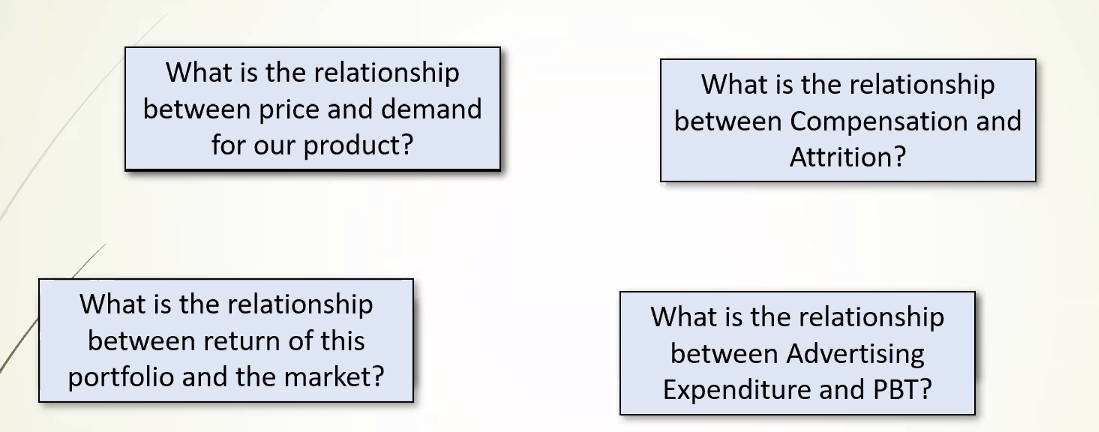
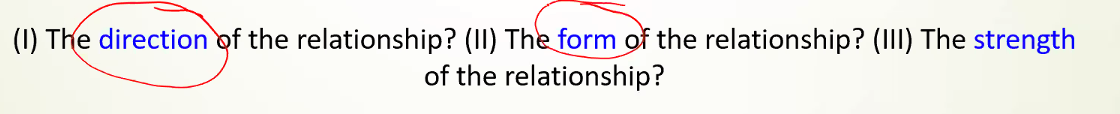
# Two variables

At some time these managerial questions are asked:



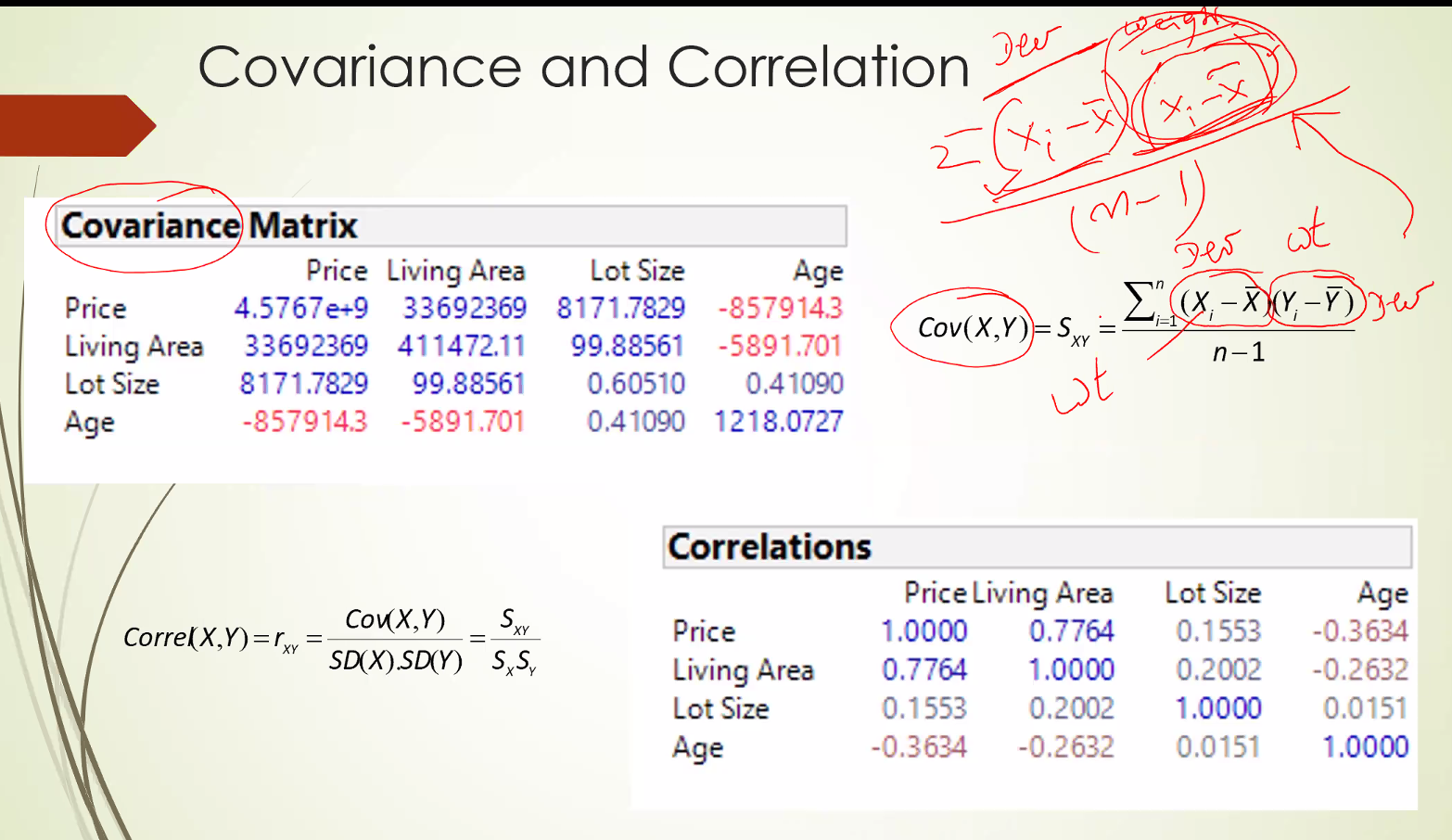
Some variables are not in control. But I know it depends on that (and that is controllable). Suppose PBT is uncontrollable but depends on working capital, employees etc.



