Week8- variable selection

* variable selection
* some extensions of regression models.

# 8.1 Introduction(WHY?):

**Factor based models**

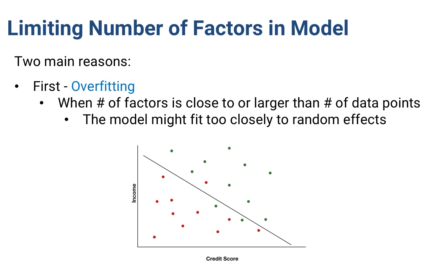
1. classification
2. clustering
3. regression

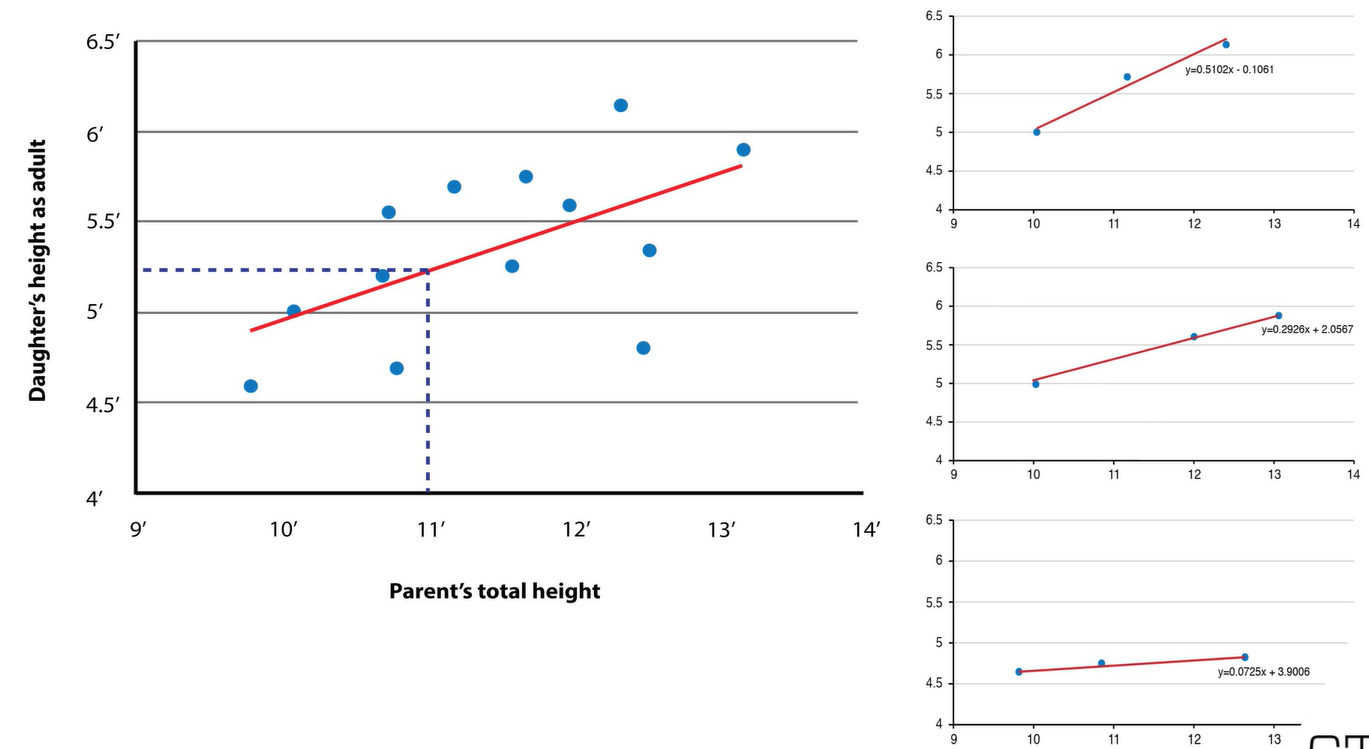
**Implicit assumption:** lot of factots in final model. But this is not ideal

**Variable selection:** Applicable to regression and other factor based model

## Why not so may factors?

1. When factors are close to data point ,it causes overfitting

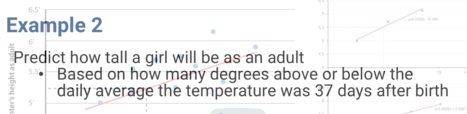




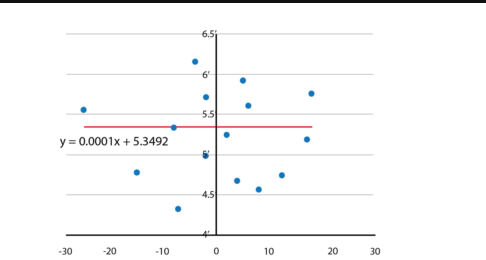
In the graphs above, when factors is almost close the data point(right side 3 graphs), it causes fit due to random effect.

