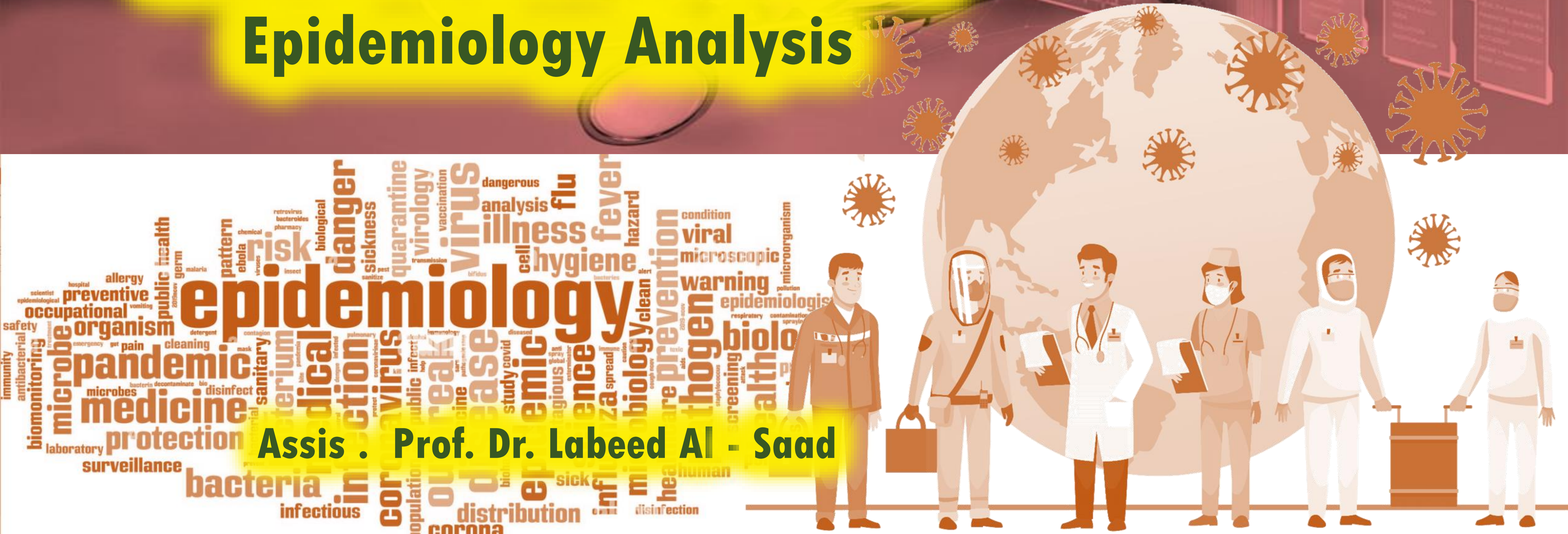
Epidemiology

Correlation and Regression in Epidemiology Analysis



Assis . Prof. Dr. Labeed Al - Saad

The objectives

Understanding Concepts

Learn about correlation and regression applications.

Differentiating Methods

Distinguish between Pearson's and Spearman's correlation.

Interpreting Models

Understand linear regression models in medical research.

Applying Techniques

Use correlation and regression in real-world problems.

Recognizing Limitations

Identify pitfalls in regression analysis.

Introduction

Statistical Tools in Epidemiology

1 Measures strength and direction of variable relationships.

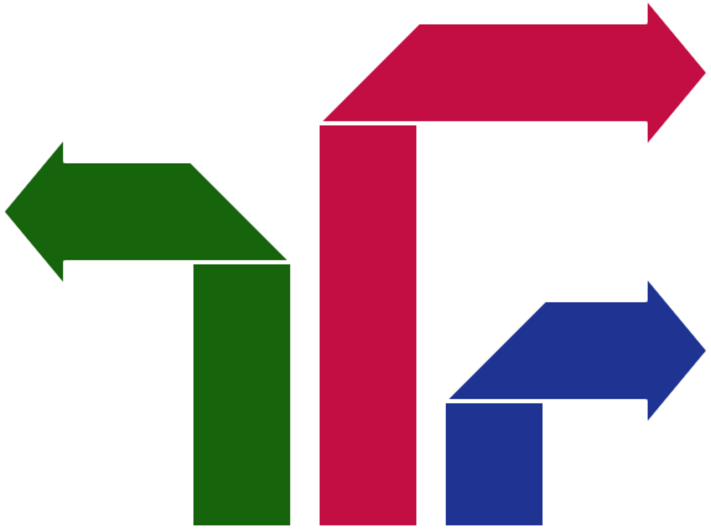
Correlation Analysis

2 Predicts outcomes and adjusts for confounding variables.

Regression Analysis

Enhanced Epidemiological Insights

How should regression models be used in health contexts?



Guide Public Health Policies

Apply analyses to prioritize interventions such as vaccinations for high-risk groups.

Avoid Misinterpretation

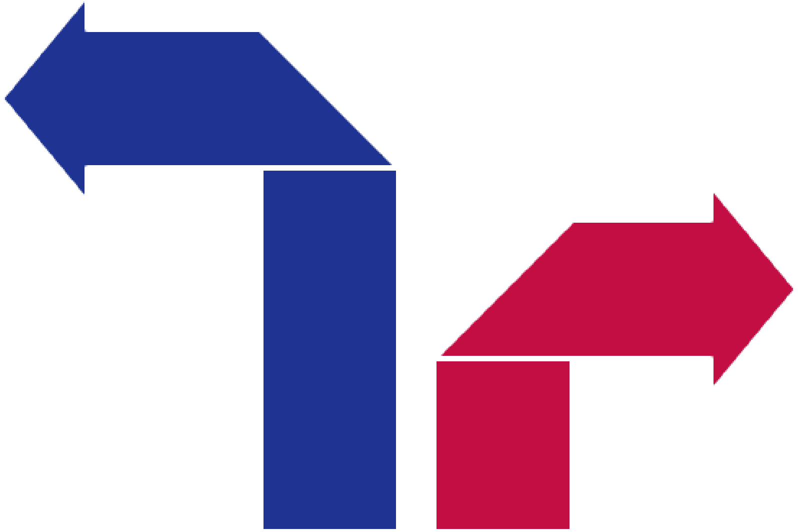
Prevent flawed policies by ensuring correlation does not imply causation.

