Advanced Regression Notes:

Summarising the important takeaways from the lecture:

1. In generalised regression models, the basic algorithm remains the same as linear regression. We compute the values of coefficients which result in the least possible error (best fit). The only difference is that we now use the features ϕ1(x),ϕ2(x),ϕ3(x)....ϕk(x) instead of the raw attributes.
2. The term 'linear' in linear regression refers to the linearity in the coefficients, i.e., the target variable y is linearly related to the model coefficients. It does not require that y should be linearly related to the raw attributes or features; feature functions could be non-linear as well.

‘Linear’ in Linear Regression:

 The model is called 'linear' because the targety**is linearly related to the** **coefficients**. To fully understand this, it is crucial to note that in regression, the**coefficients** a0,a1,a2,...,ak are your **variables**, i.e., you are trying to find the optimal coefficients that minimise some loss function. On the other hand, the **features** ϕ1(x),ϕ2(x),ϕ3(x)....ϕk(x) are actually **constants** because you are already given the dataset (i.e., the values of x, and hence ϕ(x), are fixed; so, what you are trying to tune are the coefficients).

Thus, saying that 'y is linearly related to the coefficients' implies that **only two operations**can be applied between the coefficients:

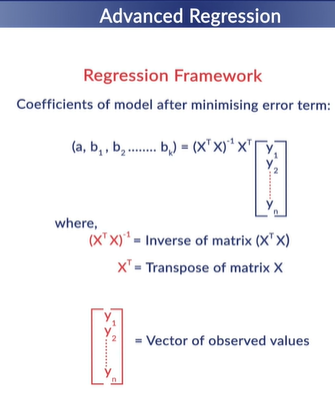
1) **Multiplying them by constants** (i.e., the features) such as a1ϕ1(x),a2ϕ2(x) and

2)**Adding the terms**with each other such asa1ϕ1(x)+a2ϕ2(x). What you **cannot** do is multiply them together, raise to one another's power, etc. That is, you cannot have terms such as a0.a1,aa32 etc.

Solving this system means to find the set of parameters a0,a1,...,a5 which satisfies all the equations. This can be done efficiently using **matrices**(and is done that way by many libraries)**.** The alternate way is to use optimisation methods such as gradient descent.

It turns out that the closest approximation of X−1 (for non-square matrices) is (XTX)−1XT. Thus, the (approximate) solution to this system is given by:





Summary:  
To summarise, you learnt that one needs to follow a 3-step process to build a regression model:

* First, we explored and visualized raw attributes to understand the shape of scatter plots.
* Second, we assessed which function of the explanatory variable would explain the shape of the data.
* Lastly, we wrote the generalized regression formula using the matrix format. We then summed up the errors between predicted and actual response variables and minimised the residual sum of error to arrive at the best-fit regression curve.

Another thing to note is that the term 'linear' in regression depicts the linear expression in the coefficients of the linear combination. It does not mean linear expression in raw attributes or features.

The ‘linear’ in linear regression does not stand for the relation between the target variable and the predictor variable. In fact, it stands for the coefficients of the predictor terms in the linear regression solution. It is entirely logical to have a linear regression solution as z= a\*sinx + b\*cosy, however, it will not be a linear regression model if the relation can be defined as y=2a\*y\*x^\*sinx + y^2\*x^2\*sinxy.

**Regularized regression**:

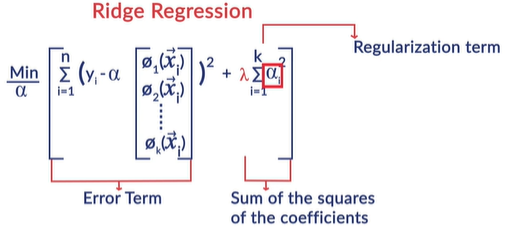
There are two kinds of regularization of a generic model, namely

* Ridge Regression
* Lasso Regression

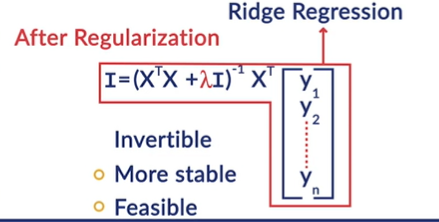
**Regularisation** is a process used to create an optimally complex model, i.e., a model that is as simple as possible while performing well on the training data. Through regularisation, one tries to strike the delicate balance between keeping the model simple and yet not making it too naive for any use.