More data points and less factors give a better fit to real and less random impact (left side graph)

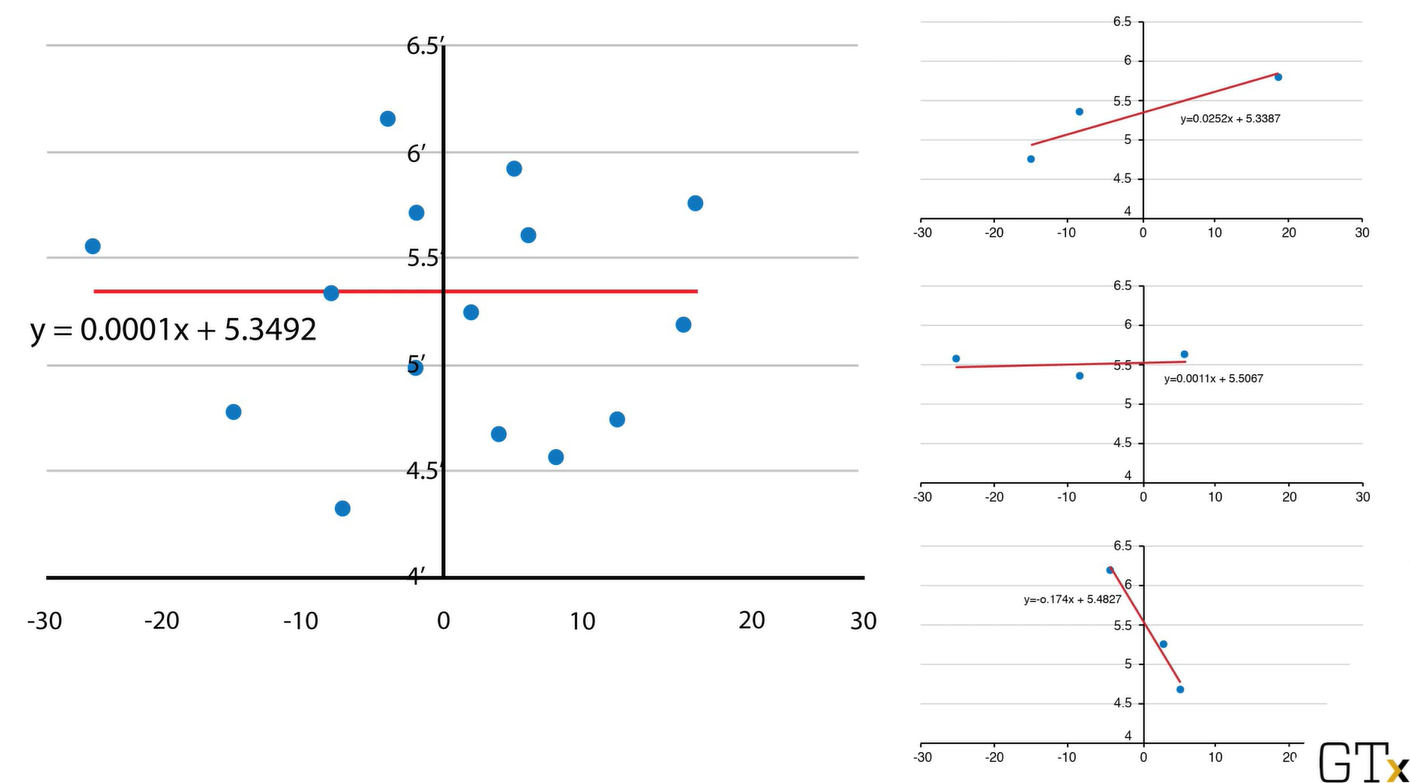
**Example:2**

****

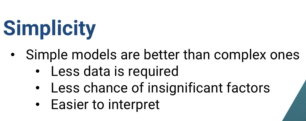
No correlation for example@ : so graph will be like this with no dependency between adult height vs. temperature