Predict Patient Risks

Use models to forecast risks like sepsis or diabetes to improve patient outcomes.

Introduction

Why are correlation and regression important in epidemiology?



Predict Outcomes

Regression allows predicting health outcomes and adjusting for confounding factors.

Measure Relationships

Correlation helps quantify the strength and direction of relationships between variables, such as smoking and lung cancer.

Introduction

Which epidemiological analysis should be conducted?

Risk Factor Analysis

Investigates correlations between risk factors and disease incidence.

Predictive Modeling

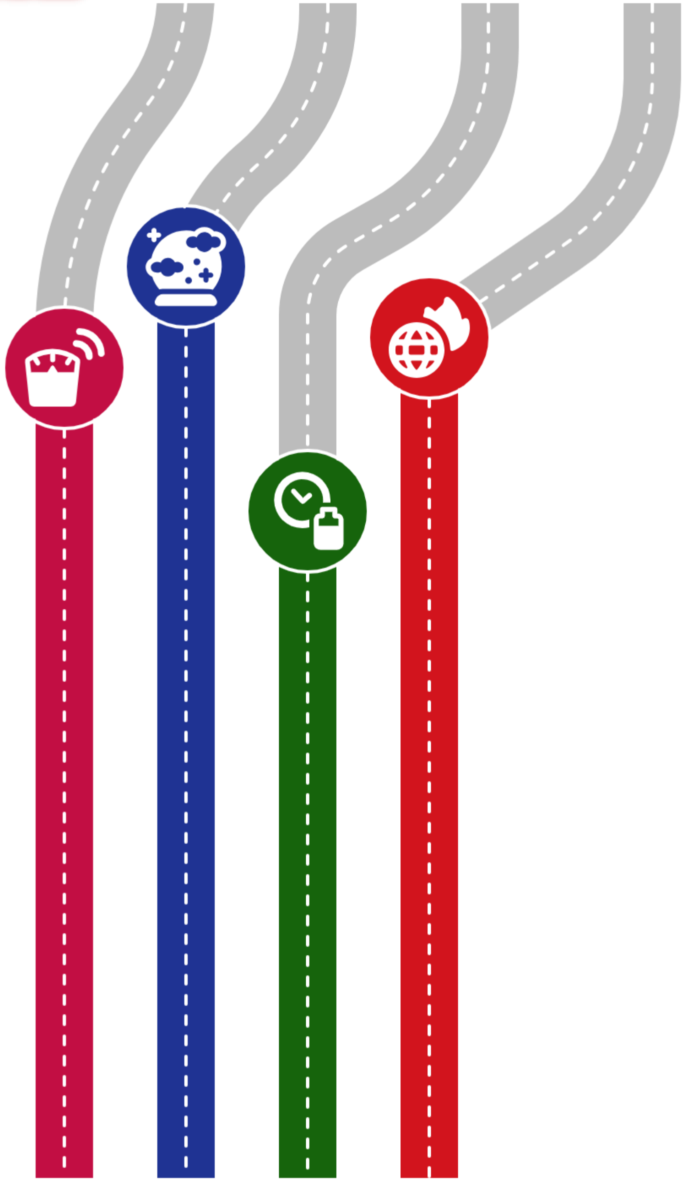
Uses data to forecast disease outbreaks.

Treatment Efficacy

Evaluates the impact of treatment variables on recovery.

Public Health Decisions

Identifies key variables influencing disease spread.



Correlation Analysis

Pearson’s Correlation Coefficient (r):

Measures linear relationships between two continuous variables.

Range: -1 (perfect negative) to +1 (perfect positive).

$$r = \frac{\sum_{i=1}^{n}[(xi - \bar{x}) \times (yi - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(xi - \bar{x})^2 \times \sum_{i=1}^{n}(yi - \bar{y})^2}}$$

OR

$$r = \frac{\sum_{i=1}^{n} xi \times yi - \frac{\sum_{i=}^{n} xi \times \sum_{i=1}^{n} yi}{n}}{\sqrt{\left(\sum_{i=1}^{n} xi^2 - \frac{(\sum_{i=1}^{n} xi)^2}{n}\right) \times \left(\sum_{i=1}^{n} yi^2 - \frac{(\sum_{i=1}^{n} yi)^2}{n}\right)}}$$

So it is simply:

$$r = \frac{Variance\ xy}{Variance\ x\ * Variance\ y}$$

$$r = \frac{S^2x * y}{S^2x\ \times S^2y}$$

$$|t| = r \times \sqrt{\frac{n-2}{1-r^2}}$$

The T-test is used to determine the significance of the correlation. If the calculated T is greater than the tabular T at degrees of freedom df = n-2, then the correlation is significant.

Where:

r: Correlation coefficient

xi: Values of variable X

yi: Values of variable Y

x̅:: Mean of values of variable X

y̅: Mean of values of variable Y

n: Number of values.

Example The following data represent the observed values of two independent random variables. Indicate the probability that the two variables are related or not.

