Session 16: Correlation and Regression

Agenda

Sl. No.	Topics For The Agenda
1.	Correlation
2.	How Is Relationship Measured
3.	Scatterplot To Analyze Correlation
4.	Strength Of Linear Association
5.	Other Strengths
6.	Examples
7.	Limitations Of Correlation
8.	Causation
9.	Least-squares Or Regression Line
10.	Linear Regression Model
11.	Estimated Regression Line

Sl. No.	Topics For The Agenda
12.	Correlation Coefficient, R
13.	Coefficient Of Determination
14.	Difference Between Correlation And Regression
15.	Limitations Of The Correlation Coefficient
16.	Multiple Linear Regression
17.	Assumptions Of A Linear Model
18.	Regression Diagnostics
19.	Detection of Collinearity: Simple Signs
20.	Detecting Multicollinearity
21.	Exercise

Correlation

Correlation

- A measure of association between two numerical variables.

Example (positive correlation)

- In the summer as the temperature increases, the thirst increases and more consumption of soft drink is there

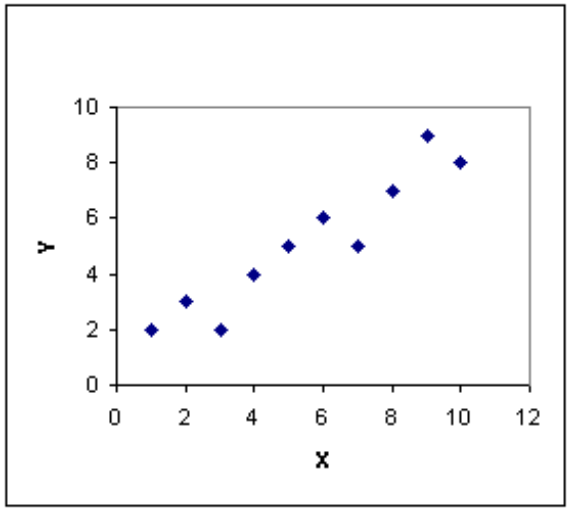
Example (negative correlation)

- As the rate of an product increases, the number of items sold decreases

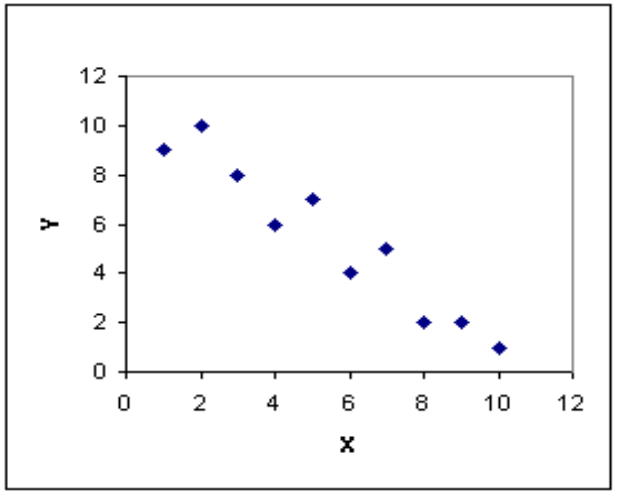
How Is Relationship Measured

- In Pearson's Sample Correlation Coefficient, r measures the direction and the strength of the linear association between two numerical paired variables