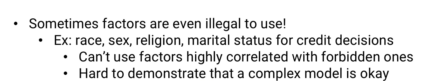
But if number of factors is same order of magnitude as data point, then we see right side graph with fit randomly and shows correlation



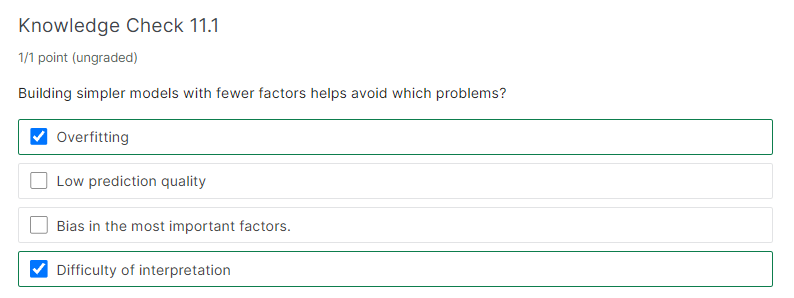
1. **Simplicity : simple model is more efficient**
2. Even if the p-value of a factor <0.05, thereis 5% chance it is meaningless.
3. For 20 of them,assuming independence, there is 64% chance that we included a factor that is meaningless
4. Easy to interpret



* Model with 100 factors is harder to explain than 10 factors



* We can remove the PII information and proceed. But, there are other factors that are correlated with the PII factors which can be used to infer PII information.
* If more factors are used, it is hard to explain if model is using any forbidden factor



# 8.2 Models for variable selection(HOW)

-regression is used..but any model can be applied

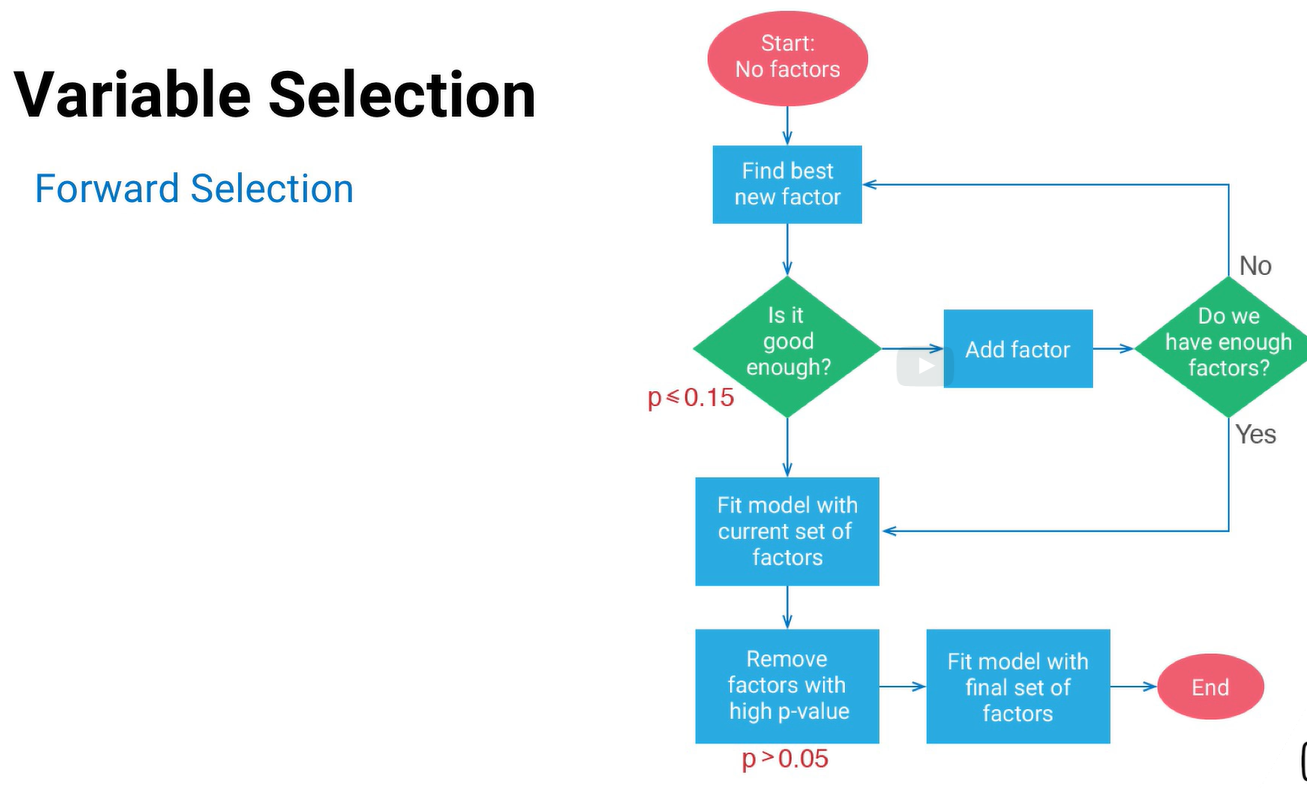
## Type#1:

-Classical

-building up or narrowing down

## 8.2.1 Forward Selection

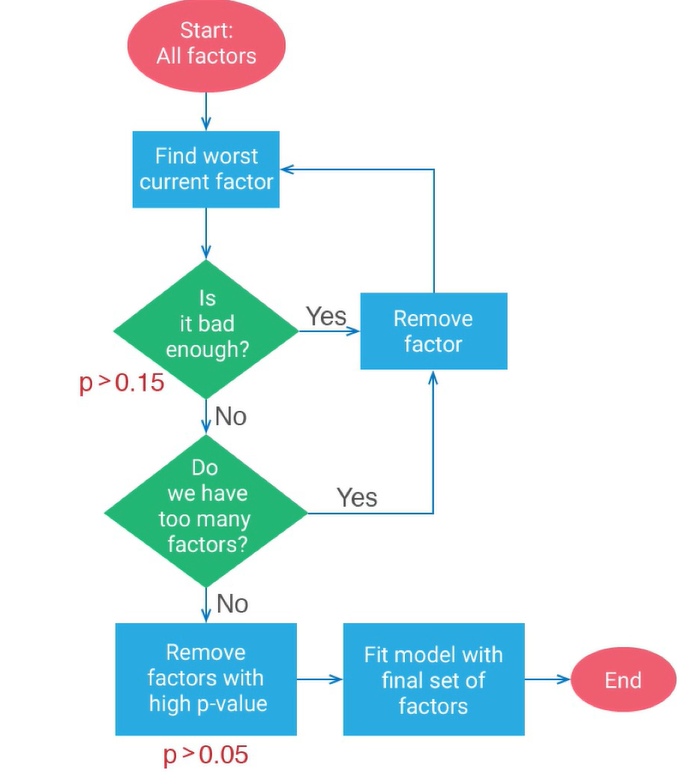
1. Start with no factors
2. Each step, add new factor as long as good enough improvement to model
3. When no factor good to add or we already have enough factors,stop
4. In the end, do a final check to remove any not good enough factors(i.er in the beginning add factors with greater p value and remove them with strict rule later)
5. Definition of good and good enough are parameters



## 8.2.2 Backward Elimination

-opposite of forward selection

-here we start with all factors and then eleiminate..bad enough is parameter we can set

