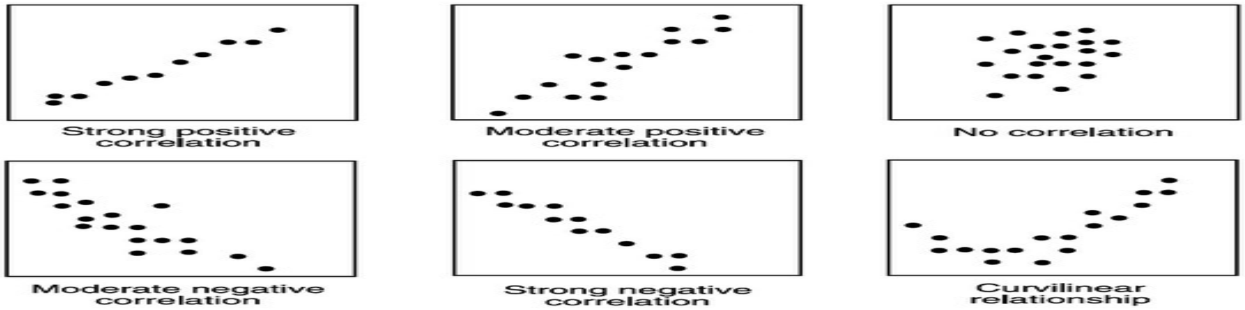
Scatter Diagram

Scatter diagrams or plots provides a graphical representation of the relationship of two continuous variables

Be Careful - Correlation does not guarantee causation. Correlation by itself does not imply a cause and effect relationship!



Strong positive correlation

Moderate positive correlation

No correlation

Moderate negative correlation

Strong negative correlation

Curvilinear relationship

Judge strength of relationship by width or tightness of scatter

Determine direction of the relationship, e.g. If X increases, and Y decreases; it is negative correlation, similarly if X increases, and Y increases, it is positive correlation

Scatter Plot can show Strong positive correlation, Moderate positive correlation, No correlation, Moderate negative Correlation, Strong

Negative correlation, curvilinear relation.

Slide-98

Correlation Analysis

Correlation Analysis measures the degree of linear relationship between two variables

Range of correlation coefficient	-1 to +1
Perfect positive relationship	+1

Perfect negative relationship	-1
No Linear relationship	0

If the absolute value of the correlation coefficient is greater than 0.85, then we say there is a good relationship

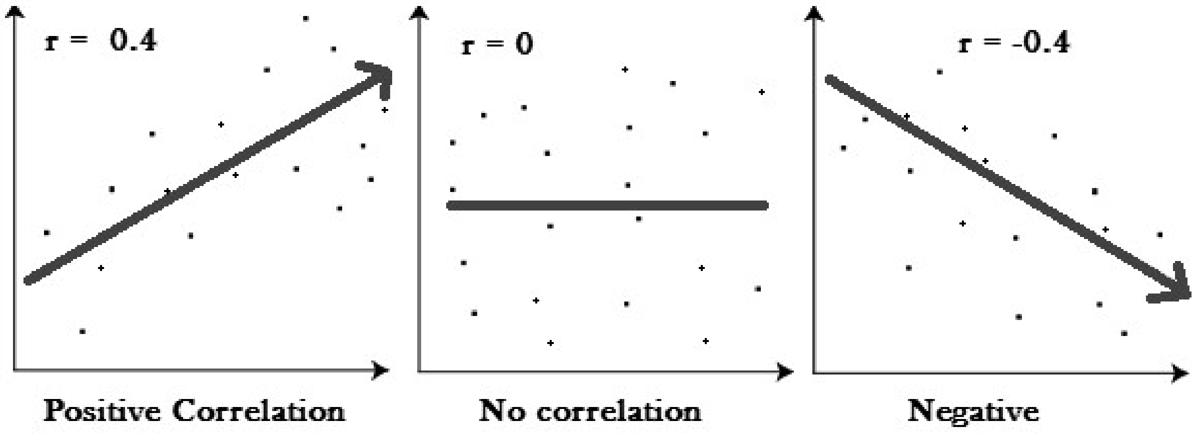
- Example: r = 0.87, r = -0.9, r = 0.9, r = -0.87 describe good relationship
- Example: r = 0.5, r = -0.5, r = 0.28 describe poor relationship