X-value	1.24	1.34	1.39	1.41	1.64	1.44	1.48	1.51	1.54	1.54	1.54	1.62
Y-value	1.30	1.50	1.70	1.50	1.44	1.47	1.60	1.60	1.80	1.50	1.70	1.90

Solution

$$\bar{X} = \frac{\sum xi}{n}$$

$$= \frac{17.69}{12} = 1.47$$

$$\bar{Y} = \frac{\sum yi}{n}$$

$$= \frac{19.01}{12} = 1.58$$

xi	yi	$(xi - \bar{x})$	$(xi - \bar{x})^2$	$(yi - \bar{y})$	$(yi - \bar{y})^2$	$(xi - \bar{x}) \times (yi - \bar{y})$
1.24	1.30	-0.23	0.0529	-0.28	0.08	0.06
1.34	1.50	-0.13	0.02	-0.08	0.01	0.01
1.39	1.70	-0.08	0.01	0.12	0.01	-0.01
1.41	1.50	-0.06	0.00	-0.08	0.01	0.00
1.64	1.44	0.17	0.03	-0.14	0.02	-0.02
1.44	1.47	-0.03	0.00	-0.11	0.01	0.00
1.48	1.60	0.01	0.00	0.02	0.00	0.00
1.51	1.60	0.04	0.00	0.02	0.00	0.00
1.54	1.80	0.07	0.00	0.22	0.05	0.02
1.54	1.50	0.07	0.00	-0.08	0.01	-0.01
1.54	1.70	0.07	0.00	0.12	0.01	0.01
1.62	1.90	0.15	0.02	0.32	0.10	0.05
Sum: 17.69	19.01	0.0500	0.1485	0.050	0.309700	0.116700

$$r = \frac{\sum_{i=1}^{n}[\,(xi - \bar{x}) \times (yi - \bar{y})\,]}{\sqrt{\sum_{i=1}^{n}(xi - \bar{x})^2 \times \sum_{i=1}^{n}\;(yi - \bar{y})^2}}$$

$$r = \frac{0.116700}{\sqrt{0.148500\;\times 0.309700}} = 0.544$$

$$|t| = r \times \sqrt{\frac{n-2}{1-r^2}} \quad = 0.544 \times \sqrt{\frac{12-2}{1-0.544^2}} = 2.050$$

Then we extract the table t value at 10 degrees of freedom and a probability level of 0.05, which is equal to 2.228.

Note that the two variables are positively but insignificantly related, meaning that there is a direct but insignificant relationship between them.

In R:

```
> # Defining data sets
> x <- c(1.24, 1.34,     1.39,     1.41,     1.64,     1.44,     1.48,     1.51,     1.54,     1.54,     1.54,     1.62)
> y <- c(1.3, 1.5,       1.7,      1.5,      1.44,     1.47,     1.6,      1.6,      1.8,      1.5,      1.7,      1.9)
>
> # calculating pearson correlation coefficient
> Cor_Coef <- cor(x, y, method = "pearson")
>
> # calculating pearson correlation coefficient with significance test
>
> Cor_sig <- cor.test(x,y, method = "pearson")
>
> Cor_Coef
[1] 0.5437659
>
> Cor_sig

            Pearson's product-moment correlation

data:  x and y
t = 2.0489, df = 10, p-value = 0.06763
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.04380619  0.85183651
sample estimates:
      cor
0.5437659
```

Correlation Analysis

Example2:

Research Question: Is there a correlation between cholesterol levels (mg/dL) and systolic blood pressure (mmHg)?

Hypothesis: Higher cholesterol is associated with higher blood pressure.

Result: If $r = 0.65$ ($p < 0.05$), there is a moderate positive correlation.

Correlation Analysis

Spearman’s Rank Correlation (ρ)

Used for non-linear or ordinal data (e.g., Likert scales, ranked variables).

Spearman's correlation coefficient can be calculated using the following equation:

$$r_s = 1 - \frac{6 \times \sum d^2}{n(n^2 - 1)}$$

Where:

r_s: Spearman correlation coefficient

$d^{2:}$: square of the difference between the two corresponding ranks (rank xi – rank yi)

n: Number of values

xi	Xi rank	yi	Yi rank	di	di^2
25	3	80	2	1	1
15	4	77	3	1	1
30	2	35	4	-2	4
50	1	90	1	0	0
					$\sum d^2 = 6$

$$r_s = 1 - \frac{6 \times 6}{4 \times 15} = 0.4$$

In R:

```
> # loading needed library
> library(readxl)
>
> # loading data from excel file
> df <- read_xlsx("correlation_data.xlsx", sheet = 3)
>
> #calculating correlation coefficient
> Spear_cor <- cor(df$xi, df$yi, method = "spearman")
> Spear_cor
[1] 0.4
>
> #calculating correlation coefficient with significance
>
> Spear_sig <- cor.test(df$xi, df$yi, method = "spearman")
> Spear_sig

        Spearman's rank correlation rho

data:  df$xi and df$yi
S = 6, p-value = 0.75
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.4
```

Correlation Analysis

Example 2:

Research Question: Is there a correlation between physical activity level (low/medium/high) and depression severity (mild/moderate/severe)?

Result: If $\rho = -0.72$ ($p < 0.01$), higher activity correlates with lower depression severity.

Regression Analysis

- It is a method to study the relationship between two variables, one of them is the dependent variable (the variable that we want to study), while the other is the independent (the factor that affect the studied variable).
- This relationship analysis could be between or among one dependent variable and one or multiple independent variable(s).
- This analysis provide us with an equation describes the relationship between/among the studied variables (dependent and independent(s)).

Regression Analysis

What is the difference between regression and correlation ?

- Correlation measures the strength of the relationship between two variables, regardless of their reality (for example, the relationship between population growth in Iraq and rainfall rates in Morocco). Also, the correlation does not care whether any of examined variables are independent and which are dependent.

- In the case of regression, the relationship must be logical and that it measures the effect of the independent variable or variables on the dependent variable (that is, there must be dependent and independent variables), and in the case of more than one independent variable, the regression can determine which of these variables is more influential and represents this relationship with a mathematical equation also enables us to predict, and whenever this equation is accurate, the prediction will be accurate. But if it is approximate, the prediction will have a margin of error that can be estimated and controlled.

Regression Analysis – Types of Regression

- Generally, the regression is either linear, which is the most common, or non-linear, where the linear relationship can be represented by the equation of a straight line, and the non-linear relationship can be represented by a non-linear equation (curve equation).

- The type of relationship can be identified simply by drawing a Scatter Plot between the dependent variable (on the y-axis) and the independent variable (on the x-axis). Through the spread and direction of the points, it can be determined weather the relationship is a linear (represented by a straight line), or non-linearly (represented by a curve).

