# Agenda

-Advanced data preparation techniques, including principal component analysis.

# 6.1 Box –cox Transformation

**Pre –requisite:** Do q-q plot to check the data distribution and then decide if box cox is required

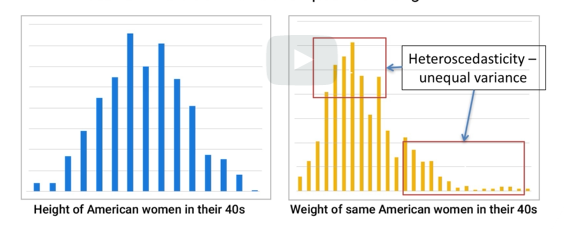
**Box –cox:** - used to transform data before using the data to fit a model

**Why?**

* Some models assume data is normally distributed
* Results are bias when the assumptions are wrong

**Example:** build model to estimate somone’ height

**Data:** women’ height and weight in 40s



**heteroscedasticity :** From second graph, smaller weights have less variance and variance is more for larger weights(wider area).

Difference in variance is called heteroscedasticity (unequal variance)

**Issue:** if we use regression model on this data, this could cause bias as higher variance for larger weights can make estimation error larger and push the model to fit that

**Resolution: Box cox**

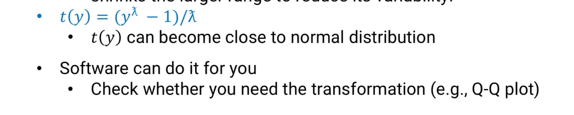
**Details of Box Cox:**

1. Logarithmic transformation
2. Stretches out the smaller range to enlarge its variability
3. Shrink the larger range to reduce variability

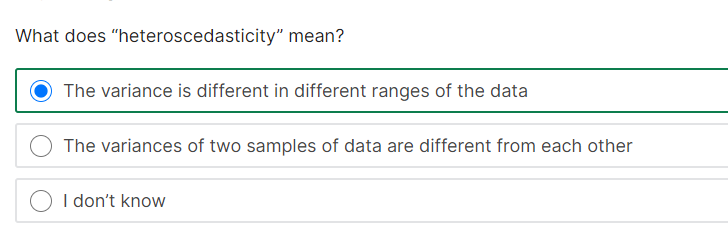
**Objective is to find best value of lamba**

**Y= vector of responses**

**T(y)=transform vector**

****

**Why box-cox:** name of the scientist(david cox and George box)

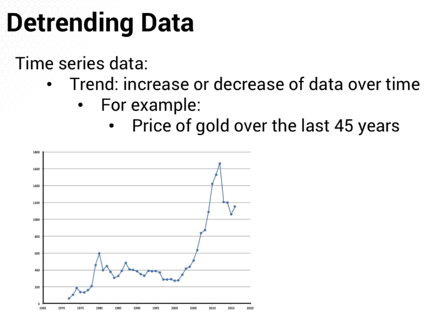
****

# 6.2 Detrending

## Why is it required?

- Time series data

**Sample**: below graph shows increasing trend, but inflation rate on dollar value decreaseis not considered in this plot



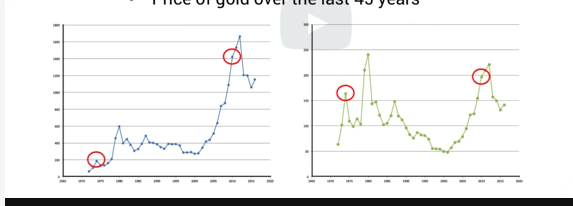
If we adjust inflation, we get below graph without upward trend. This is what detrending does



\*Detrend is used mess up in factor based analysis because of the trned

**Example:** we want to use regression to find factors impacting price of gold

1. We could have same factor value in two years separated by 2 decades , but results are different (in blue graph)
2. But in green graph after inflation adjustment, we see the results are same in the adjusted graph(green)

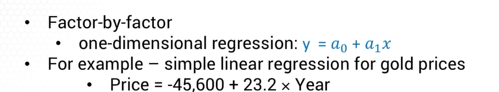


## Where can it be applied?

1. Detrending can be done to predictors and responses
2. Consider detrneding with factor based model, regression , support vector machine model

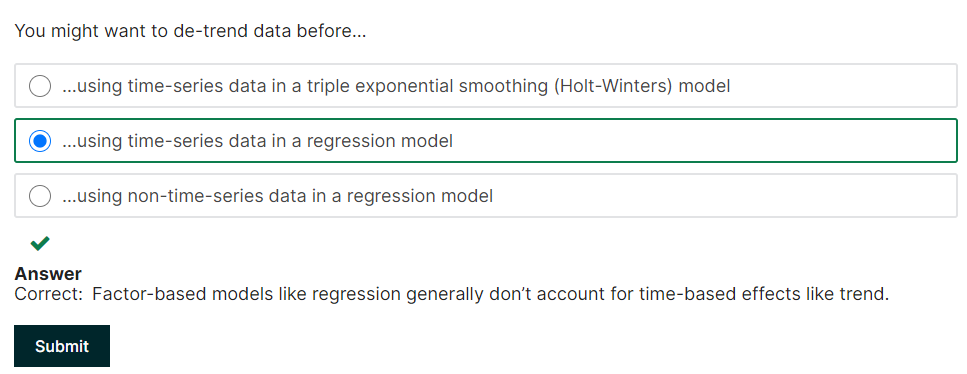
## How to detrend?