Correlation values of -1 or 1 imply an exact linear relationship. However, the real value of correlation is in quantifying less than perfect relationships

We can perform regression analysis, which attempts to further describe this type of relationship, if the correlation is good between the 2 variables

Slide- 99

Correlation Analysis:



Positive correlation: r>0

Negative correlation: r<0

No correlation: $r=0$

$r = (n (\sum xy)-(\sum x) (\sum y))/ (\text{sqrt} ([n\sum x^2-(\sum x)^2] [n\sum y^2-(\sum y)^2]))$

Slide-100

Linear Regression Model

The equation that represents how an independent variable is related to a dependent variable and an error term is a regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

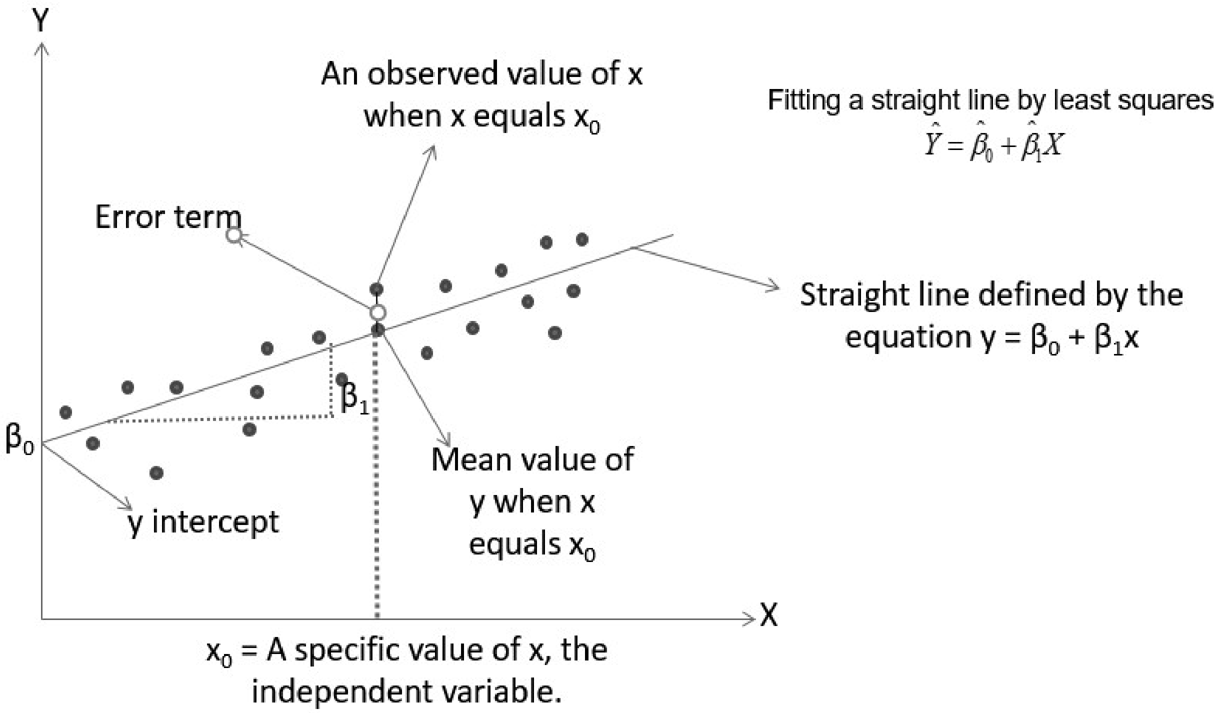
Where, β0 and β_1 are called parameters of the model,

ε is a random variable called error term.

$\beta\ o= ((\sum y)(\sum x^2)-(\sum x) (\sum xy))/ ([n\sum x^2-(\sum x)^2])$

$\beta\ 1= ((\sum xy)-(\sum x) (\sum y))/ ([n\sum x^2-(\sum x)^2])$

Slide-101


Y
An observed value of x when x equals x_0
Fitting a straight line by least squares
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
Error term
Straight line defined by the equation $y = \beta_0 + \beta_1 x$
β_1
β_0
Mean value of y when x equals x_0
y intercept
X
x_0 = A specific value of x, the independent variable.

