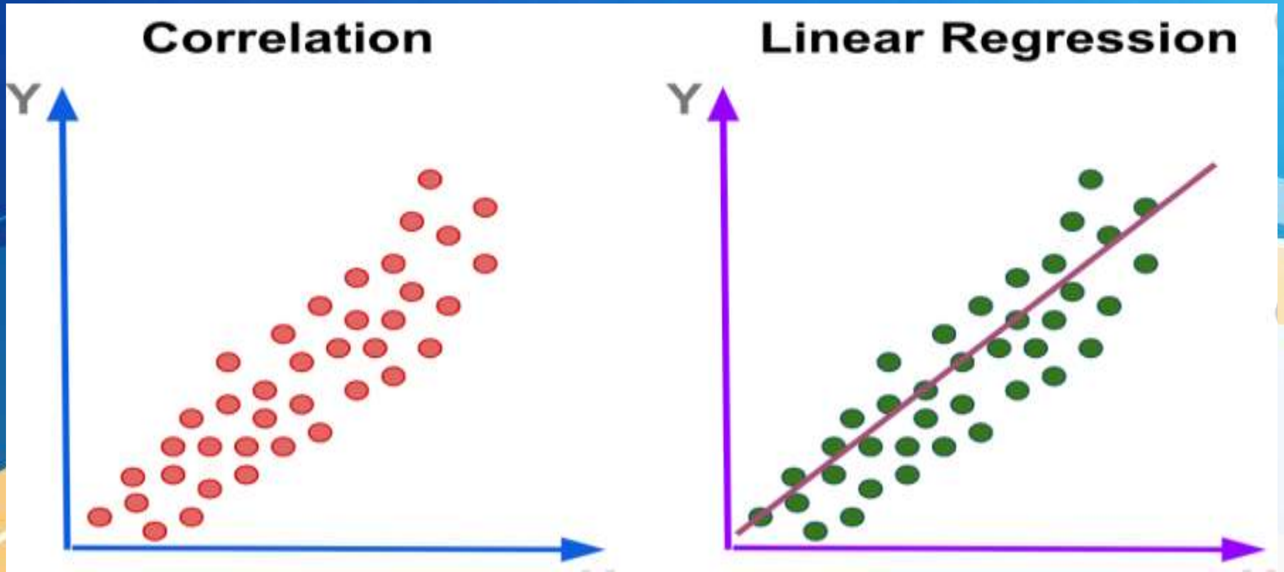
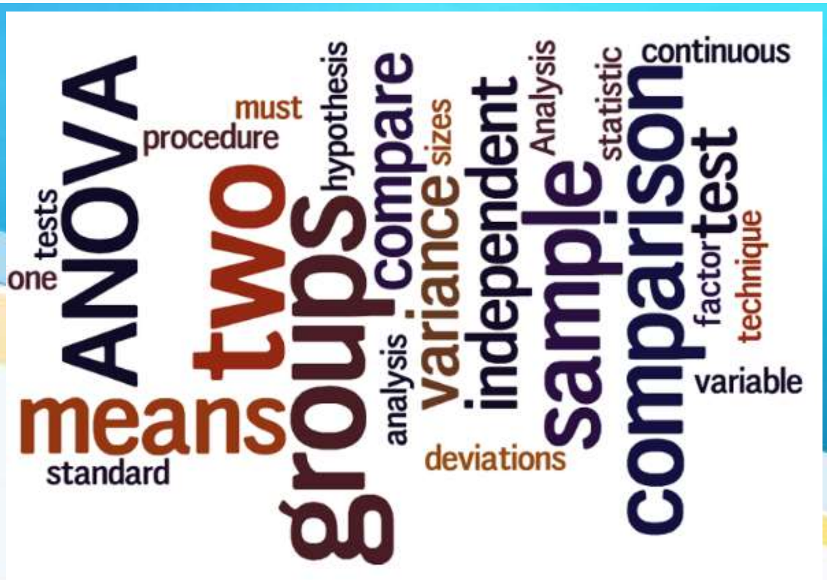
Indian College of Radiation Oncology (ICRO)
Association of Radiation Oncologists of India (AROI)

48th ICRO-SUN PG TEACHING PROGRAMME
26 & 27 OCTOBER, 2024
MAX SUPERSPECIALITY HOSPITAL, BATHINDA

CLINICAL TRIAL & CANCER STATISTICS





Correlation & Regression

Dr Rohit Malde MD, DNB. FRCR(UK)
Consultant Clinical Oncologist
Nanavati Max Institute Cancer care, Mumbai

Learning Objectives

Introduction
What is Correlation ?
Types of Correlation
- Parametric & Non Parametric

What is Regression ?
Correlation Analysis
- Pearsons Correlation
- Kendall tau correlation
- Spearman Correlation

Regression Analysis
- Simple Model

Concept of ANOVA
- Multiple Regression Analysis

Applications / Examples in Oncology Practice

Introduction

- Prediction consists of learning from data.
- Predicting the outcomes of a random process is based on observations
- Observations are independent realizations of the same random process; each observation is made of one or several variables.
- Variables are either numbers, or elements belonging to a finite set "finite number of values”
- One variable = Univariate ; 2 variables = Bivariate , > 2 variables = Multivariate
- Interaction between variables cannot be explored. For this we need to understand about Correlation & Regression

What is Correlation

Correlation quantifies the *Degree and Direction* to which two variables are related.

