Chapter 9
Correlation and Regression Analysis

Objectives

After completing this chapter, you will be able to:

- To learn Correlation and Regression Analysis
- Difference between correlation and linear regression
- Summary and Additional Information

Introduction

In this section we will first discuss correlation analysis, which is used to quantify the association between two continuous variables (e.g., between an independent and a dependent variable or between two independent variables). Regression analysis is a related technique to assess the relationship between an outcome variable and one or more risk factors or confounding variables. The outcome variable is also called the **response** or **dependent variable** and the risk factors and confounders are called the **predictors**, or **explanatory** or **independent variables**. In regression analysis, the dependent variable is denoted "*y*" and the independent variables are denoted by "*x*". [**NOTE:** The term "predictor" can be misleading if it is interpreted as the ability to predict even beyond the limits of the data. Also, the term "explanatory variable" might give an impression of a causal effect in a situation in which inferences should be limited to identifying associations. The terms "independent" and "dependent" variable are less subject to these interpretations as they do not strongly imply cause and effect.

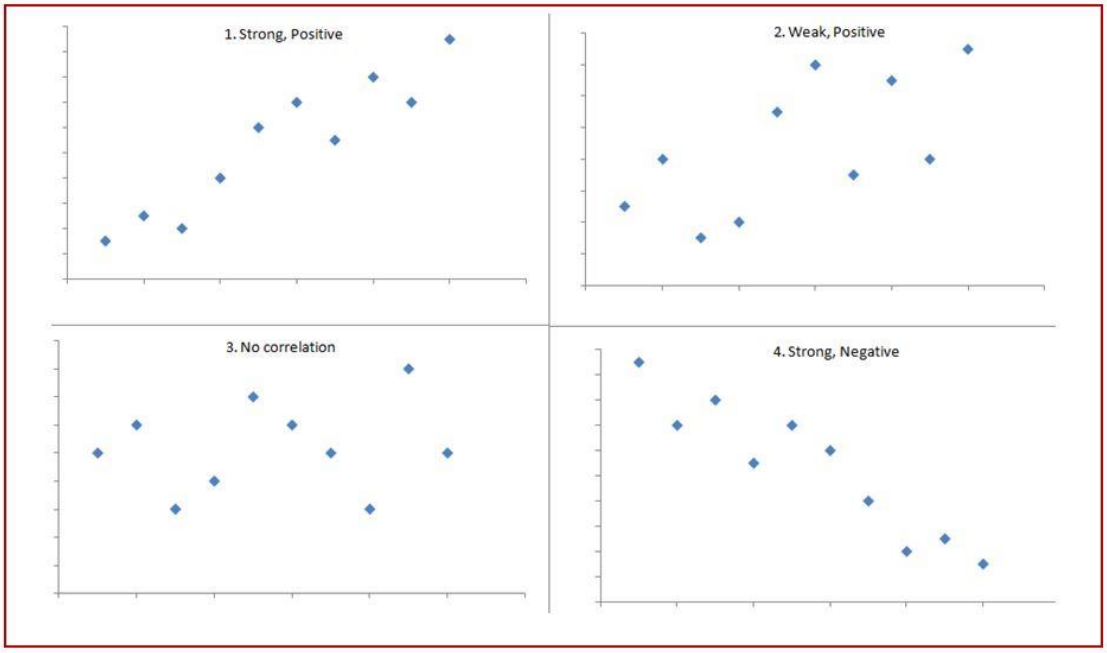
Correlation Analysis

In correlation analysis, we estimate a sample **correlation coefficient**, more specifically the **Pearson Product Moment correlation coefficient**. The sample correlation coefficient, denoted **r**, ranges between -1 and +1 and **quantifies the direction and strength of the linear association** between the two variables. The correlation between two variables can be positive (i.e., higher levels of one variable are associated with higher levels of the other) or negative (i.e., higher levels of one variable are associated with lower levels of the other).