Slide-102

Regression Analysis

R-squared-also known as Coefficient of determination, represents the % variation in output (dependent variable) explained by input variables/s or Percentage of response variable variation that is explained by its relationship with one or more predictor variables

Higher the R^2, the better the model fits your data

R^2 is always between 0 and 100%

R squared is between 0.65 and 0.8 => Moderate correlation

R squared in greater than 0.8 => Strong correlation

R^2=SSR/SST= (SSR/(SSR+SSE))

0<=R^2<=1

Mathematically

SSR =$\sum(\hat{y}$-$\overline{y})^2$ → measure of an explained variation

SSE =$\sum(y - \hat{y})^2$→ measure of an unexplained variation

SST = SSR+SSE =$\sum$(y-$\overline{y})^2$ → measure of total variation in y

Slide-103

Regression Analysis

Prediction and Confidence Interval are types of confidence intervals used for predictions in regression and other linear models

Prediction Interval: Represents a range that a single new observation is likely to fall given specified settings of the predictors

Confidence interval of the prediction: Represents a range that the mean response is likely to fall given specified settings of the predictors

The prediction interval is always wider than the corresponding confidence interval because of the added uncertainty involved in predicting a single response versus the mean response

Slide-104

Regression Techniques – Simple Linear Regression

Y-continuous, x – single & continuous

We apply simple linear Regression

Y-continuous, x – single & discrete

We create dummy variable for discrete component and

We then apply simple linear Regression

Simple Linear Regression – Dummy Variable

Slide-105

Example:

Gender	**Dummy Variable**
Male	1
Female	0
Male	1
Female	0
Male	1

Slide-106

Simple Linear Regression – R

<u>A business problem:</u>

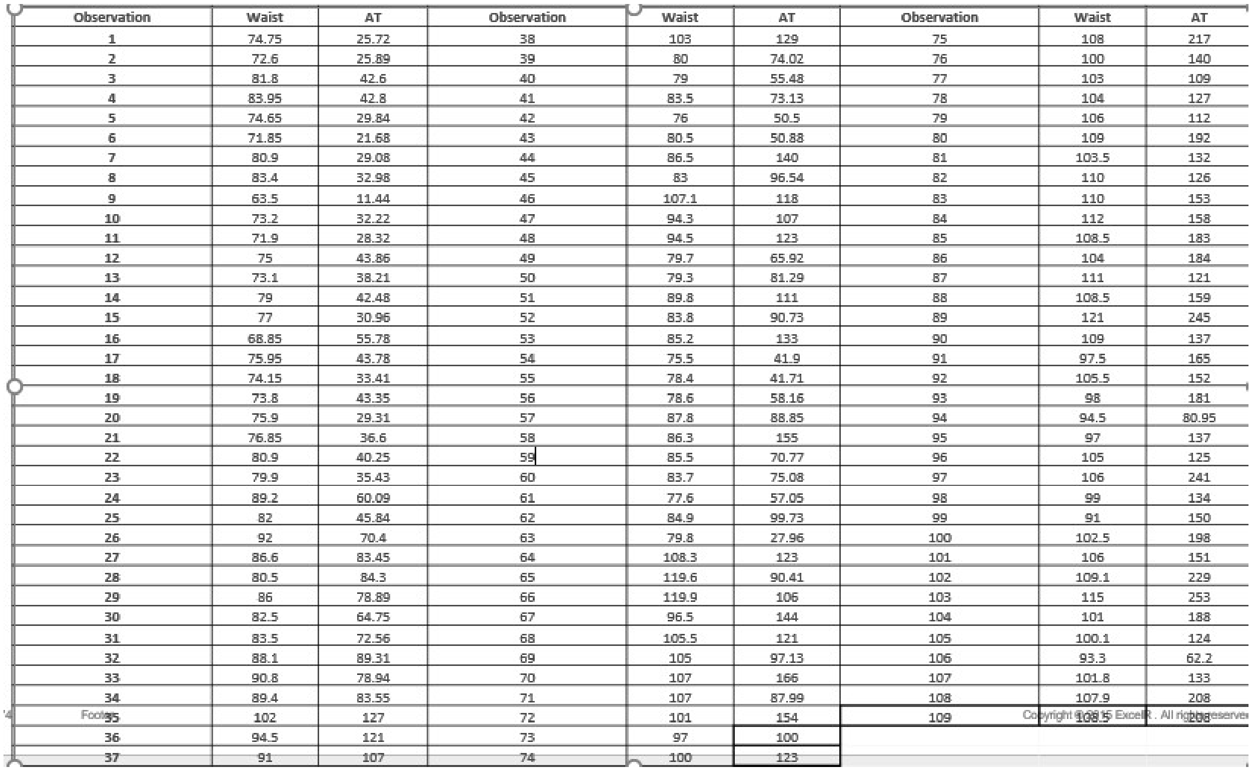
The Waist Circumference – Adipose Tissue data