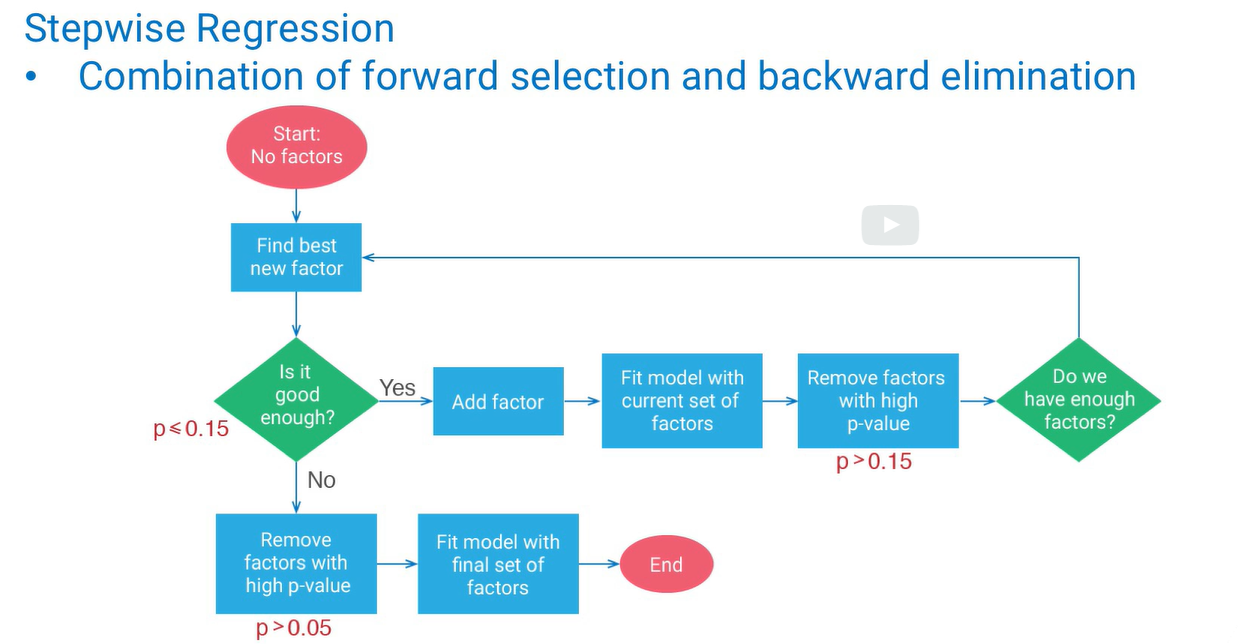


## 8.2.3 Stepwise regression

Combination for forward selection and backward elimination

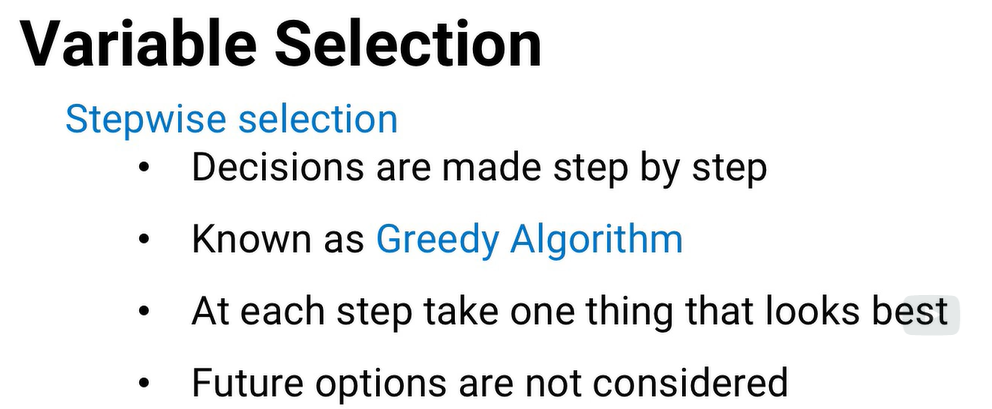
More than one variant of stepwise regression

**Steps:**

1. In flowchart below, it shows we are using forward approach
2. Difference between stepwise and forward: in stepwise, after adding new factor, we eliminate right away any factors that are not good..in “pure” forward, we add all and elimination is carried out in the end
3. 

* In all above, note we are using p value of factors ,but we can use other factors likeR2 value, AIC, BIC to pick the factor in all the above 3 variable selection procedure