The sign of the correlation coefficient indicates the direction of the association. **The magnitude of the correlation coefficient indicates the strength of the association.**

For example, a correlation of r = 0.9 suggests a strong, positive association between two variables, whereas a correlation of r = -0.2 suggest a weak, negative association. A correlation close to zero suggests no linear association between two continuous variables. It is important to note that there may be a non-linear association between two continuous variables, but computation of a correlation coefficient does not

detect this. Therefore, it is always important to evaluate the data carefully before computing a correlation coefficient. Graphical displays are particularly useful to explore associations between variables.
The figure below shows four hypothetical scenarios in which one continuous variable is plotted along the X-axis and the other along the Y-axis.



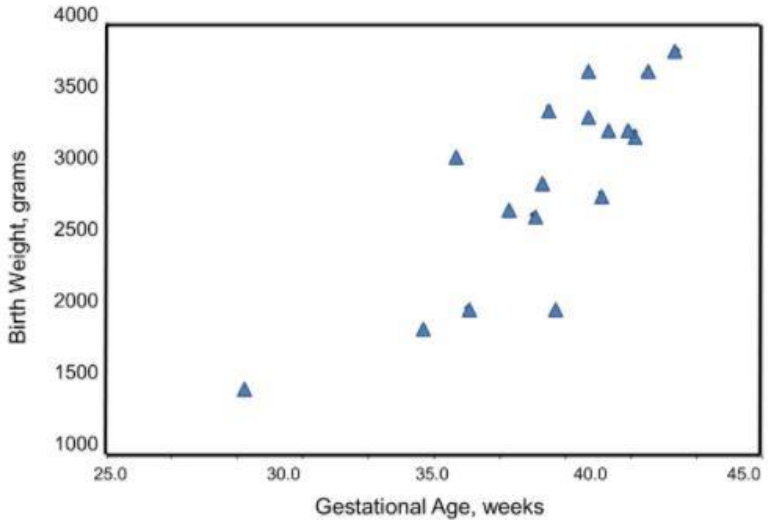
- Scenario 1 depicts a strong positive association (r=0.9), similar to what we might see for the correlation between infant birth weight and birth length.
- Scenario 2 depicts a weaker association (r=0,2) that we might expect to see between age and body mass index (which tends to increase with age).
- Scenario 3 might depict the lack of association (r approximately 0) between the extent of media exposure in adolescence and age at which adolescents initiate sexual activity.
- Scenario 4 might depict the strong negative association (r= -0.9) generally observed between the number of hours of aerobic exercise per week and percent body fat.

Example - Correlation of Gestational Age and Birth Weight

A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

We wish to estimate the association between gestational age and infant birth weight. In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus y=birth weight and x=gestational age. The data are displayed in a scatter diagram in the figure below.



Each point represents an (x,y) pair (in this case the gestational age, measured in weeks, and the birth weight, measured in grams). Note that the independent variable is on the horizontal axis (or X-axis), and the dependent variable is on the vertical axis (or Y-axis). The scatter plot shows a positive or direct association between gestational age and birth weight. Infants with shorter gestational ages are more likely to be born with lower weights and infants with longer gestational ages are more likely to be born with higher weights.

The formula for the sample correlation coefficient is

$$r = \frac{Cov(x,y)}{\sqrt{s_x^2 * s_y^2}}$$ where Cov(x,y) is the covariance of x and y defined as

$$Cov(x,y) = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n-1}$$ s_x^2 and s_y^2 are the sample variances of x and y, defined as

$$s_x^2 = \frac{\Sigma(X-\overline{X})^2}{n-1} \quad \text{and} \quad s_y^2 = \frac{\Sigma(Y-\overline{Y})^2}{n-1}$$

The variances of x and y measure the variability of the x scores and y scores around their respective sample means ($\overline{X}$ and $\overline{Y}$, considered separately). The covariance measures the variability of the (x,y) pairs around the mean of x and mean of y, considered simultaneously.

To compute the sample correlation coefficient, we need to compute the variance of gestational age, the variance of birth weight and also the covariance of gestational age and birth weight.