Positive Correlation



Negative Correlation



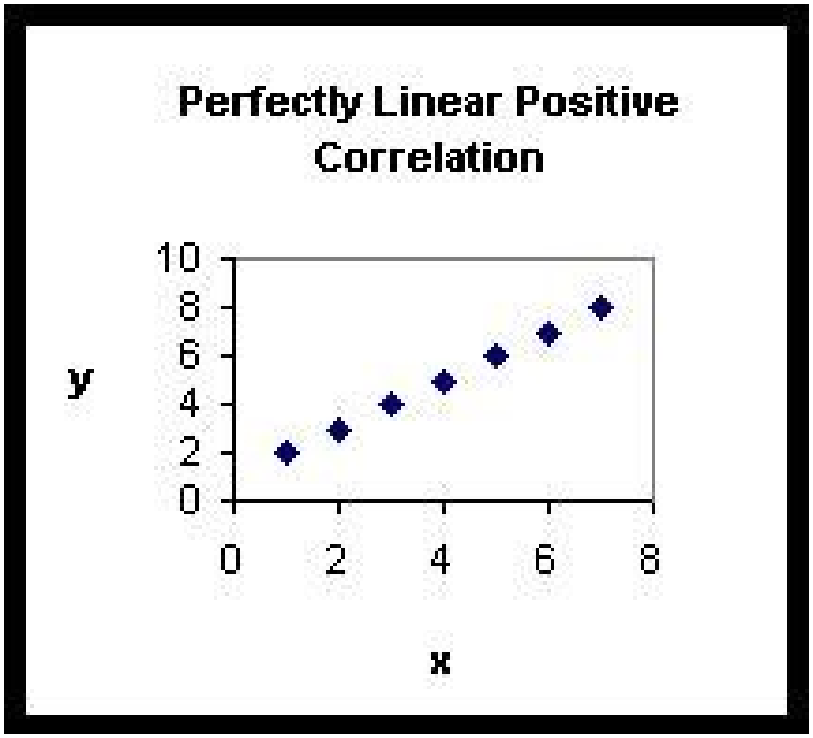
Scatterplot To Analyze Correlation

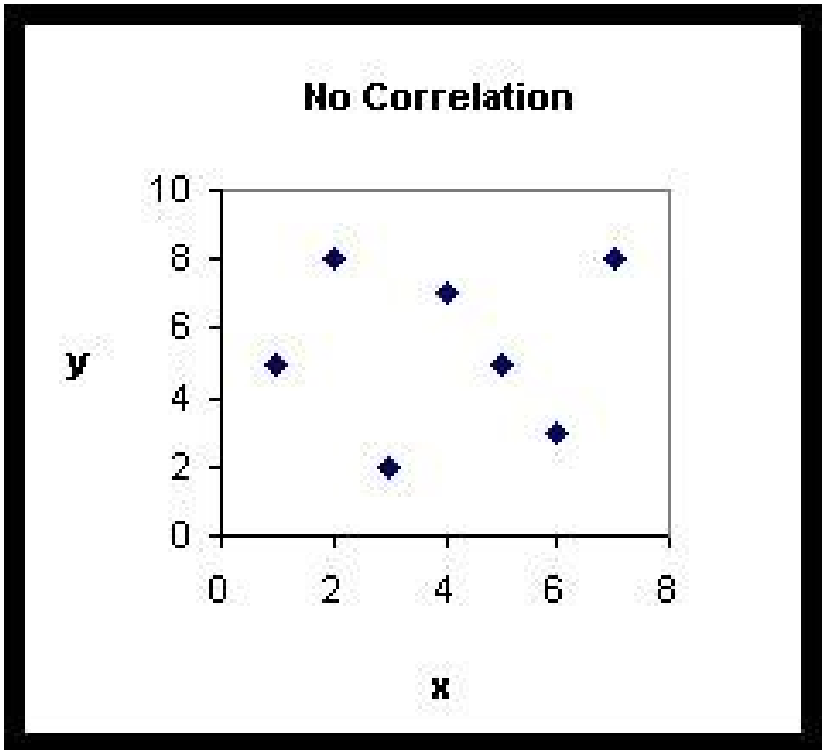
- No Correlation
  - Random or circular assortment of dots
- Positive Correlation
  - ellipse leaning to right
  - GPA and SAT
  - Smoking and Lung Damage
- Negative Correlation
  - ellipse learning to left
  - Depression & Self-esteem
  - Studying & test errors

Strength Of Linear Association

r value	Interpretation
1	perfect positive linear relationship
0	no linear relationship
-1	perfect negative linear relationship

Strength Of Linear Association (Contd.)