## Summary(Classical algorithm)

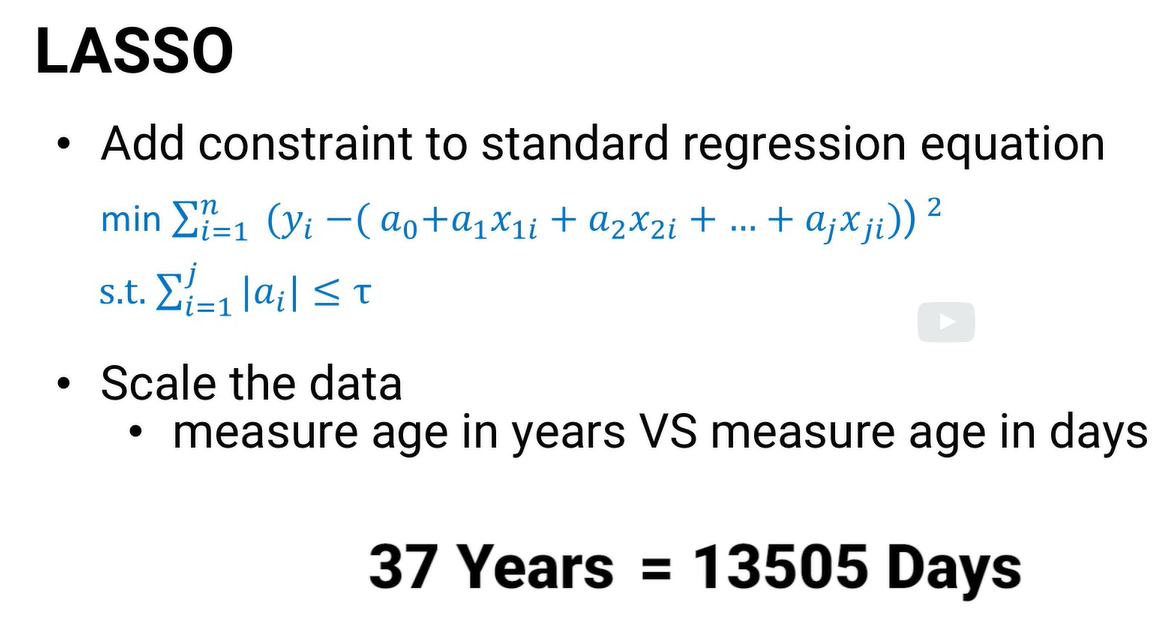
****

Global Options

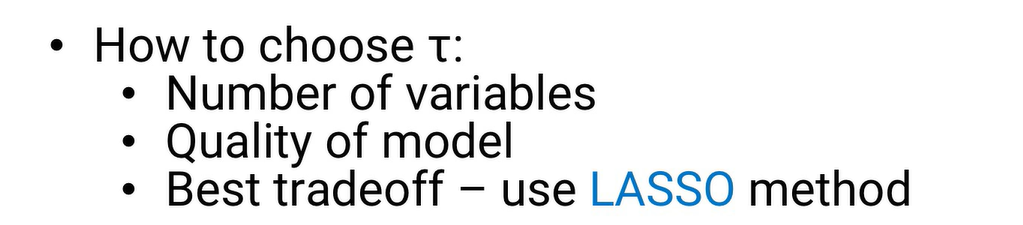
**-based on optimization models that look at all the options at the same time**

## 8.2.4 lasso Regression

* add constraints to standard regression equation (i.e sum of the co-efficients cannot be too large)
* Budget “t” is used on most importatnt co-efficients. Rest of the factors will have zero co0efficients and not part of the model
* just like SVM model, we need to SCALE

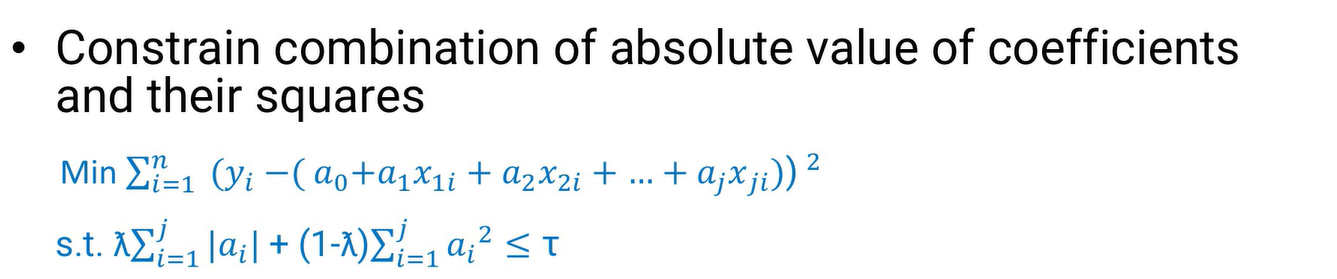


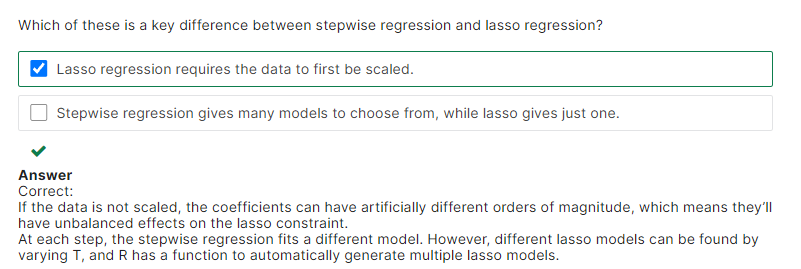
**How to choose the value of T:** use lasso for different values of T and see which gives the best value