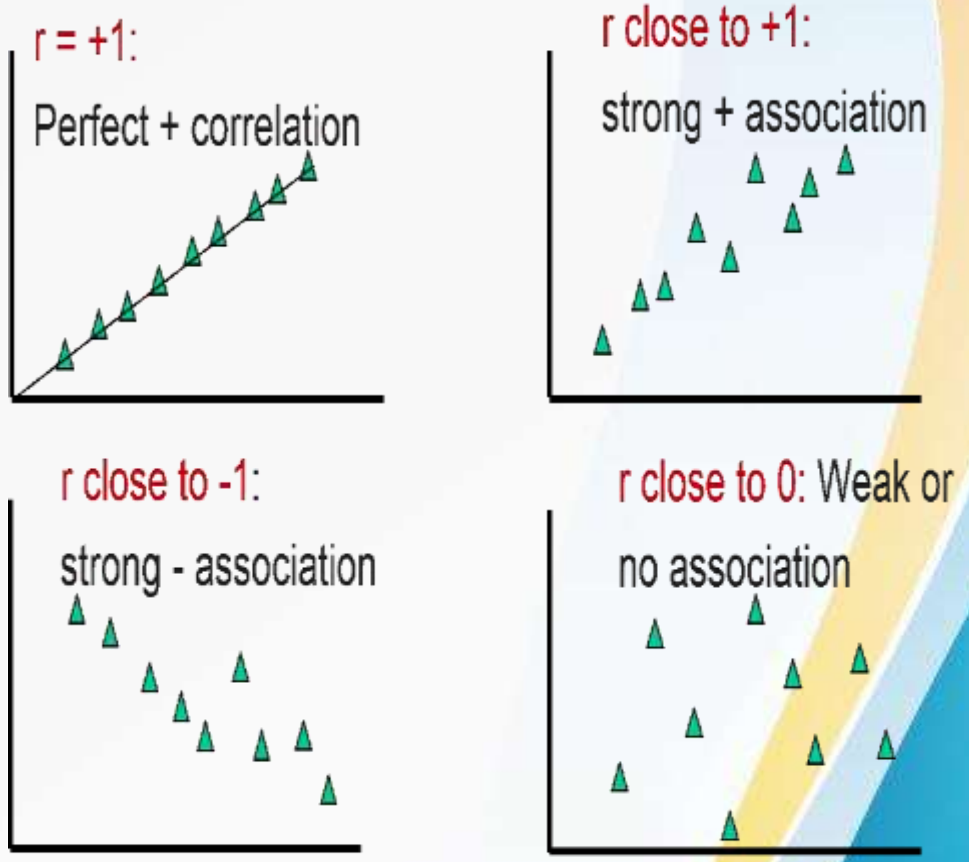
Correlation does not fit a line through the data points.

But simply is computing a correlation coefficient tells how much one variable tends to change when the other one does.

When r is 0.0, there is no relationship.

When r is positive, there is a trend that one variable goes up as the other one goes up.

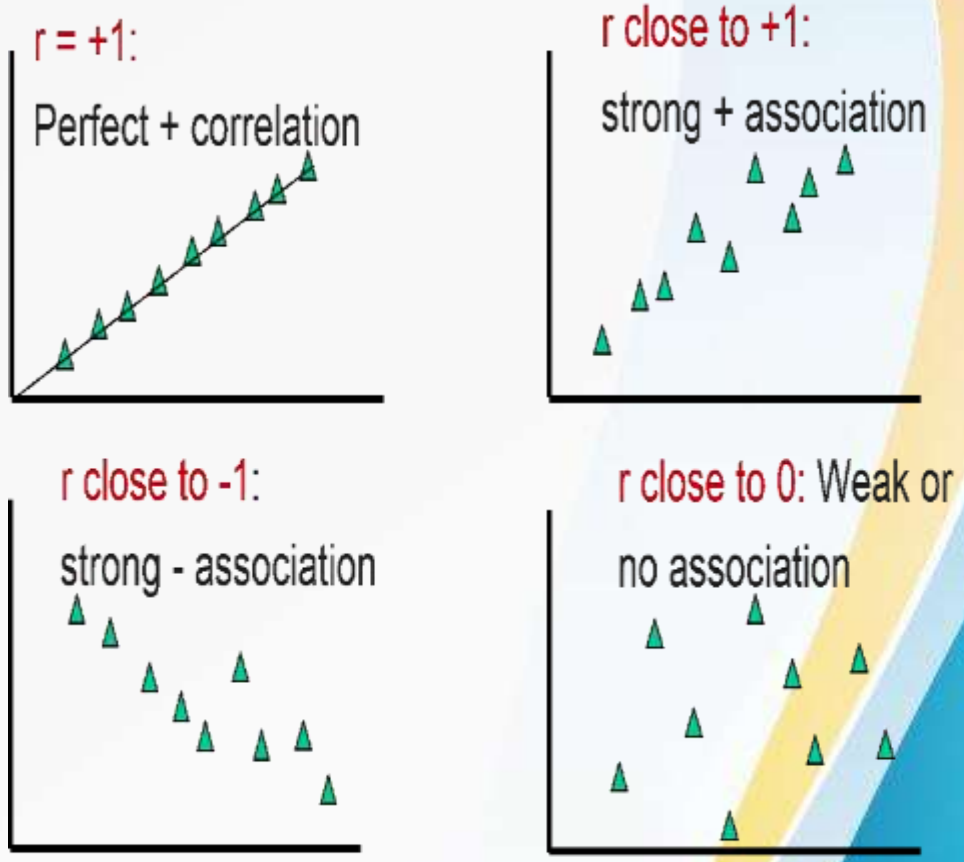
When r is negative, there is a trend that one variable goes up as the other one goes down.



What is Correlation

- If the two variables swapped the degree of correlation coefficient will be the same.
- Variables can be dependent or Independent

Correlation Coefficient (r)	Description (Rough Guideline)
+1.0	Perfect positive + association
+0.8 to 1.0	Very strong + association
+0.6 to 0.8	Strong + association
+0.4 to 0.6	Moderate + association
+0.2 to 0.4	Weak + association
0.0 to +0.2	Very weak + or no association
0.0 to -0.2	Very weak - or no association
-0.2 to – 0.4	Weak - association
-0.4 to -0.6	Moderate - association
-0.6 to -0.8	Strong - association
-0.8 to -1.0	Very strong - association
-1.0	Perfect negative association



Three types of Correlation

1. Pearson Correlation
2. Kendall Rank Correlation
3. Spearman Correlation

Which Correlation Test to use ?