* For gold trending,we used individual inflation rate for each year(using historival inflation)
* But, most cases we don’t know whats the value in the past except for the value we want to detrend
* Simple way : go factor by factor and fit 1-D regression to it
* Using regression we can get co-efficients of each predictor



Now subtract the price from the actual data that we need to detrend

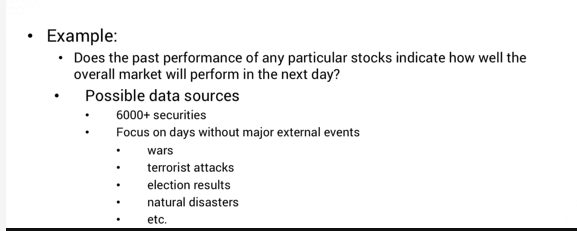




# 6.3 Principal Component Analysis(PCA)

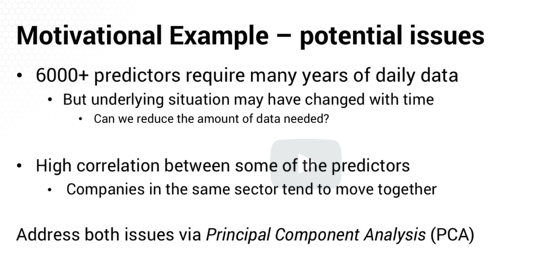
**Why is this used?**

1. feature extraction for highly correlated data
2. we have many factors and want to decide which one is most important factor to consider

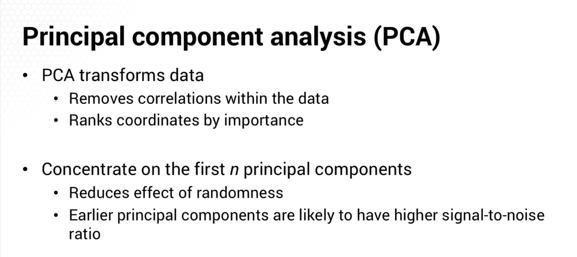


**Challenges for the example above:**

1. over 6000 predictors(not including possible interaction effects with other variables)
2. we need lots of data to avoid system being overdetermined
3. But, with so many years of data, the underlying situation could have changed. Tesla is popular now. But have no significance 10 years before ; Other hand kmart might be good predictor based on history, but after sears acquired kmart, it is closing stores
4. Some predictors are highly co-related



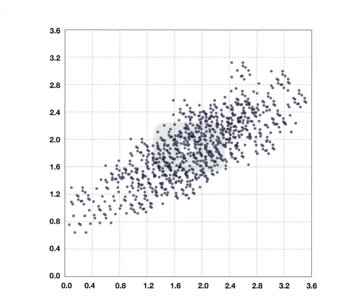
## What does PCA do : used for feature extraction



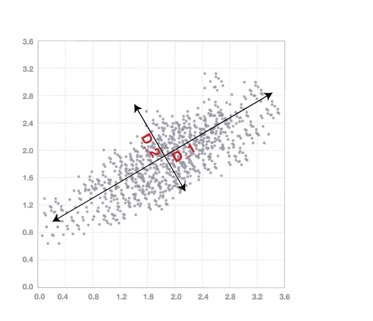
Higher signal to noise ratio means less random effect

The further down the list of principal components we go,the lower the signal to noise ration is likely to be. (more random effect

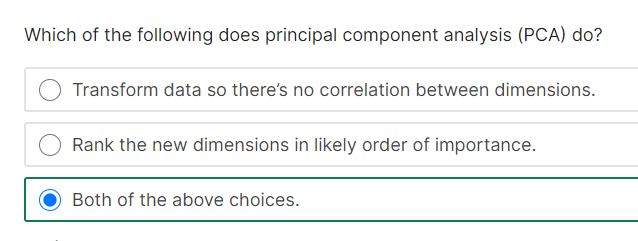
## How to do?

graphs shows two factors with strong co-orelation

1. If we rotate or add new dimensionality, it makes the data orthogonal and remove co-relation
2. We also can identify between two predictors, which has more wider spread