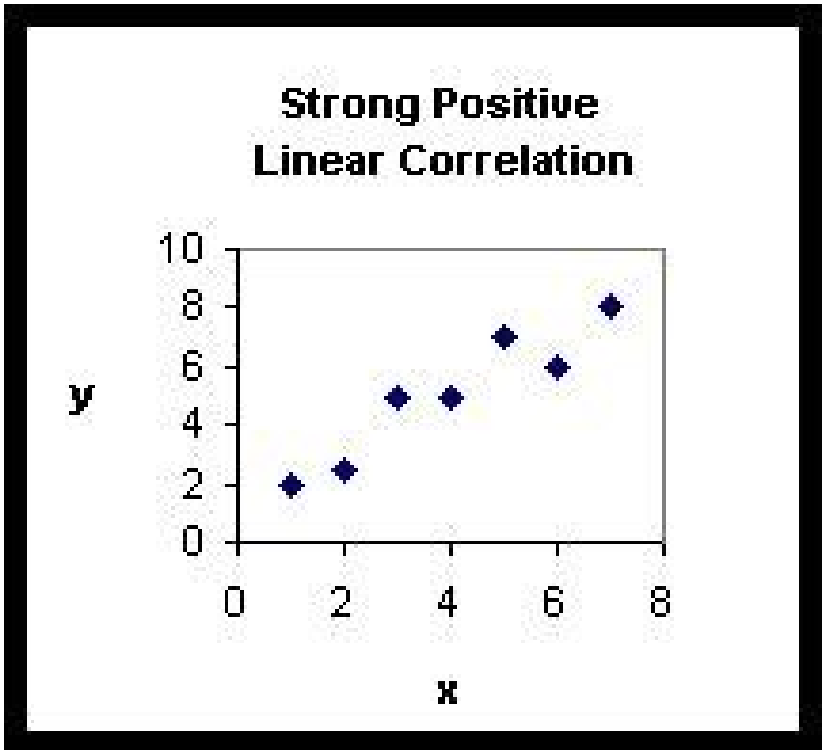


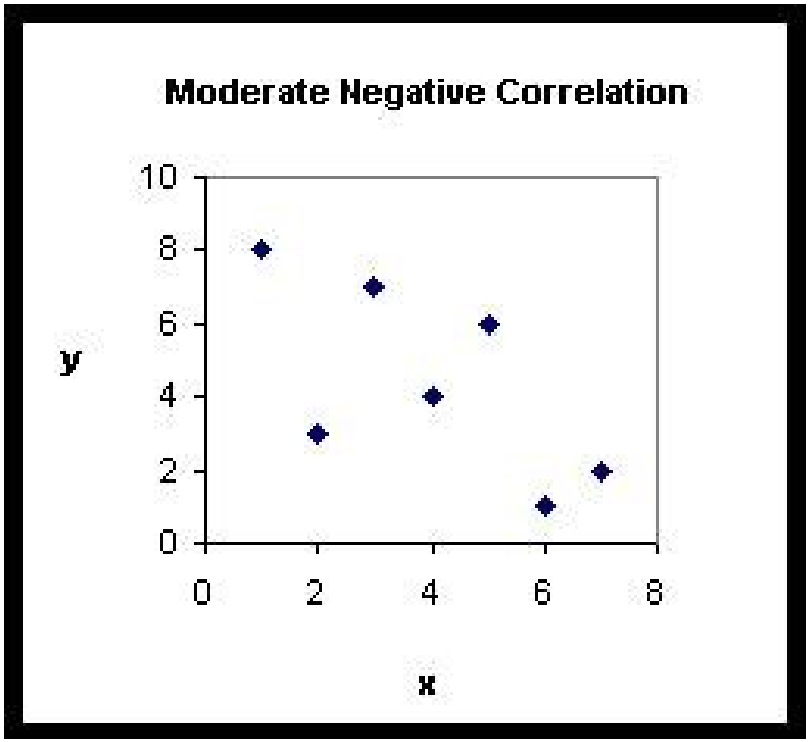


Other Strengths

r value	Interpretation
0.9	strong association
0.5	Moderate association
0.25	Weak association

Examples


Strong Positive
Linear Correlation
y
10
8
6
4
2
0
0
2
4
6
8
x


Moderate Negative Correlation
y
10
8
6
4
2
0
0
2
4
6
8
x

Limitations Of Correlation

linearity

• Can't describe non-linear relationships

e.g., relation between anxiety & performance

Truncation of range

• underestimate strength of relationship if you can't see full range of x value

No proof of causation

- third variable problem
  - could be 3rd variable causing change in both variables
  - directionality: can't be sure which way causality "flows"

Causation

- Causation, also known as cause and effect, is when an observed event or action appears to have caused a second event or action. For example, I bought a brand new bed comforter and placed it in my washing machine to be cleaned. After cleaning the comforter, my washing machine stopped working. I may assume that the first action, washing the comforter, caused the second action i.e the washing machine broke down.