Parametric tests -- Use Pearson Correlation coefficient here

1. Assumes data has been randomly selected from the population
2. The variables have a normal distribution
3. Asscoiation of data is homoscedastic (homogenous) [Std deviation is same]
4. Data is measured using an interval or ratio scale.

Non- Parametric test - use Spearman or Kendall Tau Coefficient

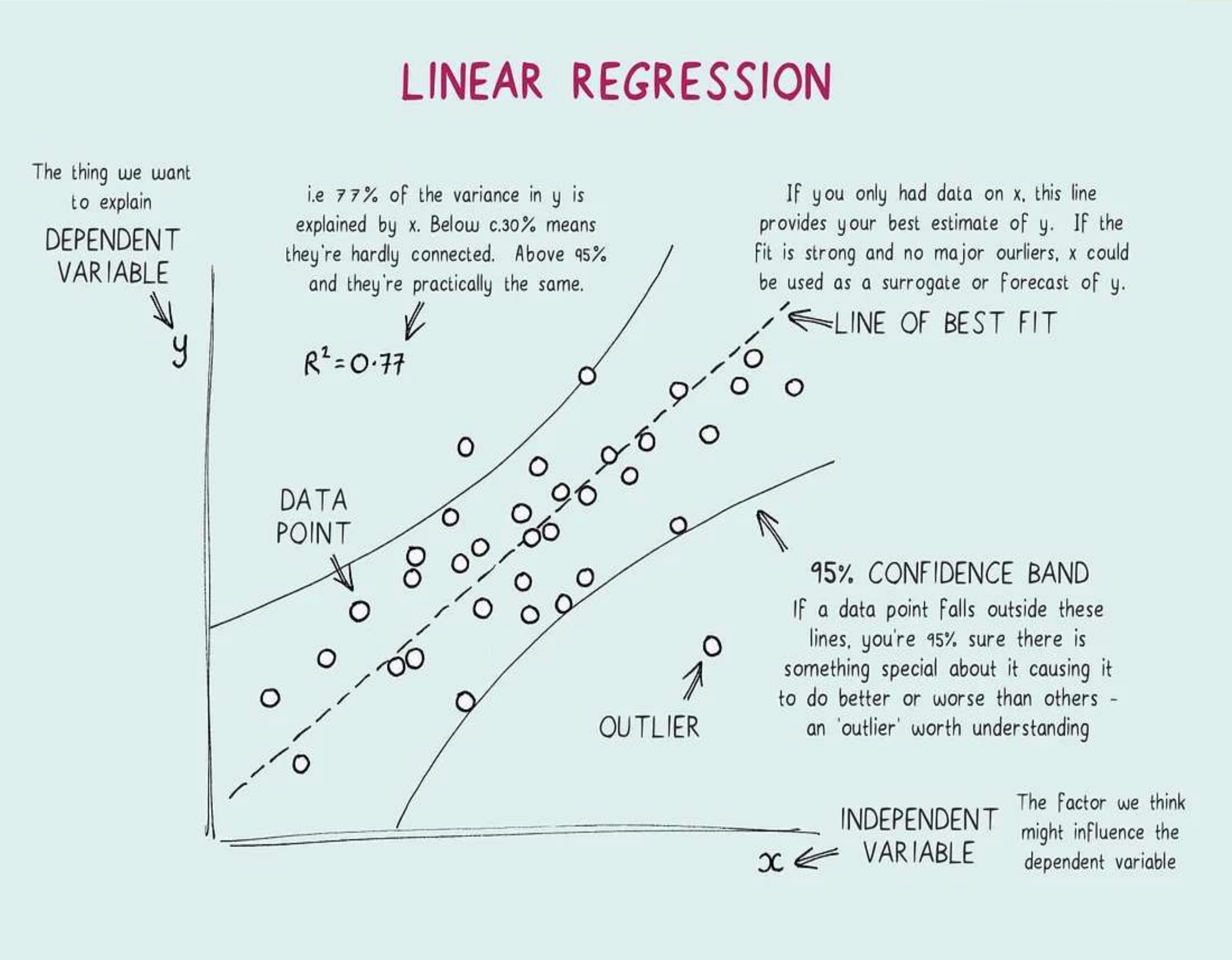
1. Data may be selective and skewed (Selective Population)
2. Variable have Skewed distribution
3. Association of data is heteroscedastic (inhomogenous) [Std deviation is varying]
4. Data is measured with nominal or ordinal scale

What is Linear Regression

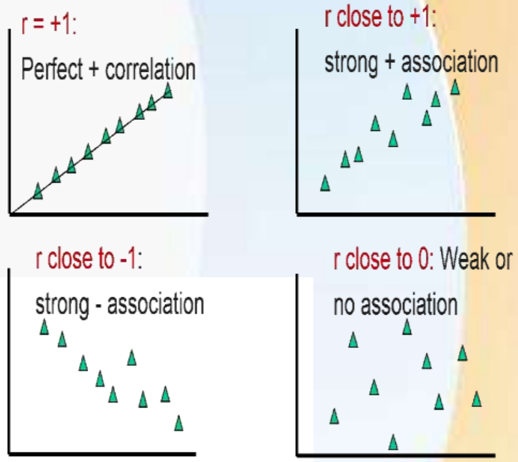
Linear regression finds the best line that predicts dependent variable from independent variable

Linear regression quantifies goodness of fit with R^2

The line that best predicts independent variable from dependent variable is not the same as the line that predicts dependent variable from independent variable. The data cannot be swapped



Pearson Correlation - (Parametric Test)



The Pearson correlation coefficient is given by the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where $\bar{x}$ is the mean of variable x values, and $\bar{y}$ is the mean of variable y values.