1. Direction: If one increases other increases/decreases.
2. Form: Either can be cause effect relationship. Another can be irrespective of cause effect relationship. Regression analysis is specifically talking about cause effect. Correlation doesn’t mean cause effect. Thus regression is subset of correlation.

Ȳ

# Covariance and Correlation

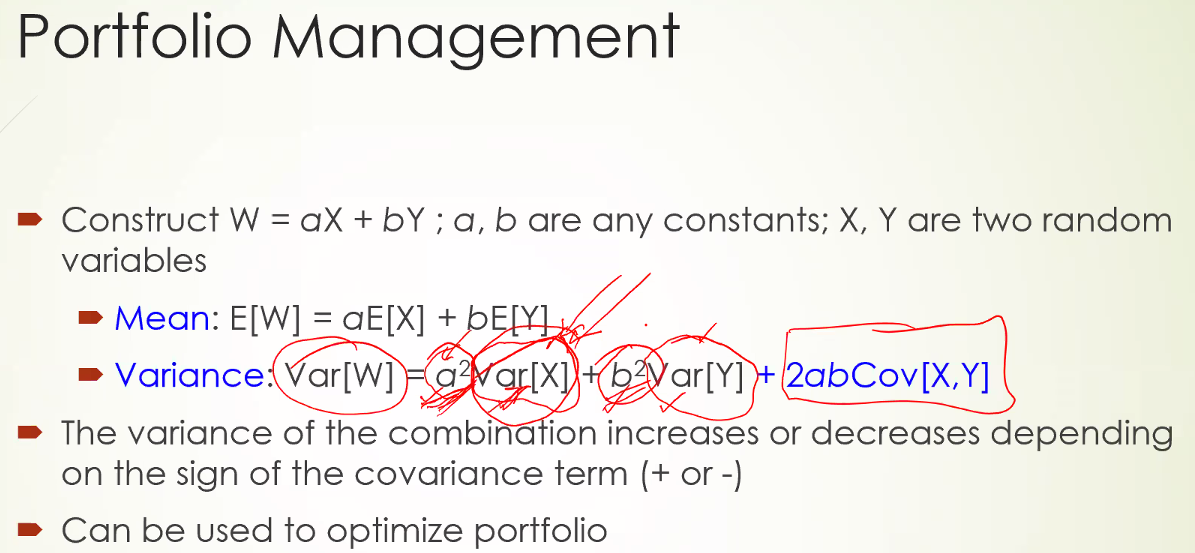


Why n-1? Why lose 1 degrees of freedom? Its because one is a variable and another is not a variable it is a weight. So lose 1 degrees of freedom.

Covariance is a general concept while variance is a special case.

Problems:

1. Units: If its variance everything gets squared. Living area sq. ft2. Also covariance be like $xsq. feet.
2. Scaling: Covariance is sensitive to scaling. And scaling it be like change in units of measurements.