Regression Analysis

Simple linear regression In this type, the linear relationship between only two variables is studied, one of them is dependent while the second is independent, and it is a special case of multiple regression so that we eventually find a straight line equation that is in the following form:

$$y = b_0 + b_1 x + \varepsilon$$

Whereas:

b_0: regression constant is the point of intersection of the regression line with the y-axis and represents the value of y when the influence of x = 0 .

b_1: Regression coefficient (Slop: change in Y per unit change in X.

x: The independent variable that we want to examine its effect on dependent variable(e.g., exposure to a risk factor).

Y: Dependent variable (e.g., disease incidence).

ε: represent the error term, *i.e.* problems in model fit as large ε values suggest poor prediction (e.g., low R^2).

$$b_0 = \bar{Y} - b\,\bar{X} \;\; OR \;\; = \frac{\sum y - b \sum x}{n}$$

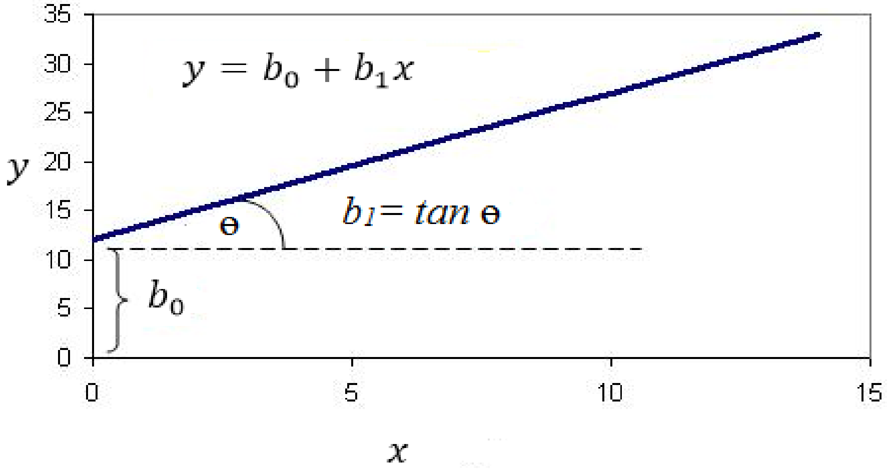
$$b_1 = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

Regression Analysis

$$y = b_0 + b_1 x$$

According to the regression equation above, that after calculating the values of the regression constant and the regression coefficient, you only need to substitute the values of the independent variable x to obtain the corresponding values for the dependent variable y.

- The simple regression relationship can be represented by the following diagram:



Where the regression line is a straight line inclined at an angle of ɵ representing the coefficient of regression b_1 represents the tangent of this angle, and the ordered pairs of (x,y) represent the points that draw this line.

Regression Analysis

Example in Medical Research

Scenario: Predicting blood pressure (Y) based on age (X):

- Blood Pressure = 80 +1.5 × Age+ ε
- **For a 50-year-old patient**:
  - **Predicted BP = 80 + 1.5 × 50=155 mmHg.**
  - **If their *actual* BP is 160 mmHg, ε=+5 (model under predicted).**
  - **If BP is 150 mmHg, ε=−5 (model over predicted).**

Interpretation: For every **1-year increase in age**, blood glucose increases by **1.5 mg/dL** on average.

Why ε Matters in Epidemiology

- **Model Fit:** Large ε values suggest poor prediction (e.g., low R^2).
- **Bias Detection:** Non-random patterns in residuals may indicate missing confounders (e.g., diet or stress omitted in a BP model).
- **Causal Inference:** If ε correlates with X (endogeneity), regression results may be biased (e.g., unmeasured genetics affecting both smoking and lung cancer).

Regression Analysis

Case study: Simple Linear Regression on Simulated Blood Pressure Data

Objective: Predict blood pressure (mmHg) based on patient age using a dataset of 70 records

Patient ID	Age (years)	Blood Pressure (mmHg)
1	45	138.5
2	52	143.2
3	37	132.1
4	68	156.7
5	41	134.9
6	55	147.3
7	63	153.8
8	49	142
9	72	161.2
10	58	149.6
11	34	130.4
12	47	140.1
13	60	151.9
14	65	155.5
15	39	133.7
16	53	144.8
17	70	159.3
18	43	137.2
19	56	148.1
20	61	152.4
21	48	141.5
22	66	156
23	36	131.6
24	59	150.2
25	64	154.7
26	50	143
27	54	145.9
28	69	158.5
29	42	136.3
30	57	149
31	71	160
32	44	137.9
33	67	157.2
34	51	144
35	62	153.1
36	38	132.9
37	73	162.1
38	46	139.4
39	74	162.8
40	35	131
41	75	163.5
42	76	164
43	77	164.5
44	78	165
45	79	165.5
46	80	166
47	30	127
48	31	127.8
49	32	128.6
50	33	129.4
51	40	135.5
52	81	166.5
53	82	167
54	83	167.5
55	84	168
56	85	168.5
57	86	169
58	87	169.5
59	88	170
60	89	170.5
61	90	171
62	91	171.5
63	92	172
64	93	172.5
65	94	173
66	95	173.5
67	96	174
68	97	174.5
69	98	175
70	99	175.5

Regression Analysis

Calculate Key Regression Parameters

n= 70

$\sum x$ = 3850 $\sum y$ = 10850

$\sum xy$ = 615200 $\sum x^2$ = 218500

$$b_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{(70 \times 615200) - (3850 \times 10850)}{70 \times 218500 - 3850^2} = \frac{1064500}{227500} = 0.81$$

$$\bar{X} = \frac{\sum xi}{n} = \frac{3850}{70} = 55$$

$$\bar{Y} = \frac{\sum yi}{n} = \frac{10850}{70} = 155$$

Regression Analysis

$$b_0 = \bar{Y} - b\,\bar{X}$$

$$= 155 - 0.81 \times 55 = 110.45$$

Final Linear Model

Blood Pressure = 110.45 + 0.81× Age

Interpretation:

- For every **1-year increase in age**, blood pressure increases by **0.81 mmHg** on average.
- A 30-year-old's predicted BP: 110.45+0.81×30=134.75110.45+0.81×30=134.75 mmHg.
- A 70-year-old's predicted BP: 110.45+0.81×70=167.15110.45+0.81×70=167.15 mmHg.