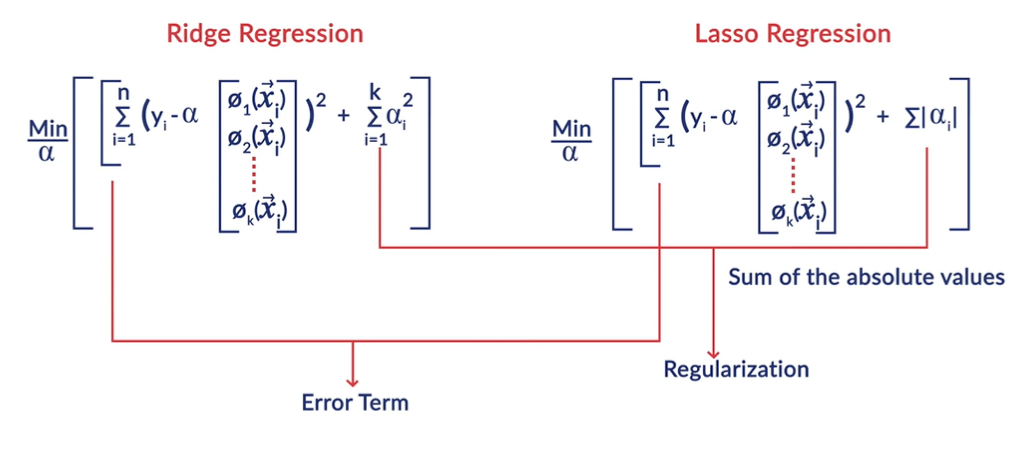
 Linear regression does not account for model complexity; it only tries to minimise the error (e.g., MSE). However, it may result in arbitrarily complex coefficients. On the other hand, in regularised regression, the objective function has two parts: the **error term** and the **regularisation term**.

1. Adding **sum of squares** of coefficients: *Ridge Regression*
2. Adding **Absolute value** of coefficients: *Lasso Regression*



Significance of Lambda in Ridge Regularized Regression:

* If the value of lambda used is high, any complexity will be clamped down
* If the value of lambda is closer to 0, then, unregularized regression is being performed (does not care about the complexity of the model)
* 



Lasso is more computationally intensive than Ridge Regression.

*Though lasso regression can be used for feature selection while ridge regression cannot be, it comes at a huge computational cost. Since it does not convert into a nice invertible function, it is to be solved using an iterative process which has significantly more computational requirements compared to ridge regression which demands a simple tweak to the simple linear regression solution and can be converted to an invertible matrix and can thus be solved using matrix operations and thus has significantly lower computational costs associated with it.*

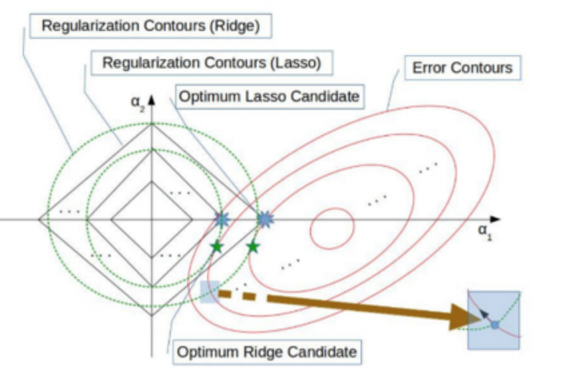
Both these methods are used to make a regression model simpler while balancing the 'bias-variance' trade-off.

In ridge regression, an additional term, 'sum of the squares of the coefficients', is added to the cost function along with the error term, whereas in case of lasso regression, a regularisation term, 'sum of the absolute value of the coefficients', is added.

Lasso Regression is helpful for Feature Selection.