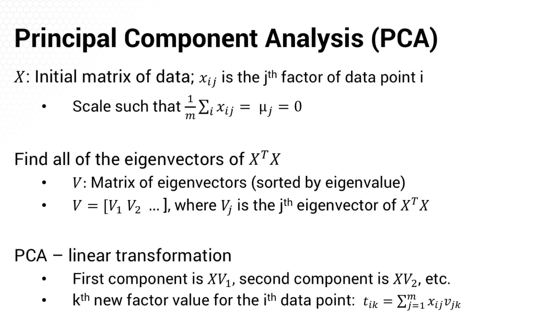
D1 has wider spread than D2. So PCA makes D1 as better bet than D2



# 6.4 Math of PCA

**Pre-requisite**: Eigen value /eigen vectors

Example: x(ij) : jth factor of data point “I” (after scaling), so average of all data points is shifted to zero



First PCA is “X” times V1

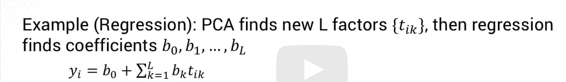
In other words,”V” is the linear transformation of data from “X” to Principal components

Each new factor will be linear combination of original factors

To remove correlation : we can use whole set of principal components

To identify “n” principal components : use first “n” principal components

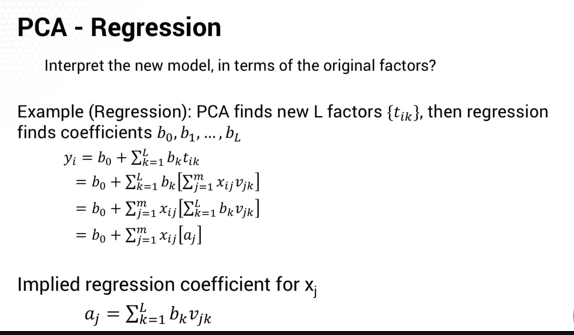
* Above math is linear. Use “kernels” for linear math  
  to use PCA regression on set of “L” factors, can we go back and use original factors to interpret the model?

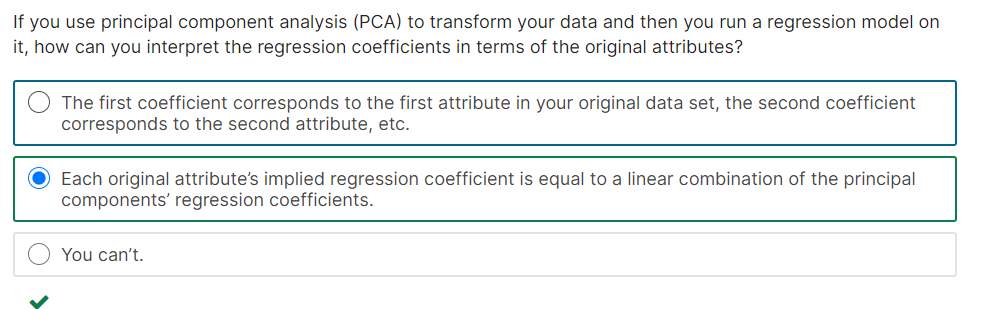


But this t-vector are themselves linear combinations of original factors. We can replace transformed features with “X” times “Eigen vectors”

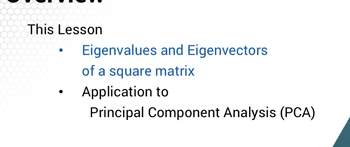


we can find the implied coefficient a j for each of our original factors j. Aj equals the sum from k equals one to l, of b k times v j k.So, we can easily find a **coefficient for each of the original factors,** and that way you give an intuitive explanation for the model



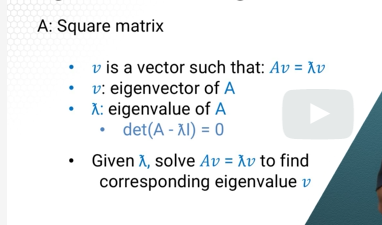


# 6.5 Eigen values and Eigen vectors



If we start with some vector “v” and use linear transformation “A” on it, we end up with a vector that goes in same direction as “v”

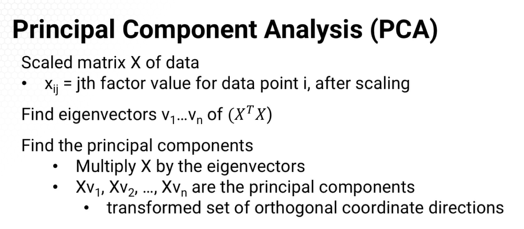
It may be longer or shorter based on “lamba”



\*Remember determinant of (A-lmba times identity matrix)=0

\* every one of the value is eigen value of “A”

Once we have the lamda value, we can plug into the equation and find the eigen vector



\*using eigen value and vector, we can transformed set that are orthogonal to each other