Least-squares Or Regression Line

- These vertical distances, i.e., the distance between y value and their corresponding estimated values on the line are called residuals
- The line which fits the best is called the regression line or, sometimes, the least-squares line
- The line always passes through the point defined by the mean of Y and the mean of X

Linear Regression Model

- The method of least-squares is available in most of the statistical packages (and also on some calculators) and is usually referred to as linear regression

- Y is also known as an outcome variable

- X is also called as a predictor

Estimated Regression Line

$$\hat{y} = \hat{\alpha} + \hat{\beta} x = 1.775351 + 0.330187 x$$

$$\hat{\alpha} = 1.775351 - is.called.y - \text{int}ercept$$

$$\hat{\beta} = 0.330187 - is.called.the.slope$$

Correlation Coefficient, R

- R is a measure of strength of the linear association between two variables, x and y.
- Most statistical packages and some hand calculators can calculate R
- For the data in our Example R=0.94
- R has some unique characteristics

Correlation Coefficient, (Contd.)

- R takes values between -1 and +1
- R=0 represents no linear relationship between the two variables
- R>0 implies a direct linear relationship
- R<0 implies an inverse linear relationship
- The closer R comes to either +1 or -1, the stronger is the linear relationship

Coefficient Of Determination

- R2 is another important measure of linear association between x and y (0 £ R2 £ 1)

- R2 measures the proportion of the total variation in y which is explained by x

- For example r2 = 0.8751, indicates that 87.51% of the variation in BW is explained by the independent variable x (BMI).

Difference Between Correlation And Regression

- Correlation Coefficient, R, measures the strength of bivariate association

- The regression line is a prediction equation that estimates the values of y for any given x

Limitations Of The Correlation Coefficient

- Though R measures how closely the two variables approximate a straight line, it does not make a valid measurement of the strength of nonlinear relationship
- When the sample size, n, is small we also have to be careful with the reliability of the correlation
- Outliers could have a marked effect on R
- Causal Linear Relationship

Multiple Linear Regression

- When there's more than one predictor variable, simple linear regression becomes multiple linear regression, and the analysis becomes more detailed.

- Rather than modeling the mean response as a straight line, as in simple regression, it is now modeled as a function of several explanatory variables.

- Several independent variables may influence the response variable we are trying to study.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Multiple Linear Regression (Contd.)

```
> datavar <- read.csv("stud_reg.csv",header=TRUE)
>
> result1 <- lm(APPLICANTS~PLACE_RATE+ NO_GRAD_STUD
+ ,data=datavar)
>
>
> summary(result1)

Call:
lm(formula = APPLICANTS ~ PLACE_RATE + NO_GRAD_STUD, data = datavar)

Residuals:
     Min       1Q   Median       3Q      Max
-1902.72  -446.67    -1.43   388.72  1701.57

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   806.6124  3764.4756   0.214  0.83393
PLACE_RATE   -140.0849    25.5515  -5.482  0.00014 ***
NO_GRAD_STUD    0.8719     0.2481   3.514  0.00427 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 884.1 on 12 degrees of freedom
Multiple R-squared:  0.813, Adjusted R-squared:  0.7819
F-statistic: 26.09 on 2 and 12 DF,  p-value: 4.27e-05
```

Multiple Linear Regression (Contd.)

Output Interpretation

- From this output, we have determined that the intercept is 806, the coefficient for the placement rate is -140 and the coefficient number of graduating students is 0.9

- Therefore, the complete regression equation is

  APPLICATIONS= 806 + (- 140 * PLACE_RATE) + 0.9 * NO_GRAD_STUD

- This equation tells us that the predicted number of applications for King's College for Master in Analytics will
  - Increase by 169 students for every one percent increase in the placement rate.
  - Increase by 1 student for every one percent increase in number of graduating students

Multiple Linear Regression (Contd.)