To summarise, one of the most important benefits of **lasso regression**is that it results in model parameters, such that the lesser important features' coefficients, becoming zero. In other words, lasso regression indirectly performs **feature selection.**

Thus, the cost function for both ridge and lasso is given by:   

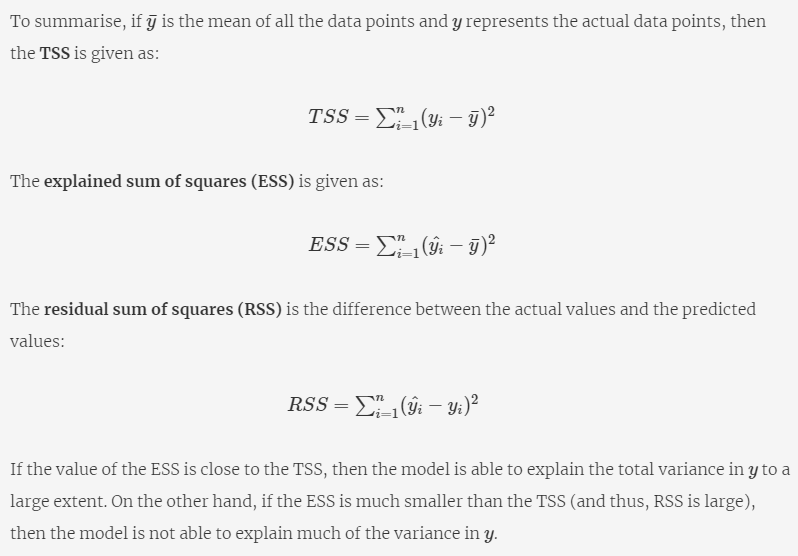



A **contour** of a function f(α) is a trace (locus) of the points that satisfy the equation f(α)=c for some constant c. The figure shows contours for various values of c: the 'inner' contours represent lower values of the error/regularisation terms. You want to find the coefficients that **minimise both the error term and the regularisation term.**

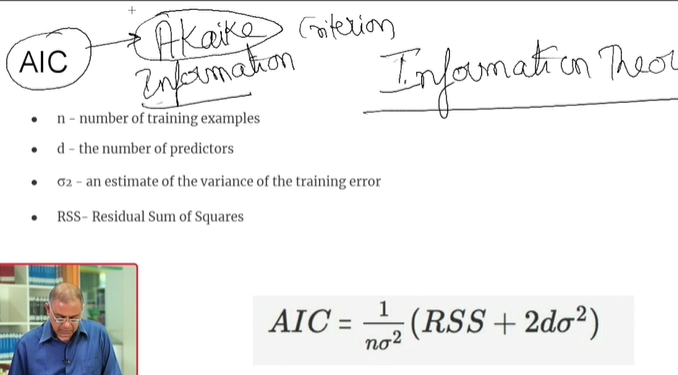
Thus, the key observation here is that at the optimum solution for α (the place where the sum of the error and regularisation terms is minimum), the corresponding regularisation contour and the error contour must 'touch' each other tangentially and not 'cross'.

*The outer contour represents the highest error value as described in the comprehension.*

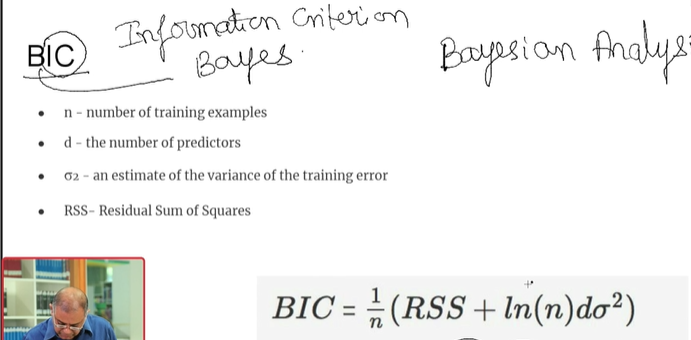
 TSS=ESS+RSS.



Akaike Information Criterion (AIC)



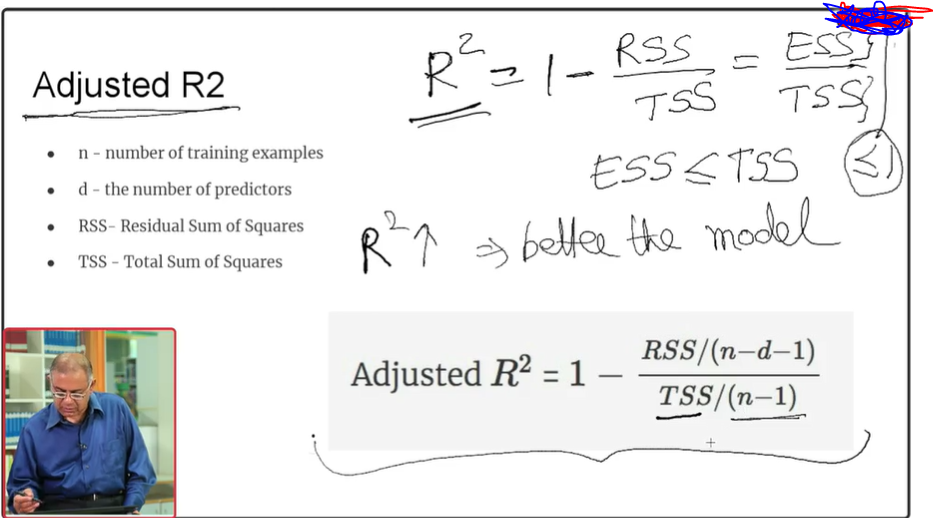
Bayesian Information Criterion (BIC)