- Studies have shown that individuals with excess Adipose tissue (AT) in the abdominal region have a higher risk of cardio-vascular diseases
- Computed Tomography, commonly called the CT Scan is the only technique that allows for the precise and reliable measurement of the AT (at any site in the body)
- The problems with using the CT scan are:
    - Many physicians do not have access to this technology
    - Irradiation of the patient (suppresses the immune system)

  - Expensive
- Is there a simpler yet reasonably accurate way to predict the AT area? i.e.,
  - Easily available
  - Risk free
  - Inexpensive
- A group of researchers conducted a study with the aim of predicting abdominal AT area using simple anthropometric measurements, i.e., measurements on the human body
- The Waist Circumference – Adipose Tissue data is a part of this study wherein the aim is to study how well waist circumference (WC) predicts the AT area

Side-107

Simple Linear Regression – Data Set



Observation	Waist	AT	Observation	Waist	AT	Observation	Waist	AT
1	74.75	25.72	38	103	129	75	108	217
2	72.6	25.89	39	80	74.02	76	100	140
3	81.8	42.6	40	79	55.48	77	103	109
4	83.95	42.8	41	83.5	73.13	78	104	127
5	74.65	29.84	42	76	50.5	79	106	112
6	71.85	21.68	43	80.5	50.88	80	109	192
7	80.9	29.08	44	86.5	140	81	103.5	132
8	83.4	32.98	45	83	96.54	82	110	126
9	63.5	11.44	46	107.1	118	83	110	153
10	73.2	32.22	47	94.3	107	84	112	158
11	71.9	28.32	48	94.5	123	85	108.5	183
12	75	43.86	49	79.7	65.92	86	104	184
13	73.1	38.21	50	79.3	81.29	87	111	121
14	79	42.48	51	89.8	111	88	108.5	159
15	77	30.96	52	83.8	90.73	89	121	245
16	68.85	55.78	53	85.2	133	90	109	137
17	75.95	43.78	54	75.5	41.9	91	97.5	165
18	74.15	33.41	55	78.4	41.71	92	105.5	152
19	73.8	43.35	56	78.6	58.16	93	98	181
20	75.9	29.31	57	87.8	88.85	94	94.5	80.95
21	76.85	36.6	58	86.3	155	95	97	137
22	80.9	40.25	59	85.5	70.77	96	105	125
23	79.9	35.43	60	83.7	75.08	97	106	241
24	89.2	60.09	61	77.6	57.05	98	99	134
25	82	45.84	62	84.9	99.73	99	91	150
26	92	70.4	63	79.8	27.96	100	102.5	198
27	86.6	83.45	64	108.3	123	101	106	151
28	80.5	84.3	65	119.6	90.41	102	109.1	229
29	86	78.89	66	119.9	106	103	115	253
30	82.5	64.75	67	96.5	144	104	101	188
31	83.5	72.56	68	105.5	121	105	100.1	124
32	88.1	89.31	69	105	97.13	106	93.3	62.2
33	90.8	78.94	70	107	166	107	101.8	133
34	89.4	83.55	71	107	87.99	108	107.9	208
35	102	127	72	101	154	109	108.5	208
36	94.5	121	73	97	100			
37	91	107	74	100	123			

Slide- 108

Simple Linear Regression – Transformation

```
reg <- lm(AT ~ Waist)                    # Linear Regression

summary(reg)

confint(reg, level=0.95)

predict(reg, interval="predict”)

reg_log <- lm(AT ~ log(Waist))       # Regression using Logarithmic Transformation

summary(reg_log)

confint(reg_log, level=0.95)

predict(reg, interval="predict”)
```

```
reg_exp <- lm(log(AT) ~ Waist)     # Regression using Exponential Transformation

summary(reg_exp)

confint(reg_exp, level = 0.95)

predict(reg, interval="predict”)
```