Suppose X: Return of Infosys

Y: Return of Reliance.

a: proportion of my investment = 0.7 invested in Infy

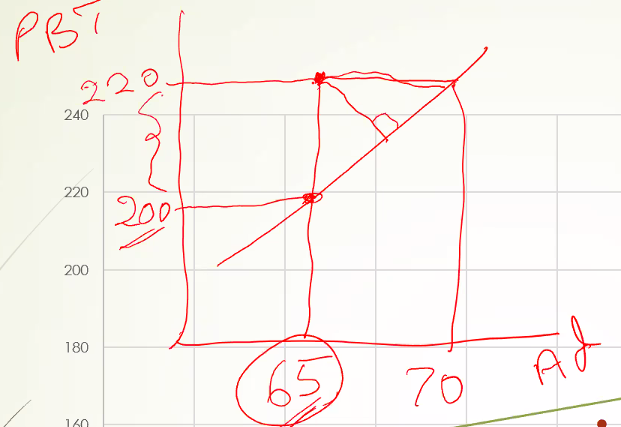
b: proportion of my investment invested in Reliance = 0.3

If I had invested all in Infosys, a = 1 b = 0. Means all the variance of Infosys (i.e. the risk) I am putting in Infosys.

But with a = 0.7, b = 0.3 means risk associated with Infosys will be 0.49 (comes down), and also for Reliance it become 0.09.

Now come to Cov(X,Y). My aim should be to see industries with negative covariance. This will result in decreasing the overall risk factor Var[W].

# Regression



How to interpret the 20M difference? Advertising in my control.

Based on the analysis I expected that PBT would be 200M but it went up by 220M. The unexplained deviation is attributable.

If I consider the horizontal distance I would say

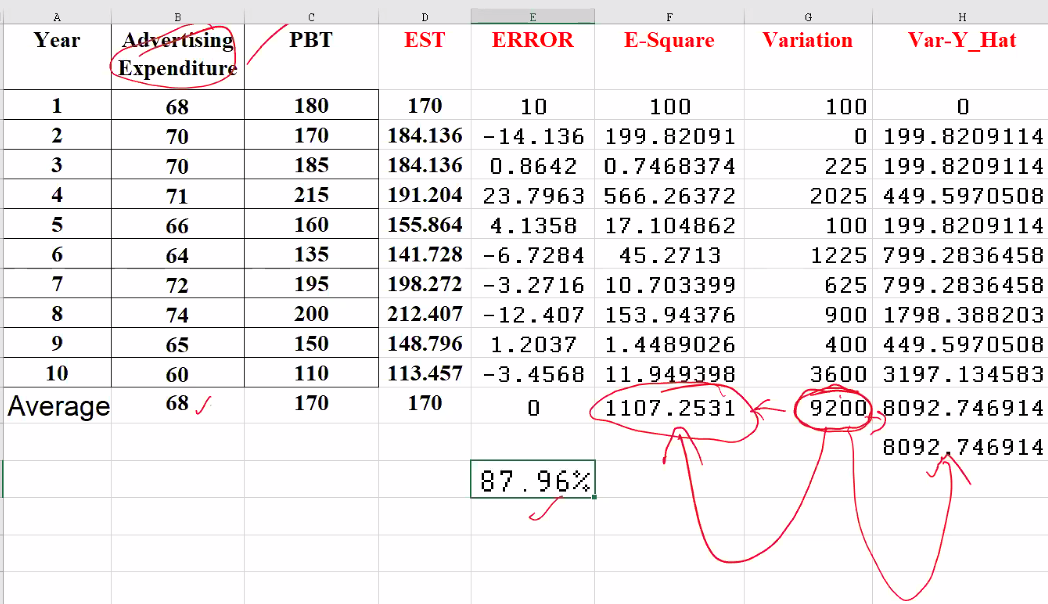
## Objective



The deviations in points should be minimized. But we need to keep a penalty just like variance that higher the deviation higher the penalty. In variance we measured the deviation from the mean. Here we measure the deviation from the line.



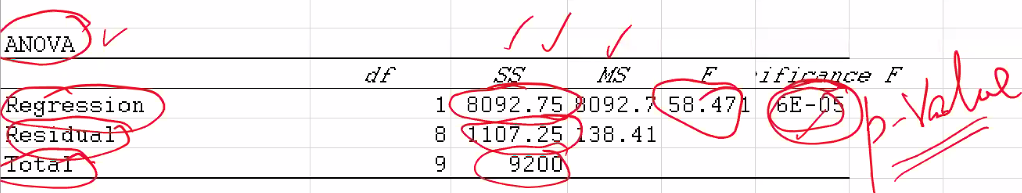
This means that regression line always passes through X\_bar and Y\_bar (means of both the variables)



9200 is the total variation in PBT out of this 8092.747 is attributable to variation in advertising expenditure, and rest 1107.2531 are due to factors other than advertising.

Thus 87.96% is the percent of variation explained by the regression equation. This is the R2. If you square root it will be exactly equal to CORR(X,Y) = r. Thus the name R2.

Advertising is acting as a proxy of other variables to explain the variability in PBT.



With ANOVA p value mean regression equation is statistically significant as it’s close to 0.