****

## 8.2.5 Elastic Net

* SCALE is important
* if we remove the absolute value term from Elastic net, we will get “ridge regression”
* Ridge regression doesn’t do variable selection.but lead to better predictive models

****

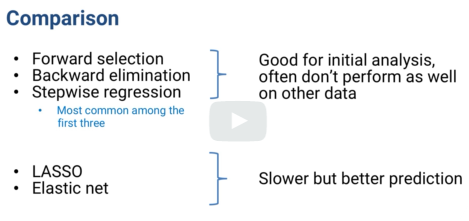


# 8.3 Choosing a variable selection model

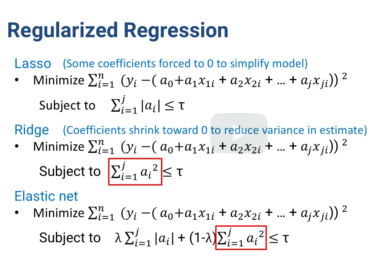
## Difference between classic vs. refined

|  |  |
| --- | --- |
| **forward, backward, stepwise** | **Lasso,Elastic net** |
| quick | slow to computer |
| Adv: good for initial data analysis | better to predict. |
| stepwise is most common | preferred unless introductory data analysis |
| disadv : most often fit random effects more For test data, they are not good |  |

**Suggestion**: use greedy methods first and then more refined global modelas later



## Math of all refined models



* Lasso has absolute coefficients
* Ridge has quadratic term
* Elastic net adds both absolute coefficients and quadratic term

**MATH**

1. Quadratic term in Ridge tends to shrink the coefficient values (whatever the initial value, quadratic term pushes towards zero) –
2. because Ridge is quadratic and not liner like lasso, its doesn’t make lot of co-efficients to zero like lasso does

(Optimization Thoery: difference between exterior point and interior point of a Polytope)

1. Shrinking the co-efficients adds bias, but reduce variance in estimates of prediction error

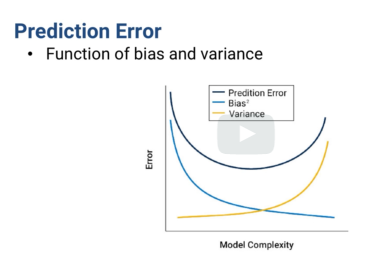
## Prediction error

Function of bias and variance

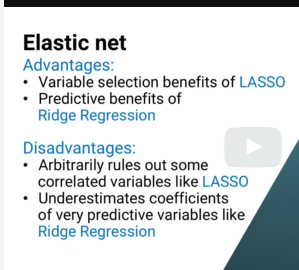
**Bias**: under-fitting

**Variance**: overfitting

Trading off bias for variance can give a better model (meaning underfitting is better than overfitting)



## Elastic net adv. And Disadv:



## Scenario#1:

Suppose the data set has two factors that are highly correlated.

## USIGN LASSO:

LASSO will usually pick just one of them to have a non zero coefficient. The other one is left out of the model.

(It's not always clear that the one of the two factors LASSO uses is the best one.)

Ridge regression will include both and because both coefficients shrink,they get closer to each other.

**Example/DISADV:** among relevant patients in your data set, two medical tests might give highly correlated results and either one would lead to a good predictive model.

But one test could be very inexpensive while the other is very costly.

Obviously, you'd prefer for your model to use the inexpensive test but LASSO might include the expensive one instead.

**Workaround:** check the correlations manually remove the expensive test and rerun the model.

## USIGN RIDGE:

Ridge regression includes more factors and because larger coefficients contribute much more to the constraint term, it tends to shrink the coefficients toward each other.

That means if you have some really good predictors, their coefficients might be underestimated by this method.

**Suggestion:**

Not single good rule of thumb when to try one vs. another. Try more than one and compare the differences between the models they suggest.