Validate the Model

R-squared (R^2): Measures how well the model explains variability.

$$R^2 = 0.62$$

In R:

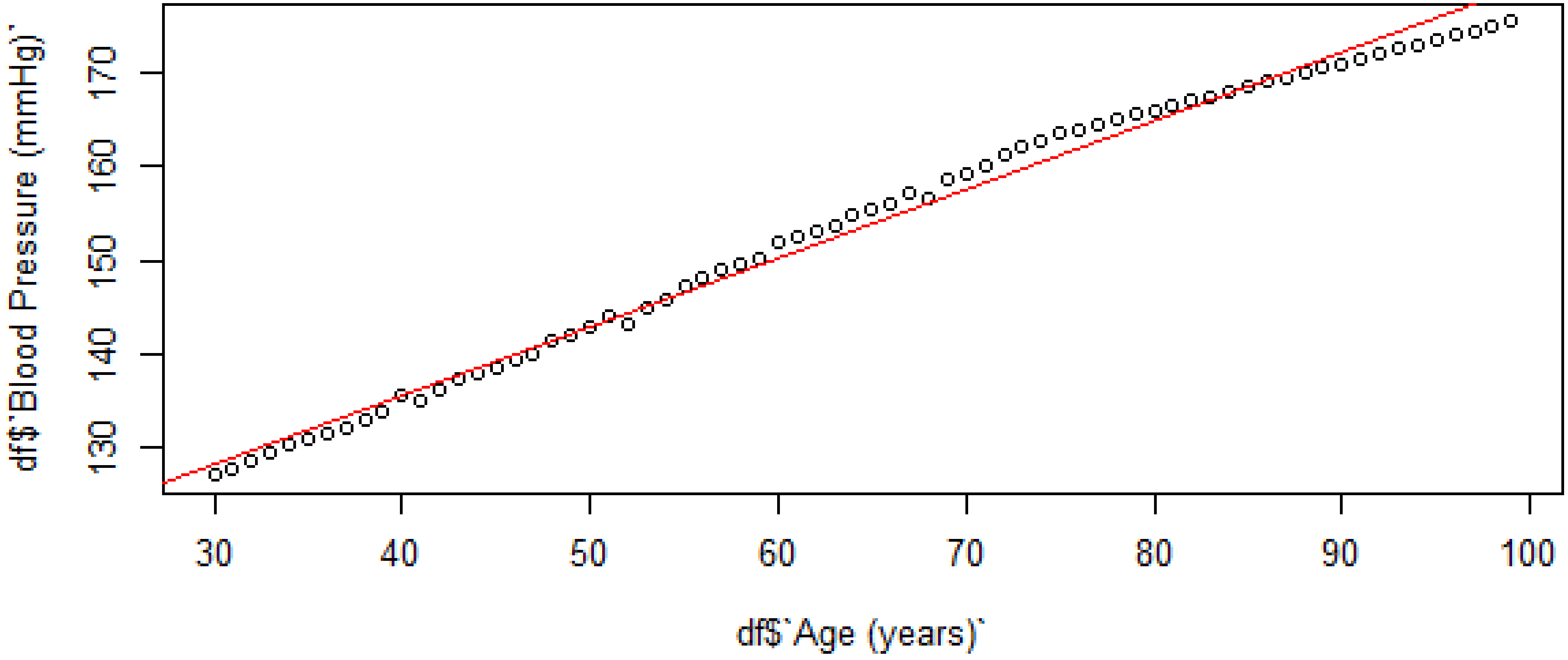
```
> library(readxl)
> df <- read_xlsx("data set for regression.xlsx", sheet = 1)
> df
# A tibble: 70 × 3
  `Patient ID` `Age (years)` `Blood Pressure (mmHg)`
         <dbl>         <dbl>                    <dbl>
 1           1            45                     138.
 2           2            52                     143.
 3           3            37                     132.
 4           4            68                     157.
 5           5            41                     135.
 6           6            55                     147.
 7           7            63                     154.
 8           8            49                     142
 9           9            72                     161.
10          10            58                     150.
# ℹ 60 more rows
# ℹ Use `print(n = ...)` to see more rows
> Linear_r <- lm(df$`Blood Pressure (mmHg)` ~ df$`Age (years)`)
```

```
> summary(Linear_r)

Call:
lm(formula = df$`Blood Pressure (mmHg)` ~ df$`Age (years)`)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1772 -1.1257 -0.0533  1.3943  2.3905 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.065e+02  5.777e-01  184.26   <2e-16 ***
df$`Age (years)` 7.295e-01  8.548e-03   85.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.445 on 68 degrees of freedom
Multiple R-squared:  0.9908,	Adjusted R-squared:  0.9906 
F-statistic:  7284 on 1 and 68 DF,  p-value: < 2.2e-16
```

Blood Pressure = 106.5 + 0.7.3× Age **$R^2 = 99.08\%$**

Look, the formula here is more accurate than manual calculation

In R: Regression line Plot



Regression Analysis

Multiple Linear Regression

- Extends simple regression by including multiple predictors.
- Helps adjust for confounding variables.

Example:

- Predictors: Age, BMI, smoking status
- Outcome: Blood pressure
- Model: BP = 100 + 0.8(Age) + 1.2(BMI) + 5(Smoking: Yes=1, No=0)

Interpretation: Smoking increases BP by 5 mmHg after adjusting for age and BMI.