Inspecting Results

What is the expected applications given this year's placement rate of 90% and college graduating class of 15000

806.61 – 140.08 * 90 + 0.9 * 15000

=1700 students

Create a table with fitted values and residuals

```
> data.frame(datavar , fitted.value=fitted(result1),residual=resid(result1))
   APPLICANTS PLACE_RATE NO_GRAD_STUD fitted.value     residual
1        5945         61        13742     4243.431  1701.569228
2        6500         50        14744     6658.034  -158.033562
3        5888         53        13588     5229.833   658.166639
4        4000         55        13000     4436.972  -436.971566
5        4700         50        12500     4701.433    -1.433378
6        6300         44        12800     5803.520   496.479753
7        6200         45        13100     5925.013   274.987073
8        7000         44        13850     6719.042   280.958277
9        5000         43        13900     6902.723 -1902.722868
10       3000         57        12000     3284.877  -284.876583
11       1000         62        11000     1712.527  -712.526948
12       4000         55        11531     3156.113   843.886575
13       4600         54        12788     4392.208   207.791696
14       3000         62        13000     3456.377  -456.377378
15       1000         79        13500     1510.897  -510.896958
```

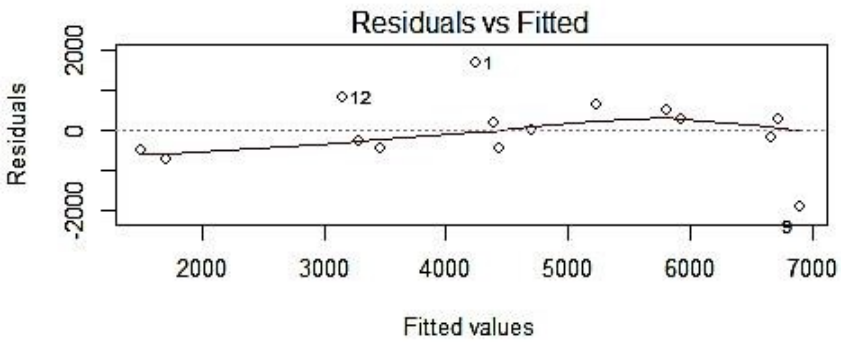
Assumptions Of A Linear Model

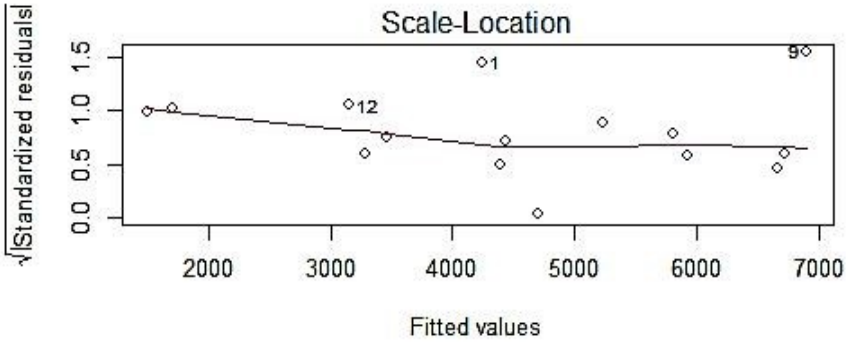
Linear Regression Model Assumptions

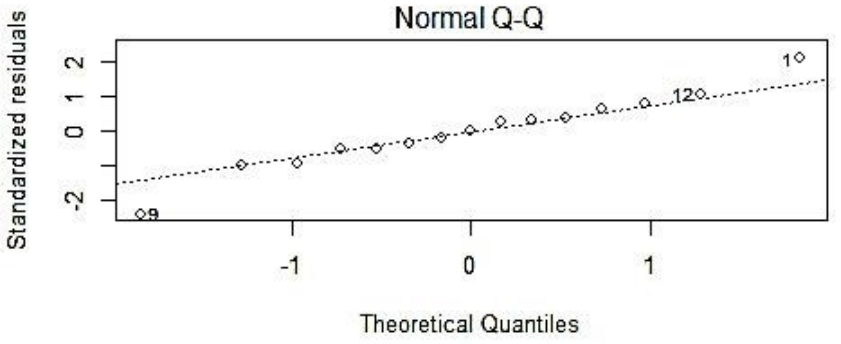
- The regression model is linear in parameters.
- The explanatory variable X is assumed to be nonstochastic.
- The error term, εi, follows a normal distribution.
- Given the value of X, the variance of εi is constant (Homoscedasticity). That is, the conditional variance value of εi is constant.
- There is no autocorrelation between two εi values.
- Low correlation between Xi and εi.
- The X values in a given sample must not be the same.
- There is no perfect multicollinearity (no perfect linear relationship) among explanatory variables.