We first summarize the gestational age data. The mean gestational age is:

$$\overline{X} = \frac{\Sigma X}{n} = \frac{652.1}{17} = 38.4.$$ To compute the variance of gestational age, we need to sum the squared deviations (or differences) between each observed gestational age and the mean gestational age. The computations are summarized below.

Infant ID #	Gestational Age	$(X-\overline{X})$	$(X-\overline{X})^2$
1	34.7	-3.7	13.69
2	36.0	-2.4	5.76
3	29.3	-9.1	82.81
4	40.1	1.7	2.89
5	35.7	-2.7	7.29
6	42.4	4.0	16.00
7	40.3	1.9	3.61
8	37.3	-1.1	1.21
9	40.9	2.5	6.25
10	38.3	-0.1	0.01
11	38.5	0.1	0.01
12	41.4	3.0	9.00
13	39.7	1.3	1.69
14	39.7	1.3	1.69
15	41.1	2.7	7.29
16	38.0	-0.4	0.16
17	38.7	0.3	0.09
	$\Sigma X = 652.1$	$\Sigma(X-\overline{X}) = 0$	$\Sigma(X-\overline{X})^2 = 159.45$

The variance of gestational age is:

$$s_x^2 = \frac{\Sigma(X-\overline{X})^2}{n-1} = \frac{159.45}{16} = 10.0.$$

Next, we summarize the birth weight data. The mean birth weight is:

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{49{,}334}{17} = 2902.$$

The variance of birth weight is computed just as we did for gestational age as shown in the table below.

Infant ID #	Birth Weight	$(Y-\overline{Y})$	$(Y-\overline{Y})^2$
1	1895	-1007	1,014,049
2	2030	-872	760,384
3	1440	-1462	2,137,444
4	2835	-67	4,489
5	3090	188	35,344
6	3827	925	855,625
7	3260	358	128,164
8	2690	-212	44,944
9	3285	383	146,689
10	2920	18	324
11	3430	528	278,784
12	3657	755	570,025
13	3685	783	613,089
14	3345	443	196,249
15	3260	358	128,164
16	2680	-222	49,284
17	2005	-897	804,609
	Σ Y = 49,334	$\Sigma\ (Y-\overline{Y})=0$	$\Sigma(Y-\overline{Y})^2$ = 7,767,660

The variance of birth weight is:

$$s_y^2 = \frac{\Sigma(Y-\overline{Y})^2}{n-1} = \frac{7{,}767{,}660}{16} = 485{,}578.8.$$

Next we compute the covariance,

$$Cov(x,y) = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n-1}$$

To compute the covariance of gestational age and birth weight, we need to multiply the deviation from the mean gestational age by the deviation from the mean birth weight for each participant (i.e.,

$(X-\overline{X})(Y-\overline{Y}))$

The computations are summarized below. Notice that we simply copy the deviations from the mean gestational age and birth weight from the two tables above into the table below and multiply.

Infant Identification Number	$(X-\overline{X})$	$(Y-\overline{Y})$	$(X-\overline{X})(Y-\overline{Y})$
1	-3.7	-1007	3725.9
2	-2.4	-872	2092.8
3	-9.1	-1462	13,304.2
4	1.7	-67	-113.9
5	-2.7	188	-507.6
6	4.0	925	3700.0
7	1.9	358	680.2
8	-1.1	-212	233.2
9	2.5	383	957.5
10	-0.1	18	-1.8
11	0.1	528	52.8
12	3.0	755	2265.0
13	1.3	783	1017.9
14	1.3	443	575.9
15	2.7	358	966.6
16	-0.4	-222	88.8
17	0.3	-897	-269.1
			$\Sigma (X-\overline{X})(Y-\overline{Y})$= 28,768.4

The covariance of gestational age and birth weight is:

$$s_y^2 = \frac{\Sigma(Y-\overline{Y})^2}{n-1} = \frac{7,767,660}{16} = 485,578.8.$$

We now compute the sample correlation coefficient:

$$r = \frac{Cov(x,y)}{\sqrt{s_x^2 * s_y^2}} = \frac{1798.0}{\sqrt{10.0 * 485,578.8}} = \frac{1798.0}{2199.4} = 0.82.$$