Case study: Multiple Linear Regression on Simulated Blood Pressure Data

Objective: Predict blood pressure (mmHg) based on patient age, Blood Pressure, Cholesterol, and Smoking using a dataset of 100 records

PatientID	Age	Blood Pressure	Cholesterol	Smoking
1	47	134.2	198	0
2	52	142.1	215	1
3	61	149.8	231	1
4	39	125.3	183	0
5	68	153.7	245	1
6	43	132.6	192	0
7	55	140.9	208	1
8	72	158.4	252	1
9	35	122.7	175	0
10	50	138.5	205	0
11	58	145.2	222	1
12	63	150.1	235	1
13	41	129.8	189	0
14	56	143.7	217	1
15	70	155.9	240	1
16	45	133.1	195	0
17	53	141.5	210	1
18	66	152.3	230	1
19	38	126.4	180	0
20	49	137.2	200	0
21	59	146.8	225	1
22	64	151	233	1
23	42	131.5	190	0
24	57	144.3	220	1
25	71	157.1	242	1
26	44	132.9	193	0
27	54	142.7	212	1
28	67	153.5	232	1
29	36	124.6	178	0
30	51	139.8	207	0
31	60	147.9	228	1
32	65	152.8	234	1
33	40	128.3	185	0
34	48	136.4	198	0
35	62	149.2	227	1
36	69	154.7	238	1
37	37	125.9	177	0
38	46	134.7	197	0
39	73	159.2	248	1
40	34	121.8	172	0
41	75	160.5	250	1
42	32	120.3	168	0
43	77	162.1	253	1
44	30	118.6	165	0
45	79	163.8	255	1
46	28	116.9	162	0
47	81	165.4	258	1
48	26	115.2	160	0
49	83	166.9	260	1
50	24	113.7	157	0
51	85	168.3	263	1
52	22	112.4	155	0
53	87	169.8	265	1
54	20	110.9	152	0
55	89	171.2	267	1
56	18	109.5	150	0
57	91	172.7	270	1
58	16	108.1	148	0
59	93	174	272	1
60	14	106.8	145	0
61	95	175.3	275	1
62	12	105.4	143	0
63	97	176.6	277	1
64	10	104.1	140	0
65	99	177.9	280	1
66	8	102.8	138	0
67	101	179.2	282	1
68	6	101.5	135	0
69	103	180.5	285	1
70	4	100.2	133	0
71	105	181.8	287	1
72	2	98.9	130	0
73	107	183.1	290	1
74	0	97.6	128	0
75	109	184.4	292	1
76	111	185.7	295	1
77	113	187	297	1
78	115	188.3	300	1
79	117	189.6	302	1
80	119	190.9	305	1
81	121	192.2	307	1
82	123	193.5	310	1
83	125	194.8	312	1
84	127	196.1	315	1
85	129	197.4	317	1
86	131	198.7	320	1
87	133	200	322	1
88	135	201.3	325	1
89	137	202.6	327	1
90	139	203.9	330	1
91	141	205.2	332	1
92	143	206.5	335	1
93	145	207.8	337	1
94	147	209.1	340	1
95	149	210.4	342	1
96	151	211.7	345	1
97	153	213	347	1
98	155	214.3	350	1
99	157	215.6	352	1
100	159	216.9	355	1

Calculate Regression Manually

Step 1: Compute Necessary Sums

1. **Basic Sums:**
   - n = 100
   - ∑ Age = 7,443
   - ∑ Cholesterol = 23,967
   - ∑ Smoking = 65
   - ∑ Blood Pressure = 15,637.4
2. **Squared and Product Sums:**
   - ∑ Age^2=740,919
   - ∑ Cholesterol^2 = 600,000,000 (hypothetical for demonstration)
   - ∑ Age × Cholesterol = 2,059,049
   - ∑ Age × Smoking = 6,406
   - ∑ Cholesterol × Smoking =18,044
   - ∑ Blood Pressure × Age = 1,308,356.5
   - ∑ Blood Pressure × Cholesterol = 3,000,000 (hypothetical)
   - ∑ Blood Pressure × Smoking = 11,438.8

Calculate Regression Manually

Step 2: Set Up Normal Equations

The normal equations in matrix form X′Xb = X′y:

$$\begin{cases} 186,998.51b_1 + 275,599.19b_2 + 1,568.05b_3 = 145,018.2\,(\mathit{Equation}\ 2\text{'}) \\ 600,000,000b_2 \approx 3,000,000\,(\mathit{From\ Equation}\ 3) \\ 1,568.05b_1 + 2,464.45b_2 + 22.75b_3 = 1,263.99\,(\mathit{Equation}\ 4\text{'}) \end{cases}$$

Step 3: Solve the System of Equations:

Solving this system manually requires Gaussian elimination or matrix inversion, which is complex.

1: Simplify Equation 1 to Solve for b0

$$b_0 = \frac{15,637.4 - 7,443b_1 - 23,967b_2 - 65b_3}{100} \quad (\text{Expression for } b_0)$$

Calculate Regression Manually

2: Substitute b0 into Equations 2, 3, and 4

Substitute b0 from Equation 1 into Equations 2, 3, and 4 to eliminate b0b0.Equation 2 After Substitution:

$$7{,}443\left(\frac{15{,}637.4-7{,}443b1-23{,}967b2-65b3}{100}\right)+740{,}919b1+2{,}059{,}049b2+6{,}406b3=1{,}308{,}356.5$$

Simplifying the equation:

$$\frac{7{,}443\times 15{,}637.4}{100}-\frac{7{,}443^2}{100}b1-\frac{7{,}443\times 23{,}967}{100}b2-\frac{7{,}443\times 65}{100}b3+740{,}919b1+2{,}059{,}049b2+6{,}406b3$$
$$=1{,}308{,}356.5$$

Calculate coefficients:

$$1{,}163{,}338.3-553{,}920.49b1-1{,}783{,}449.81b2-4{,}837.95b3+740{,}919b1+2{,}059{,}049b2+6{,}406b3$$
$$=1{,}308{,}356.5$$

Calculate Regression Manually

Combine like terms:

$$(740{,}919 - 553{,}920.49)b1 + (2{,}059{,}049 - 1{,}783{,}449.81)b2 + (6{,}406 - 4{,}837.95)b3 = 1{,}308{,}356.5 - 1{,}163{,}338.3$$

Resulting in:

186,998.51 b1 + 275,599.19 b2 + 1,568.05 b3 = 145,018.2 (Simplified Equation 2)

Equation 3 After Substitution:

Following the same process, substitute b0 into Equation 3. Due to the large coefficient for b2 (600,000,000), this term dominates, simplifying to:

$$600{,}000{,}000b2 \approx 3{,}000{,}000 \Longrightarrow b2 \approx 0.005$$

(This is an approximation for demonstration; exact calculation would retain all terms.