A study is conducted involving 10 students to investigate the association between statistics and science tests. The question arises here; is there a relationship between the degrees gained by the 10 students in statistics and science tests?

Table (2.1) Student degree in Statistic and science

Students	1	2	3	4	5	6	7	8	9	10
Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

Notes: the marks out of 30

Suppose that (x) denotes for statistics degrees and (y) for science degree

Calculating the mean $(\bar{x}\ ,\ \bar{y})$;

$$\bar{x}=\frac{\sum x}{n}=\frac{173}{10}=17.3\ ,\quad \bar{y}=\frac{\sum y}{n}=\frac{200}{10}=20$$

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}=\frac{228}{\sqrt{356.1}\sqrt{252}}$$

$$=\frac{228}{(18.8706)(15.8745)}=\frac{228}{299.5614}=0.761$$

Table (2.2) Calculating the equation parameters

Statistics	Science					
x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
20	20	2.7	7.29	0	0	0
23	25	5.7	32.49	5	25	28
8	11	-9.3	86.49	-9	81	83
29	24	11.7	136.89	4	16	46
14	23	-3.3	10.89	3	9	-9.9
12	16	-5.3	28.09	-4	16	21.2
11	12	-6.3	39.69	-8	64	50.4
21	21	3.7	13.69	1	1	3.7
17	22	-0.3	0.09	2	4	-0.6
18	26	0.7	0.49	6	36	4.2
173	200	0	356.1	0	252	228

$$\sum(x-\bar{x})^2=356.1\ ,\ \sum(y-\bar{y})^2=252\ ,$$

$$\sum(x-\bar{x})(y-\bar{y})=228$$

The calculation shows a strong positive correlation (0.761) between the student's statistics and science degrees. This means that as degrees of statistics increases the degrees of science increase also. Generally the student who has a high degree in statistics has high degree in science and vice versa.

Kendall tau Rank Correlation - (Non-Parametric Test)

The following formula is used to calculate the value of Kendall rank correlation:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

τ = Kendall rank correlation coefficient
nc = number of concordant (Ordered in the same way).
nd= Number of discordant (Ordered differently)

Let x1, …, x_n be a sample for random variable x and
Let y1, …, y_n be a sample for random variable y of the same size n.

Select distinct pairs according to rank (x1 , y1) and (x1 , y2), (x2,y2) ,……..

For any such assignment of pairs, define each pair as concordant, discordant or neither as follows:
Concordant (C) if (x1> x2 and y1> y2) or (x1 < x2 and y1 < y2)
Discordant (D) if (xi > xj and yi < yj) or (xi < xj and yi > yj)
Neither if x1 = x2 or y1 = y2 (i.e. ties are not counted).

Students	1	2	3	4	5	6	7	8	9	10
Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

Set rank to the data

data			
statistics (degree)	science (degree)	Rank (statistics)	Rank (science)
20	20	4	7
23	25	2	2
8	11	10	10
29	24	1	3
14	23	7	4
12	16	8	8
11	12	9	9
21	21	3	6
17	22	6	5
18	26	5	1

Arranged Rank	
Rank (science)	Rank (statistics)
1	5
2	2
3	1
4	7
5	6
6	3
7	4
8	8
9	9
10	10

	Concordant	Discordant
1	5	4
2	7	1
3	7	0
4	3	3
5	3	2
6	4	0
7	3	0
8	3	0
9	2	0
10	1	0

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

$$\tau = \frac{35 - 10}{\frac{1}{2} * 10(10-1)}$$

$$\tau = \frac{25}{45} = 0.556$$

Kendall's Tau coefficient τ = 0.556; this indicates a moderate positive relationship between the ranks individuals obtained in the statistics and science exam. This means the higher you ranked in statistics, the higher you ranked in science also, and vice versa.

Spearman Rank Correlation - (Non-Parametric Test).

Spearman rank correlation test does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal

ρ = Spearman rank correlation coefficient
di= the difference between the ranks of corresponding values Xi and Yi
n= number of value in each data set

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Students	1	2	3	4	5	6	7	8	9	10
Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

Calculating the Parameters of Spearman rank Equation:

statistics (degree)	science (degree)	Rank (statistics)	Rank (science)	|d|	d^2
20	20	4	7	3	9
23	25	2	2	0	0
8	11	10	10	0	0
29	24	1	3	2	4
14	23	7	4	3	9
12	16	8	8	0	0
11	12	9	9	0	0
21	21	3	6	3	9
17	22	6	5	1	1
18	26	5	1	4	16

$$\sum d_i^2 = 9+0+0+4+9+0+0+9+1+16 = 48$$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad ; \quad \rho = 1 - \frac{6*48}{10(10^2-1)}$$

$$\rho = 1 - \frac{288}{990} \quad ; \quad \rho = 1 - 0.2909$$

$$\rho = 0.71$$

indicates a strong positive relationship between the ranks individuals obtained in the statistics and science exam

Regression Analysis

- Involves identifying and evaluating the relationship between a dependent variable and one or more independent variables, which are also called predictor or explanatory variables.

- Useful to assess and adjusting for confounding

- Explores relationships that can be readily described by straight lines or their generalization to many dimensions