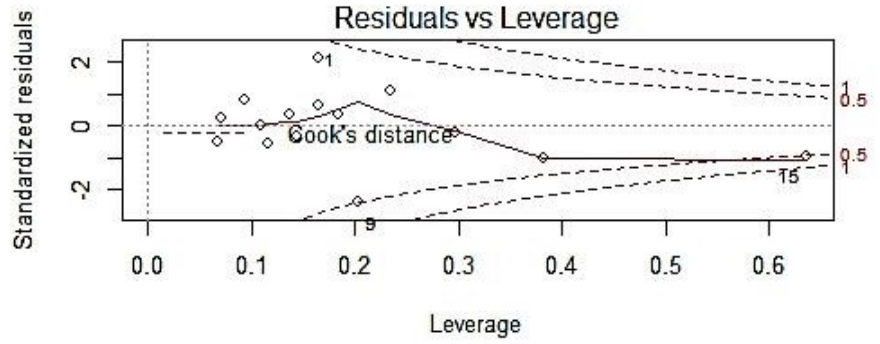
Regression Diagnostics

```
layout(matrix(c(1:4),2,2))
plot(result1)
```









Regression Diagnostics (Contd.)

Multicollinearity

- High correlation between X variables (explanatory variable is called multicollinear).
- Regression coefficients in presence of collinearity measure combined effect.
- Leads to unstable coefficients.
- Always exists; it's a matter of degree

Effects of Multicollinearity

- The variances of regression coefficient estimators are inflated.
- The magnitudes of regression coefficient estimates may be different.
- Adding and removing variables produce large changes in the coefficient estimates.
- Regression coefficient may have negative sign.
- Removing a data point causes large change in the coefficient estimate.

Detection of Collinearity: Simple Signs

- Large changes in the estimated coefficients when a variable is added or deleted
- Large changes in the estimated coefficients when a data point is added or dropped
- Algebraic signs of the estimated coefficients do not conform to prior expectations
- Coefficients of variables that are expected to be important, have large standard errors (small t-values)
- The F ratio might be significant while none of the t-value is significant

Detecting Multicollinearity

- The variance inflation factor (VIF) is a relative measure of the increase in the variance because of collinearity.
- A VIF greater than 10 indicates that collinearity is a problem.
  Few researchers put an upper limit of 4 for VIF.
- Variance inflation factor associated with introducing a new variable Xj is given by:

$$\mathrm{VIF}(X_j) = \frac{1}{1 - R_j^2}$$

R_j^2 is the coefficient of determination for the regression of X_j as dependent variable

Next Class
Assumptions in Regression

Sl. No.	Topics For The Agenda
1.	The Assumptions
2.	Assumption 1
3.	A Look At Normal Distributions
4.	Examining Univariate Distributions
5.	Boxplots
6.	QQ Plot Of Residuals
7.	Regression Example Using R Code
8.	Residuals Are Independent Of X
9.	What About Non-normality
10.	Assumption 2
11.	Heteroscedasticity
12.	Plot Of Pred And Res
13.	Assumption 3
14.	Additivity
15.	Assumption 4
16.	Linearity

Sl. No.	Topics For The Agenda
17.	Detecting Non-Linearity
18.	Residual Plot
19.	Linearity: A Case Of Additivity
20.	Assumption 5
21.	Independence Assumption
22.	Residual Plots
23.	How Does It Arise?
24.	Why Does It Matter?
25.	Result, With Line Of Best Fit
26.	What The Result Shows
27.	Some Difference
28.	Assumption 6
29.	Uncorrelated With The Error Term
30.	No Perfect Multicollinearity
31.	Assumption 8
32.	Mean Of The Error Term = 0


Q & A time