To summarise, AIC is defined as follows:

AIC=−lnP(D|M)+d

Here, **d** is the number of parameters in the model, **D** is the entire data set, **M** is the model, and lnP(D|M) is the model likelihood. In the next lecture, you will learn about the **adjusted R-squared.**



To summarise, R2 is given as:

R2=1−RSSTSS=ESSTSS≤1

A model is said to be a good model when it is able to explain most of the TSS; the larger the R2, the better the model. But the problem with using R2 is that it increases with the number of features in the model and does **not have any regularisation**mechanism to keep a check on the complexity.

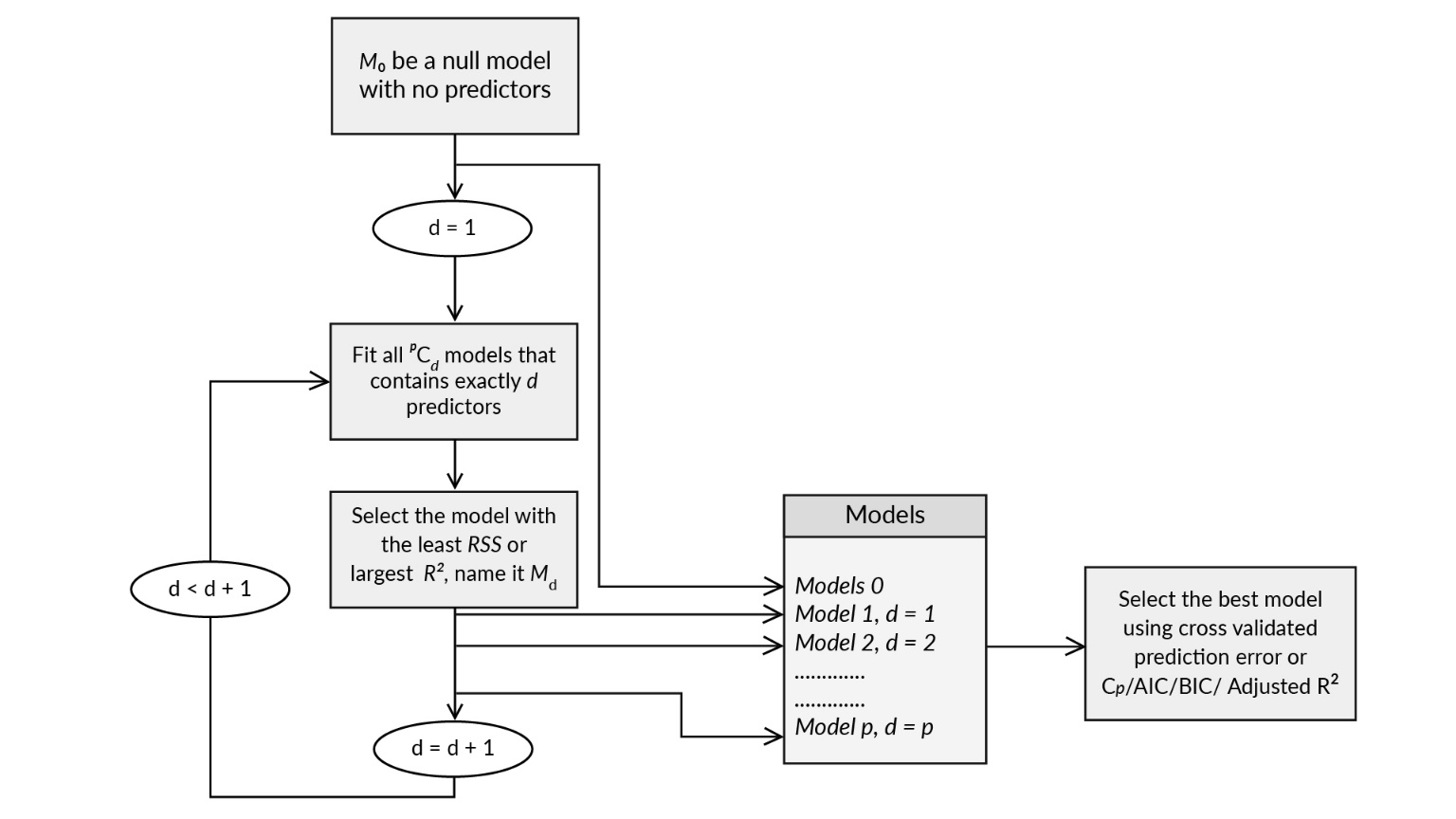
So, in practice, we use the**adjusted R-squared:**