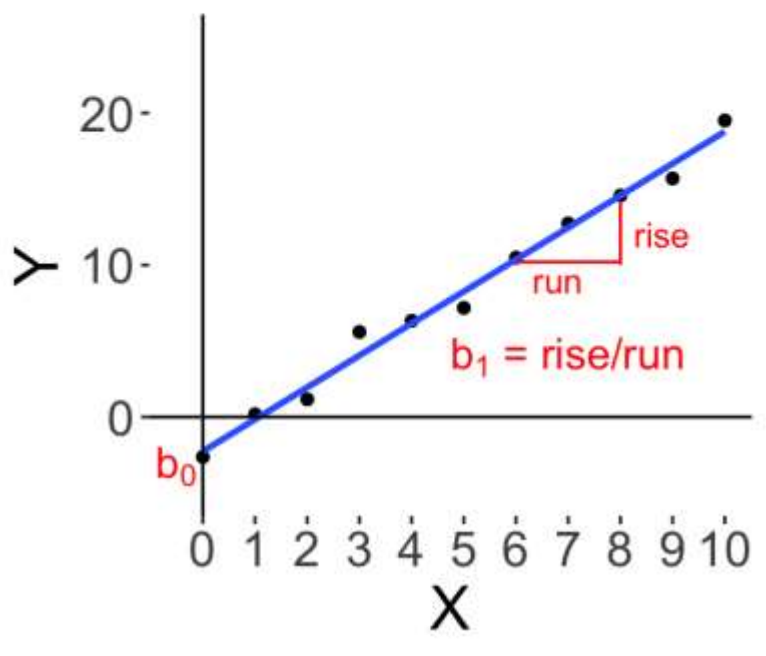
- Single continuous dependent variable + single independent variable, the analysis is called a simple linear regression analysis

- Multiple regression explores relationship between several independent or predictor variables and a dependent variable.

Linear Regression Analysis

In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect. Least squares linear regression is a method for predicting the value of a dependent variable y, based on the value of an independent variable x.

$$\mathbf{y} = \beta 0 \pm \beta 1\ \boldsymbol{x}1$$



Where

- x independent variable.
- $\mathbf{y}$ dependent variable.
- β_1 The Slope of the regression line
- $\boldsymbol{\beta_0}$ The intercept point of the regression line and the y axis.

$$\boldsymbol{\beta_1} = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

- $\boldsymbol{n}$ Number of cases or individuals.
- $\sum \boldsymbol{xy}$ Sum of the product of dependent ar independent variables.
- $\sum \boldsymbol{x}$ = Sum of independent variable.
- $\sum \mathbf{y}$ = Sum of dependent variable.
- $\sum \boldsymbol{x}^2$ = Sum of square independent variable.

$$\boldsymbol{\beta_0} = \bar{\mathbf{y}} - \boldsymbol{\beta_1}\bar{\boldsymbol{x}}$$

Linear Regression Analysis

A study is conducted involving 10 patients to investigate the Relationship & Effects of patient's Age and their BP

calculating the linear regression of patient's age and blood pressure

Obs	Age x	BP y	xy	x^2
1	35	112	3920	1225
2	40	128	5120	1600
3	38	130	4940	1444
4	44	138	6072	1936
5	67	158	10586	4489
6	64	162	10368	4096
7	59	140	8260	3481
8	69	175	12075	4761
9	25	125	3125	625
10	50	142	7100	2500
Total	491	1410	71566	26157

Required calculation

$\sum x = 491$

$\sum y = 1410$

$\sum xy = 71566$

$\sum x^2 = 26157$

Calculating the mean $(\bar{x}\ ,\ \bar{y})$;

$$\bar{x} = \frac{\sum x}{n} = \frac{491}{10} = 49.1\ ,\ \bar{y} = \frac{\sum y}{n} = \frac{1410}{10} = 141$$

Calculating the regression coefficient;

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \qquad \beta_1 = \frac{10 * 71566 - 491 * 1410}{10 * 26157 - (491)^2}$$

$$\beta_1 = \frac{715660 - 692310}{261570 - 241081} \qquad \beta_1 = \frac{23350}{20489} = 1.140$$

$$\beta_0 = \bar{y} - \beta_1\bar{x} \qquad \beta_0 = 141 - 1.140 * 49.1$$

$$\beta_0 = 141 - 55.974 \qquad \beta_0 = 85.026$$

Estimated blood pressure $(\hat{Y})$ = 85.026 + 1.140 age

$\beta 0$ = 85.026 indicates that blood pressure at age zero.

Regression coefficient $\beta 1$ = 1.140 indicates that as age increase by 1 year the blood pressure increase by 1.140

ANOVA (Analysis of variance) Test

A statistical method that examine whether there are significant differences in the means among three or more groups.

By evaluating the variance within and between groups, ANOVA helps determine if the observed distinctions likely stem from genuine group variations or mere chance.

It’s frequently used in experimental studies to assess how independent variables impact a dependent variable.

Types:

- One-Way ANOVA: Compares means of three or more independent groups based on one independent variable.
- Two-Way ANOVA: Examines the influence of 2 independent variables on a dependent variable, and can assess interaction effects.

$$F = \frac{Variance\ between\ groups}{Variance\ within\ groups}$$

Error / random chance differences among individuals within single groups

ANOVA Table

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma nj(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	$MSB = SSB / (k - 1)$	$f = MSB / MSE$ or, $F = MST/MSE$
Error	$SSE = \Sigma nj(\bar{X} - \bar{X}_j)^2$	$df_2 = N - k$	$MSE = SSE / (N - k)$	
Total	$SST = SSB + SSE$	$df_3 = N - 1$		

Applying the value of age to the regression Model to calculate the estimated blood pressure ($\hat{Y}$) coefficient of determination (R^2) as follows:

Estimated blood pressure ($\hat{Y}$) = 85.026 + 1.140 age

Obs	Age	BP	Est.	Est-Mean		Actual – Est		Actual - Mean	
	x	y	$\hat{y}$	$\hat{Y} - \bar{y}$	$(\hat{Y} - \bar{y})^2$	$(y - \hat{Y})$	$(y - \hat{Y})^2$	$(y - \bar{y})$	$(y - \bar{y})^2$
1	35	112	124.926	-16.074	258.373	-12.926	167.081	-29	841
2	40	128	130.626	-10.374	107.620	-2.626	6.896	-13	169
3	38	130	128.346	-12.654	160.124	1.654	2.736	-11	121
4	44	138	135.186	-5.814	33.803	2.814	7.919	-3	9
5	67	158	161.406	20.406	416.405	-3.406	11.601	17	289
6	64	162	157.986	16.986	288.524	4.014	16.112	21	441
7	59	140	152.286	11.286	127.374	-12.286	150.946	-1	1
8	69	175	163.686	22.686	514.655	11.314	128.007	34	1156
9	25	125	113.526	-27.474	754.821	11.474	131.653	-16	256
10	50	142	142.026	1.026	1.053	-0.026	0.001	1	1
Total	491	1410	1410	0.000	$\boxed{2662.750}$	0.000	$\boxed{622.950}$	0	$\boxed{3284}$

