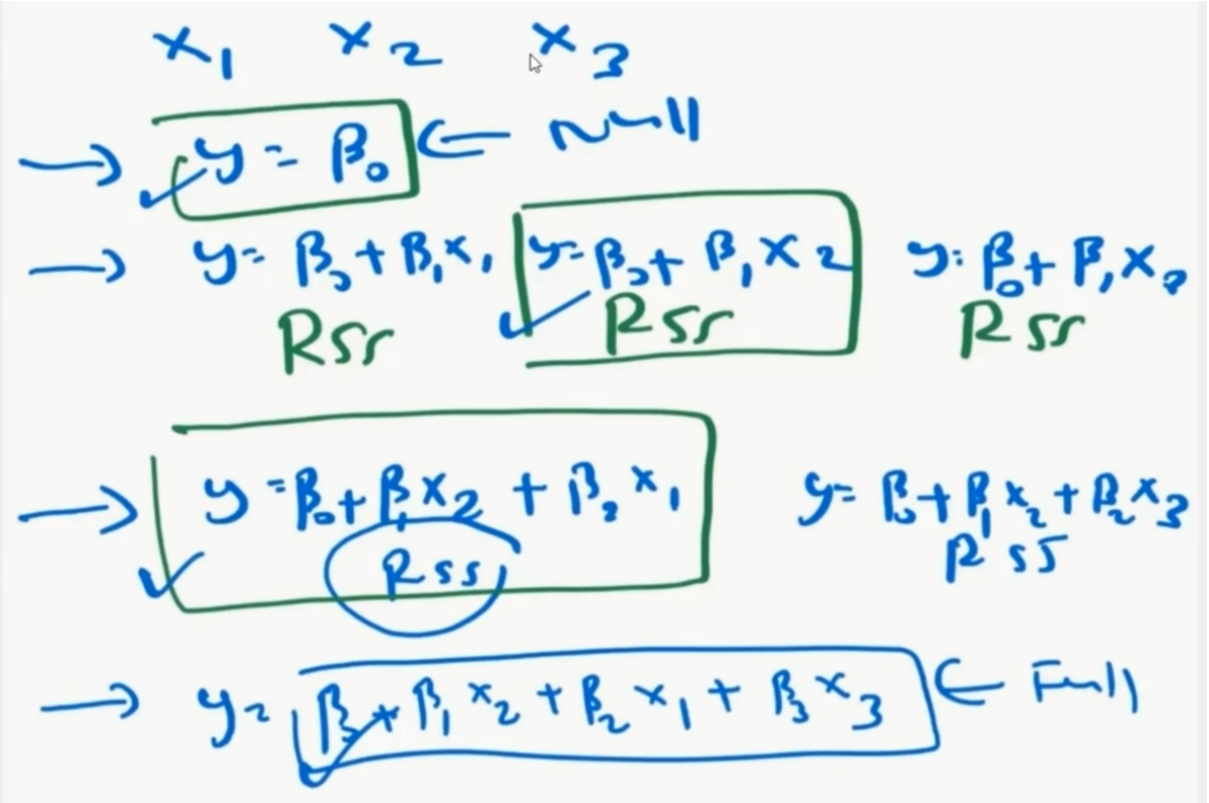
$$MSE = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2$$

- 2 to the power P will give number of Models
- If the P is increasing there will be more models so we have to find the error in more models .



Forward Selection

- start with null model
- Fit P simple linear regression models and get the lowset RSS model .
- Increase that model with rest of the variables (P-1) with the lowest RSS .
- Construct until some stopping rule is satisfied .
-



Backward Selection

Backward Selection

- Start with the model with all P variables.
- Remove one variable and select the most significant model (P-1).
- Continue until a stopping rule is reached