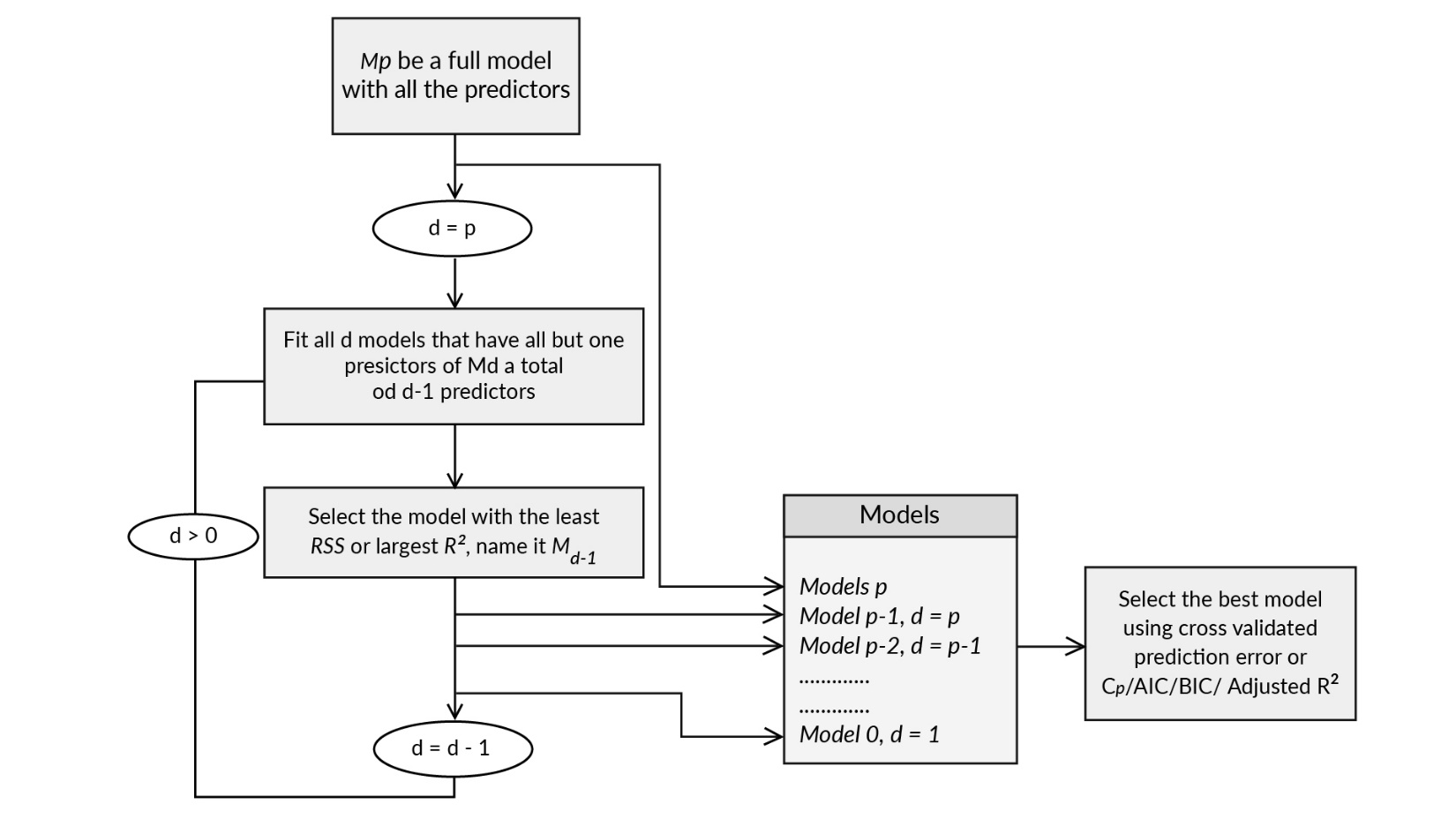
Adjusted R2=1−RSS(n−d−1)TSS(n−1)