We can say that 81% of the variation in the BP rate is explained by age

Calculating the coefficient of determination (R^2)

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\text{Regression Sum of Square (SSR)}}{\text{Total Sum of Square (SST)}}$$

$$R^2 = \frac{2662.75}{3284} = 0.810$$

ANOVA Table

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma nj(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	MSB = SSB / (k - 1)	f = MSB / MSE or, F = MST/MSE
Error	$SSE = \Sigma nj(\bar{X} - \bar{X}_j)^2$	$df_2 = N - k$	MSE = SSE / (N - k)	
Total	SST = SSB + SSE	$df_3 = N - 1$		

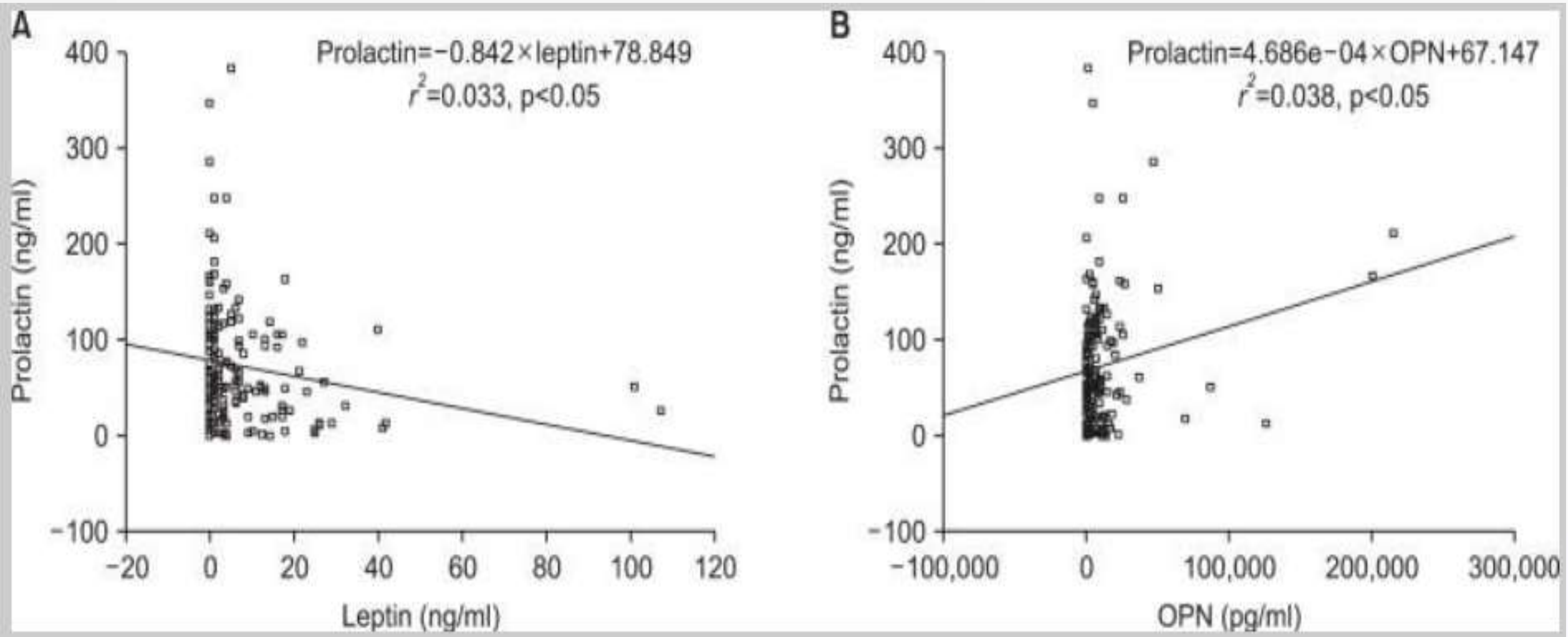
			F = MST / MSE
2662.75	1	2662.75	34.195
622.95	8	77.86875	
3284	9		

Correlation between preoperative serum levels of five biomarkers and relationships between these biomarkers and cancer stage in epithelial overian cancer

56 newly diagnosed epithelial ovarian cancer patients. Preoperative serum levels of leptin, prolactin, osteopontin (OPN), insulin-like growth factor-II, and CA-125 were determined by ELISA

Correlation between the five biomarkers was assessed using Pearson's correlation analysis.

There was a significant negative correlation between prolactin and leptin and a significant positive correlation between prolactin and OPN. No significant correlation was found between any of the other biomarkers



	Leptin	Prolactin	OPN	IGF-II	CA-125
Leptin					
CC	1	−0.182*	−0.022	0.102	−0.125
Significance	0.021	0.779	0.197	0.116	
Prolactin					
CC	−0.182*	1	0.195*	0.061	0.133
Significance	0.021		0.014	0.442	0.093
OPN					
CC	−0.022	0.195*	1	−0.020	0.072
Significance	0.779	0.014		0.798	0.363
IGF-II					
CC	0.102	0.061	−0.020	1	0.109
Significance	0.197	0.442	0.798		0.217
CA-125					
CC	−0.125	0.133	0.072	0.028	1
Significance	0.116	0.093	0.363	0.722	