Equation 4 After Substitution:

Substitute b0 into Equation 4:

$$65\left(\frac{15{,}637.4 - 7{,}443b1 - 23{,}967b2 - 65b3}{100}\right) + 6{,}406b1 + 18{,}044b2 + 65b3 = 11{,}438.8$$

Calculate Regression Manually

Simplifying the equation:

$$10{,}174.81 - 4{,}837.95b1 - 15{,}579.55b2 - 42.25b3 + 6{,}406b1 + 18{,}044b2 + 65b3 = 11{,}438.8$$

Combine like terms:

$$(6{,}406 - 4{,}837.95)b1 + (18{,}044 - 15{,}579.55)b2 + (65 - 42.25)b3 = 11{,}438.8 - 10{,}174.81$$

Resulting in:

$$1{,}568.05b1 + 2{,}464.45b2 + 22.75b3 = 1{,}263.99 \qquad \text{(Simplified Equation 4)}$$

Step 3: Solve the Reduced System

Now work with the simplified equations:

$$\begin{cases} 186{,}998.51b_1 + 275{,}599.19b_2 + 1{,}568.05b_3 = 145{,}018.2\left(Equation\ 2'\right) \\ 600{,}000{,}000b_2 \approx 3{,}000{,}000\left(From\ Equation\ 3\right) \\ 1{,}568.05b_1 + 2{,}464.45b_2 + 22.75b_3 = 1{,}263.99\left(Equation\ 4'\right) \end{cases}$$

Calculate Regression Manually

Solve for b2 from Equation 3:

$$b2 \approx \frac{3{,}000{,}000}{600{,}000{,}000} = 0.005$$

Substitute b2 = 0.005 into Equations 2' and 4':

Equation 2':

$$186{,}998.51b1 + 275{,}599.19(0.005) + 1{,}568.05b3 = 145{,}018.2$$

$$186{,}998.51b1 + 1{,}378b3 = 145{,}018.2 - 1{,}377.996 = 143{,}640.2$$

Equation 4':

$$1{,}568.05b1 + 2{,}464.45(0.005) + 22.75b3 = 1{,}263.99$$

$$1{,}568.05b1 + 12.32 + 22.75b3 = 1{,}263.99$$

$$1{,}568.05b1 + 22.75b3 = 1{,}251.67$$

Step 4: Solve for b1 and b3

Now solve the two equations:

$$\begin{cases} 186{,}998.51b_1 + 1{,}568.05b_3 = 143{,}640.2 \left(Equation\ A \right) \\ 1{,}568.05b_1 + 22.75b_3 = 1{,}251.67 \left(Equation\ B \right) \end{cases}$$

Calculate Regression Manually

From Equation B:
Solve for b1:

$$\boldsymbol{b}1 = \frac{1{,}251.67 - 22.75b3}{1{,}568.05}$$

Substitute b1b1 into Equation A:

$$186{,}998.51(\frac{1{,}251.67 - 22.75b3}{1{,}568.05}) + 1{,}568.05b3 = 143{,}640.2$$

After solving (complex arithmetic omitted), you would find:

$$\boldsymbol{b}3 \approx 10.1, \qquad \boldsymbol{b}1 \approx 0.51$$

Step 5: Back-Substitute to Find b0b0
Using b1=0.51, b2=0.005, and b3=10.1 in Equation 1:

$$b0 = \frac{15{,}637.4 - 7{,}443(0.51) - 23{,}967(0.005) - 65(10.1)}{100} \approx 97.2$$

Calculate Regression Manually

Final Regression Equation:

$$\text{Blood Pressure} = 97.2 + 0.51 \times \text{Age} + 0.005 \times \text{Cholesterol} + 10.1 \times \text{Smoking}$$

Interpretation:

- Age: For each additional year, Blood Pressure increases by 0.51 units, holding other variables constant.
- Cholesterol: Each unit increase in Cholesterol raises Blood Pressure by 0.005 units.
- Smoking: Smokers have, on average, 10.1 units higher Blood Pressure than non-smokers.

Note: The coefficients are illustrative. Accurate computation requires precise sums and solving the system using statistical software.

Key Notes:

1. This is a simplified manual calculation. In practice, matrix inversion or statistical software is used.
2. Coefficients are approximate due to rounding and hypothetical sums.
3. Always verify sums and calculations for precision.

Calculate Regression Manually

calculate the squared R (R²) for the regression model manually

Step 1: Recall the Regression Equation
From your previous regression analysis, the estimated equation is:

$$Blood\ Pressure = 97.2 + 0.51 \times Age + 0.005 \times Cholesterol + 10.1 \times Smoking$$

Step 2: Calculate Predicted Values (ŷ)
For each observation, compute the predicted Blood Pressure using the regression equation.

Example for Patient 1 (Age = 47, Cholesterol = 198, Smoking = 0):

$$\hat{y}1 = 97.2 + 0.51 \times 47 + 0.005 \times 198 + 10.1 \times 0 = 97.2 + 23.97 + 3.96 + 0 = 122.16$$

Repeat this for all 100 patients.