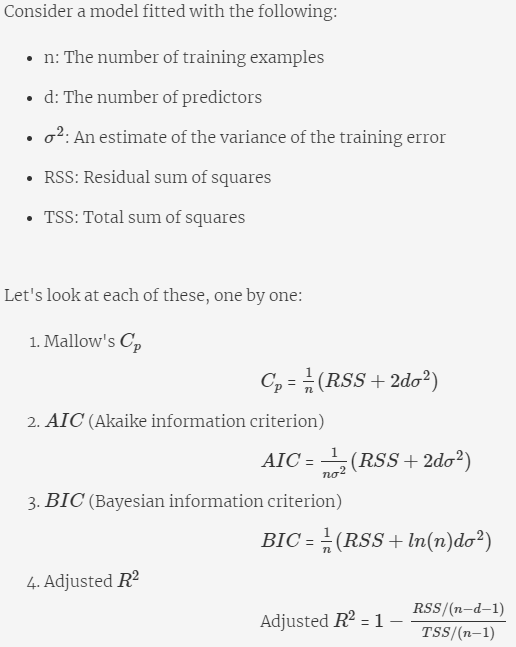
Adjusted R-squared penalises the model for having a large number of variables, d. When d increases, the term (n-d-1) decreases, which increases the term RSS(n−d−1), and consequently, the adjusted R-squared will decrease.

If the dataset is large, BIC can be use, else, use AIC.

**Forward Selection Process**:



Backward Feature Selection: 



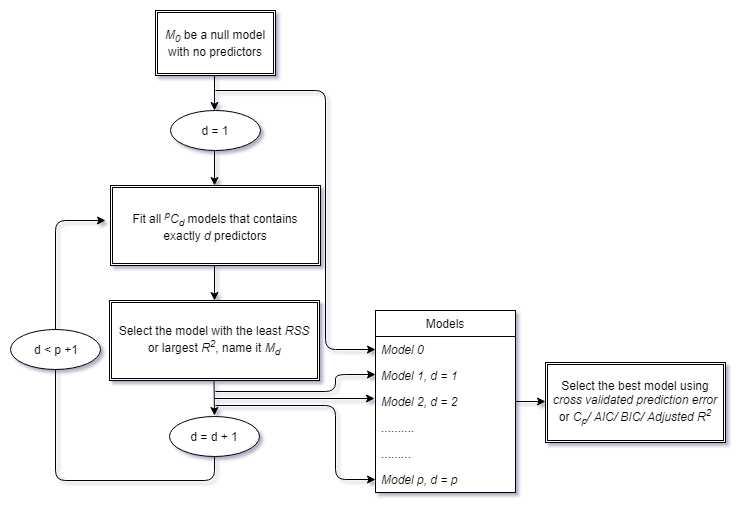
You can see that as we increase the number of predictors, d, the penalty term in Cp, AIC and BIC increases, while the adjusted R2 decreases. Hence, the lower the value of  Cp, AIC and BIC, the better is the fit of the model.

*If the number of observations,*n*> 7, then*Cp*will give a lower value; else,*BIC*.*

Feature subset selection can be done in the following ways:

1. Best Subset Selection
2. Stepwise Selection

Best Subset Selection:

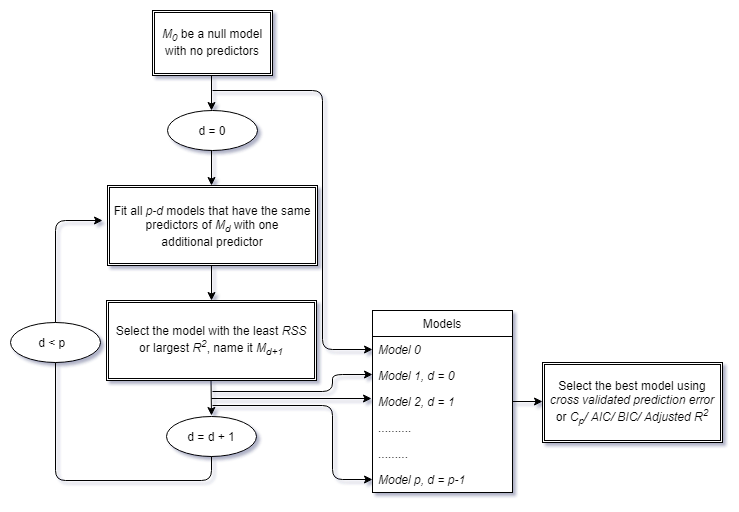


**Suggested Answer**

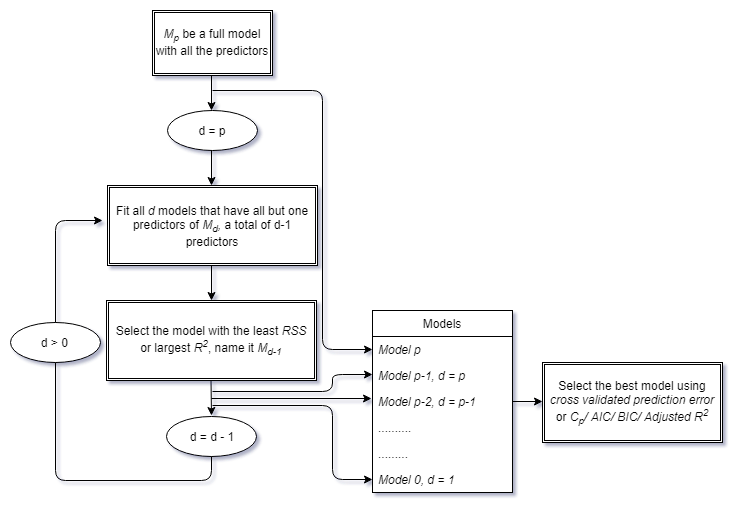
In the Best Subset Selection algorithm, we start with 0 features, i.e. a null model M0 with no features. Now, as we increase the number of features, we consider every model that has all combinations of a certain number of features. and select a model which results in the least RSS (or largest R2). This gives us a model Md with d features. We continue this iteration by increasing the value of d by one till you reach d is equal to the number of features in the dataset and find the models M0, M1, M2,....., Mp. Out of all these models M0, M1, M2,....., Mp, select the best one, as measured by measures such as Cp, AIC, BIC, Adjusted R2 or mean cross-validated error.

We can see that the total number of models that need to be analysed for Best Subset Selection is 2p where p is the total number of predictors.

**Stepwise Selection - Forward**



Stepwise Selection – backward



We can see that the total number of models that need to be analysed for forward stepwise selection is 1+p(p+1)2, where p is the total number of predictors. It is the same for backward stepwise selection.

In backward stepwise selection, how many models do we fit at each step as we decrease 'd'? (The total number of predictors is p.)   
p,p−1,p−2,......,2,

*When n < p and the number of predictors > 40, we can only use forward stepwise selection.*

Suppose we are regressing an independent variable y on 18 predictors on a data set with 400 observations. Which method will be able to give us the best model, i.e., one with the lowest test error?  
*Best subset selection can get the best model, as it tries each and every combination, and here, the number of predictors is also less than 40.*

Graded Questions:

Why is multicollinearity a problem in linear regression? Select the correct option.

**Singularity of the matrices that need to be inverted for the least squares solution causes instability/breakdown of the algorithm.**

**Feedback :**

*Multicollinearity means that some of the attributes (features) are linearly dependent. This means that in a feature matrix, one or more columns can be written as a linear combination of the other features. As a result, the determinant of the matrix will become 0 causing the matrix to be singular.*

Ridge regression is a technique for analysing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large; so, they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard test errors.

As λ increases from 0 to infinity, select the correct option that describes the pattern of the variance of the model.  
**Decreases steadily**

**Feedback :**

*When λ=0, the alphas have their least square estimate values. The actual estimates heavily depend on the training data, and hence, the variance is high. As we increase λ, the alphas start decreasing and the model becomes simpler. In the limiting case of λ approaching infinity, all betas reduce to zero, and the model predicts a constant and has no variance.*