Slide-109

Regression Techniques – Multiple Linear Regression

Y-continuous, x – Multiple & continuous

We apply Multiple linear Regression

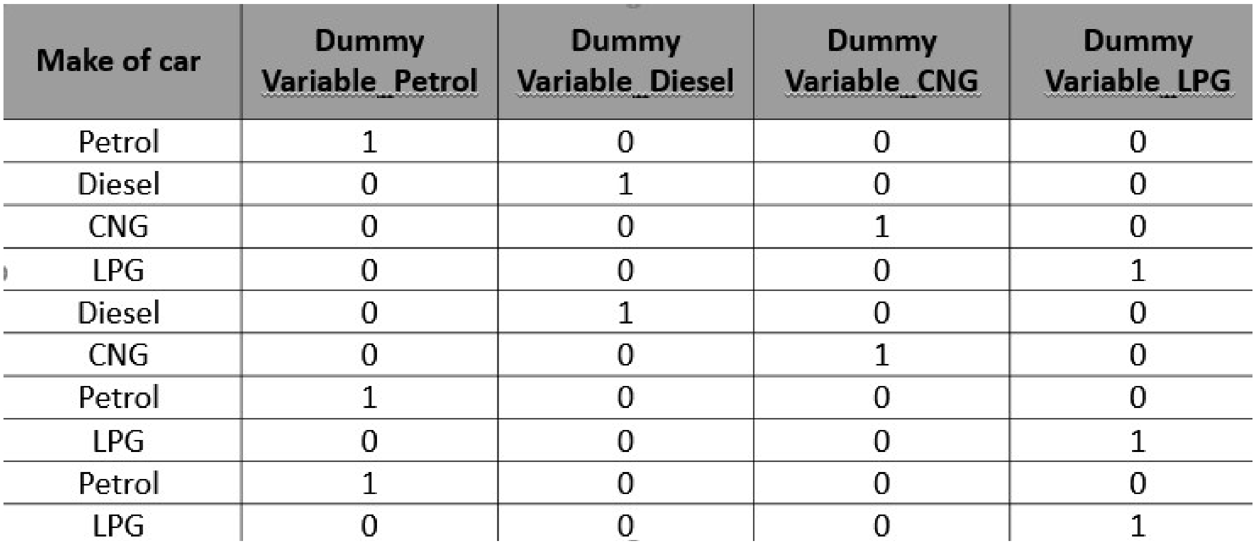
Y-continuous, x – Multiple & discrete

We create dummy variable for discrete component and

We then apply Multiple linear Regression

Slide-110

Multiple Linear Regression – Dummy Variable



Make of car	Dummy Variable_Petrol	Dummy Variable_Diesel	Dummy Variable_CNG	Dummy Variable_LPG
Petrol	1	0	0	0
Diesel	0	1	0	0
CNG	0	0	1	0
LPG	0	0	0	1
Diesel	0	1	0	0
CNG	0	0	1	0
Petrol	1	0	0	0
LPG	0	0	0	1
Petrol	1	0	0	0
LPG	0	0	0	1

Slide- 111

Multiple Regression Model

DATA : CARS, 81 observations, *“cars.csv”*

- VOL = cubic feet of cab space

- HP = engine horsepower
- MPG = average miles per gallon
- SP = top speed, miles per hour
- WT = vehicle weight, hundreds of pounds

Our interest is to model the MPG of a car based on the other variables.

Slide-112

Model and Assumptions

Our Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$

Linear

Independent

Normal

Equal Variance

Linearity (Assumptions about the form of the model):

- Linear in parameters
- Assumptions about the errors:
- IID Normal (Independently & identically distributed)
- Zero mean
- Constant variance (Homoscedasticity)
- If no constant variance (HETEROSCEDASTICITY)
- Independent of each other. If not independent, it is called as AUTO CORRELATION problem
- Assumptions about the predictors:
- Non-random
- Measured without error
- Linearly independent of each other. If not it is called as COLLINEARITY problem
- Assumptions about the observations:
- Equally reliable