Staging with computed tomography of patients with colon cancer

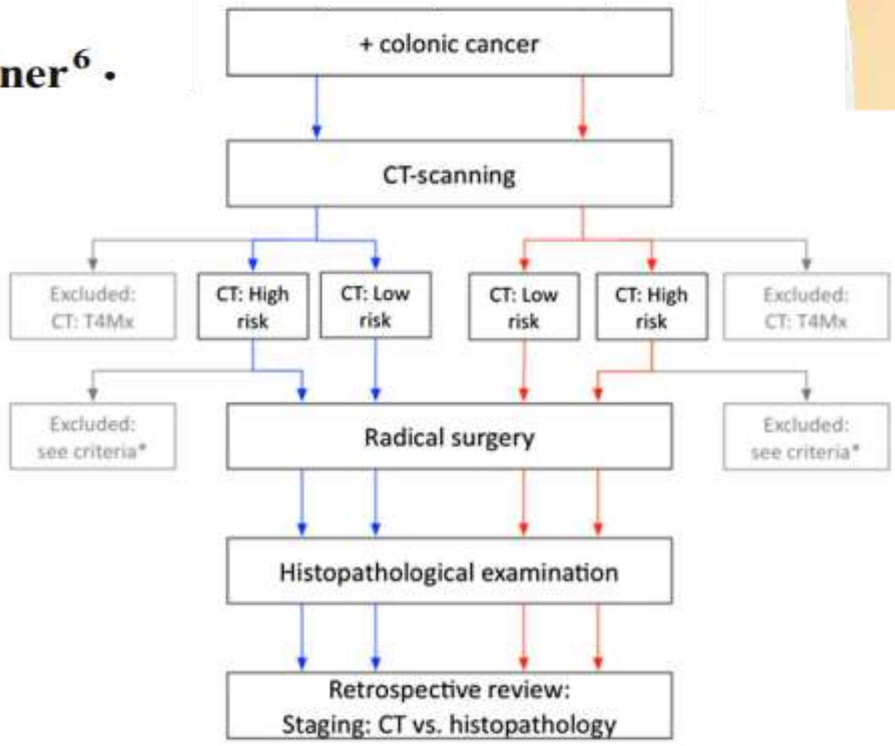
M. L. Malmstrøm[1,2] · S. Brisling[3] · T. W. Klausen[4] · A. Săftoiu[5] · T. Perner[6] ·
P. Vilmann[1] · I. Gögenur[2]

615 consecutive patients operated for colonic cancer. (Screened & Non screened Pts)
Patients were stratified into high-risk and low-risk groups based on T stage.

The Kendall tau correlation coefficient was used to calculate concordance between radiological (r)T-stage obtained at CT & pathological (p)T-stage from the final pathology

No significant differences in the Kendall tau values for diagnostic measures between the groups at the 95% (CI) level: 49% (95% CI, 43–55) for all individuals, 48% (95% CI, 40–56) for screened individuals, & 47% (95% CI, 37-56) for non-screened individuals.

CT-based T-staging showed no differences between the screened and symptomatic pts.



The individual radiologists (A–E) had staged > 20 individuals from the included population. These radiologists were experienced in colon cancer staging as follows: A and E > 2 years and B, C, and D > 4 years. The "other" group consisted of 31 different radiologists who all staged < 20 individuals from the included population; further, these radiologists were not necessarily experienced in cancer staging. Diagnostic measures are given for *finding a high-risk tumor*.

	A	B	C	D	E	Other
Sensitivity, 95% CI	0.556 [0.212; 0.863]	0.636 [0.308; 0.891]	0.576 [0.392; 0.745]	0.733 [0.449; 0.922]	0.783 [0.563; 0.925]	0.611 [0.435; 0.769]
Specificity, 95% CI	0.957 [0.852; 0.995]	0.923 [0.749; 0.991]	0.937 [0.880; 0.972]	0.781 [0.660; 0.875]	0.830 [0.702; 0.919]	0.862 [0.746; 0.939]
Kendall tau, 95% CI	0.512 [0.289; 0.666]	0.554 [0.266; 0.725]	0.463 [0.339; 0.571]	0.405 [0.237; 0.552]	0.575 [0.402; 0.692]	0.504 [0.358; 0.619]

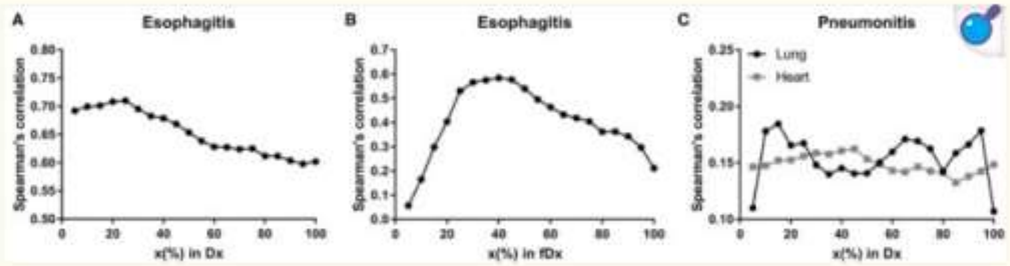
▸ Front Oncol. 2020 Aug 6;10:1395. doi: 10.3389/fonc.2020.01395

Predictive Modeling of Thoracic Radiotherapy Toxicity and the Potential Role of Serum Alpha-2-Macroglobulin

Baseline A2M levels were obtained for 258 patients prior to thoracic radiotherapy (RT).
Dose-volume characteristics were extracted from treatment plans.
Spearman's correlation (Rs) test was used to correlate clinical and dosimetric variables with toxicities.

Spearman's correlation test between dosimetric variables in esophagus and esophagitis showed that all variables had Rs > 0.60 ($p < 0.0001$) as shown in Figure 1A. For the fractional dose, fD40 was the highest correlated variable (Rs = 0.58/$p < 0.0001$) as shown in Figure 1B.

Figure 1.



Open in a new tab

Spearman's correlation coefficients. Spearman's correlation coefficients between radiation-induced injuries (≥ grade 2) and Dx in esophagus for **(A)** esophagitis, fDx in esophagus for **(B)** esophagitis, and Dx in lung and heart for **(C)** pneumonitis.

Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest

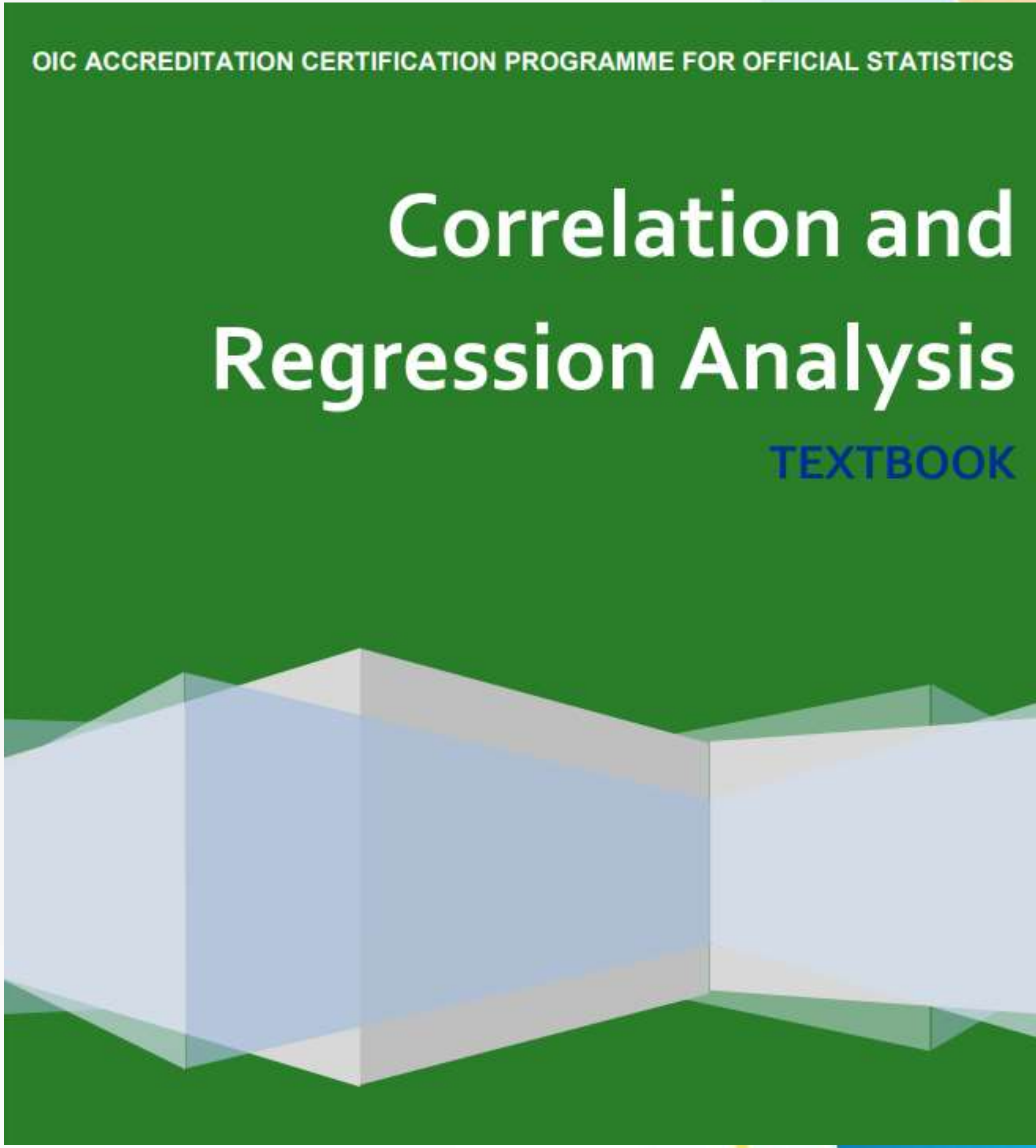
The attribute values ranges from
Clump Thickness 1 - 10
Uniformity of Cell Size 1 - 10
Uniformity of Cell Shape 1 – 10
Marginal Adhesion 1 - 10
Single Epithelial Cell Size 1 - 10
Bare Nuclei 1 - 10
Bland Chromatin 1 - 10
Normal Nucleoli 1 - 10
Mitoses 1 – 10

Class = Benign Vs Malignant

```
Call: lm(formula = class ~ clump_thickness +
shape_uniformity + size_uniformity +
marginal_adhesion + epithelial_size +
bare_nucleoli + bland_chromatin +
normal_nucleoli, data = wbcd)
Residuals:
     Min       1Q   Median       3Q      Max
-1.67976 -0.16600 -0.02453  0.11442  1.52764
Coefficients:  Estimate Std. Error t value
Pr(>|t|)
(Intercept)       1.505412   0.032613  46.160
< 2e-16 ***
clump_thickness   0.063518   0.007108   8.936
< 2e-16 ***
shape_uniformity  0.031286   0.012464   2.510
0.012300 *
size_uniformity   0.043806   0.012723   3.443
0.000611 ***
marginal_adhesion 0.016693   0.007910   2.110
0.035194 *
epithelial_size   0.020559   0.010261   2.004
0.045509 *
bare_nucleoli     0.090711   0.006429  14.109
< 2e-16 ***
bland_chromatin   0.038179   0.010043   3.801
0.000157 ***
normal_nucleoli   0.037237   0.007379   5.046
5.8e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3801 on 674 degrees
of freedom
Multiple R-squared:  0.8433, Adjusted R-
squared:  0.8415
F-statistic: 453.5 on 8 and 674 DF,  p-value:
< 2.2e-16
```

The success rate of classification is 84.33% obtained by linear regression.

Thank You


OIC ACCREDITATION CERTIFICATION PROGRAMME FOR OFFICIAL STATISTICS
Correlation and
Regression Analysis
TEXTBOOK