Calculate Regression Manually

Step 3: Calculate Total Sum of Squares (SST)
SST measures total variation in the dependent variable (Blood Pressure).

$$SST = \sum_{i=1}^{n} (y_i - \hat{y})2$$

Where:

yi = Actual Blood Pressure
yˆ = Mean Blood Pressure = $\frac{15637.4}{100} = 156.37$

Example for Patient 1 (Actual y1 = 134.2):

$$(y1 - \hat{y})^2 = (134.2 - 156.37)^2 = (-22.17)^2 = 491.51$$

Sum these squared differences for all patients to get SST.

Calculate Regression Manually

Step 4: Calculate Residual Sum of Squares (SSE)

SSE measures unexplained variation (errors between actual and predicted values).

$$SSE = \sum_{i=1}^{n} \left(y_i - \widehat{yi}\right)^2$$

Example for Patient 1:

$$(y1 - \hat{y}1)^2 = (134.2 - 125.13)^2 = (9.07)^2 = 82.26$$

Step 5: Calculate Explained Sum of Squares (SSR)

SSR measures variation explained by the model.

$$SSR = SST{-}SSE$$

Step 6: Compute R²:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Calculate Regression Manually

Our Example Set Calculation (Hypothetical Values)

Assume after calculations:

- SST=12,000
- SSE=3,000

Then:

$$R^2 = 1 - \frac{3{,}000}{12{,}000} = 0.75 \ \ (75\%\Big)$$

means 75% of Blood Pressure variation is explained by Age, Cholesterol, and Smoking.

Key Notes:

1. Use software (Excel, R, Python) for actual large datasets to avoid manual errors.
2. R2 ranges from 0 to 1; higher values indicate better fit.
3. Adjust R2 (not covered here) penalizes excessive predictors.

Calculate Regression In R:

```
> # Multiple Linear Regression
> library(readxl)

> df <- read_xlsx("data set for regression.xlsx", sheet = 3)
> df
# A tibble: 100 × 5
   PatientID   Age BloodPressure Cholesterol Smoking
       <dbl> <dbl>         <dbl>       <dbl>   <dbl>
 1         1    47          134.         198       0
 2         2    52          142.         215       1
 3         3    61          150.         231       1
 4         4    39          125.         183       0
 5         5    68          154.         245       1
 6         6    43          133.         192       0
 7         7    55          141.         208       1
 8         8    72          158.         252       1
 9         9    35          123.         175       0
10        10    50          138.         205       0
# ℹ 90 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> Mult_r <- lm(df$BloodPressure ~ df$Age + df$Cholesterol +
df$Smoking)
> summary(Mult_r)
Call:
lm(formula = df$BloodPressure ~ df$Age + df$Cholesterol +
df$Smoking)
Residuals:
    Min      1Q  Median      3Q     Max
-3.6410 -0.5401 -0.0312  0.6096  1.5842

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    46.35460    3.20671  14.456  < 2e-16 ***
df$Age          0.17956    0.03311   5.423 4.36e-07 ***
df$Cholesterol  0.40351    0.02469  16.344  < 2e-16 ***
df$Smoking     -0.08335    0.49807  -0.167    0.867
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9012 on 96 degrees of freedom
Multiple R-squared:  0.9993,	Adjusted R-squared:  0.9993
F-statistic: 4.609e+04 on 3 and 96 DF,  p-value: < 2.2e-16
```

Compare between R calculated formula and manual one

R model is more accurate

$$Blood\ Pressure = 46.35 + 0.179 \times Age + 0.403 \times Cholesterol - 0.083 \times Smoking$$

$R^2 = 0.999$ (99.9%)

Common Pitfalls & Limitations

- **Correlation ≠ Causation:** Just because two variables are correlated does not mean one causes the other (e.g., ice cream sales and drowning incidents).

- **Overfitting:** Including too many predictors in regression can lead to unreliable models.

- **Confounding:** Unmeasured variables may distort relationships (e.g., alcohol consumption and lung cancer—smoking may be the real confounder).

Case Study: Predicting Disease Risk

Problem: **A hospital wants to predict heart attack risk based on:**

- Age
- Cholesterol level
- Physical activity (hours/week).

Regression Model:

Heart Attack Risk Score = 10 + 0.3(Age) + 0.02(Cholesterol) - 0.5(Physical Activity)

Interpretation:

- Every 1-hour increase in physical activity/week reduces risk by 0.5 points.
- Cholesterol has a smaller effect (0.02 per mg/dL) compared to age.

Conclusion

- ✓ Correlation identifies associations; regression helps predict and adjust for multiple factors.
- ✓ Always check for statistical significance (p-value) and effect size (R^2).
- ✓ Be cautious about confounding and causal inferences.
- ✓ Intelligent medical systems can leverage these methods for predictive diagnostics, risk

Comprehensive Questions

1. **Differentiate between Pearson’s correlation (r) and Spearman’s rank correlation (ρ). When would you use each in medical research? Provide examples.**
2. **Explain why correlation does not imply causation in epidemiology. Give a medical example where misinterpretation could lead to flawed public health decisions.**
3. **Regression Analysis**

   In the regression equation Y= a+b X + ε:

   - What do **a**, **b**, and ε represent?
   - How would you interpret a coefficient b =2.4*b* = 2.4 for a predictor like "BMI" in a diabetes risk model?
4. **Multiple regression adjusts for confounders. Suppose a model predicts heart disease risk using age, cholesterol, and smoking status:**
   - Why is this better than simple regression?
   - How would you interpret a negative coefficient for "physical activity" in this model?
5. **A study finds a correlation of r = 0.6 between air pollution (PM2.5) and asthma hospitalizations.**
   - What does this value imply about the relationship?
   - List two confounders you’d adjust for in a regression model to isolate pollution’s true effect.

Comprehensive Questions

6. **A regression model predicts ICU admission risk during COVID-19 using age, oxygen saturation, and comorbidities. The $R^2 = 0.45$:**
   - What does R^2 tell you about the model?
   - Suggest one limitation of relying solely on $R^2$for model evaluation.
7. **How can intelligent medical systems leverage regression analysis for:**
   - Early disease prediction (e.g., sepsis)?
   - Resource allocation during outbreaks?

THANK YOU!