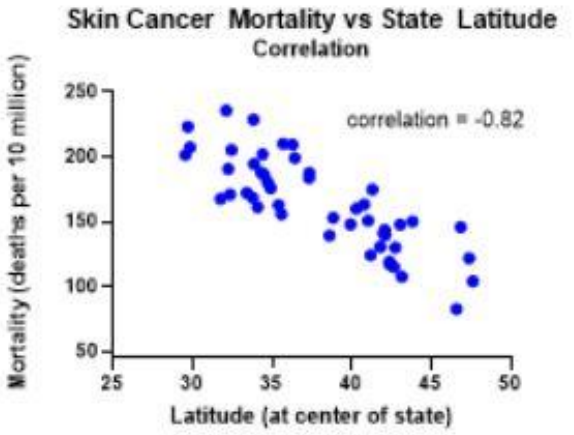
Not surprisingly, the sample correlation coefficient indicates a strong positive correlation.

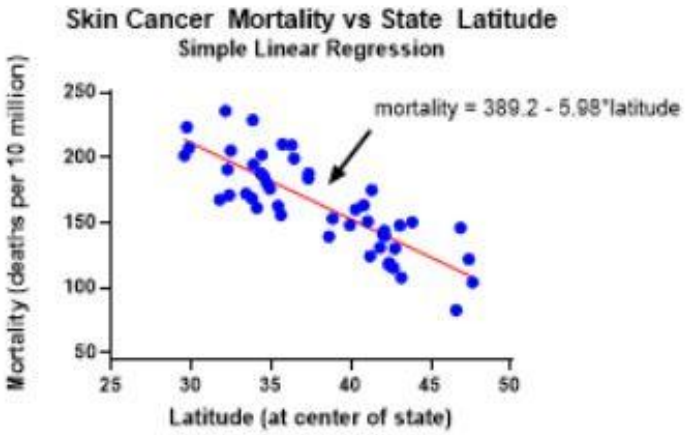
As we noted, sample correlation coefficients range from -1 to +1. In practice, meaningful correlations (i.e., correlations that are clinically or practically important) can be as small as 0.4 (or -0.4) for positive (or negative) associations.

What is the difference between correlation and linear regression?

When investigating the relationship between two or more numeric variables, it is important to know the difference between correlation and regression. The similarities/differences and advantages/disadvantages of these tools are discussed here along with examples of each.

Correlation quantifies the direction and strength of the relationship between two numeric variables, X and Y, and always lies between -1.0 and 1.0. **Simple linear regression** relates X to Y through an equation of the form Y = a + bX.





Correlation coefficient

The correlation coefficient is measured on a scale that varies from + 1 through 0 to - 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative.

Significance test

To test whether the association is merely apparent, and might have arisen by chance use the *t* test in the following calculation:

More advanced methods

More than one independent variable is possible - in such a case the method is known as multiple regression. (3,4)This is the most versatile of statistical methods and can be used in many situations.

Spearman rank correlation

A plot of the data may reveal outlying points well away from the main body of the data, which could unduly influence the calculation of the correlation coefficient. Alternatively the variables may be quantitative discrete such as a mole count, or ordered categorical such as a pain score.

This results in a simple formula for Spearman's rank correlation, Rho.

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

Table 11.2 Derivation of Spearman rank correlation from data of table 11.1

Child number	Rank height	Rank dead space	d	d^2
1	1	3	2	4
2	2	1	-1	1
3	3	2	-1	1
4	4	4	0	0
5	5	5.5	0.5	0.25
6	6	11	5	25
7	7	7	0	0
8	8	5.5	-2.5	6.25
9	9	8	-1	1
10	10	13	3	9
11	11	10	-1	1
12	12	9	-3	9
13	13	12	-1	1
14	14	15	1	1
15	15	14	-1	1
Total				60.5

From this we get that

$$r_s = 1 - \frac{6 \times 60.5}{15 \times (225 - 1)} = (0.8920)$$

In this case the value is very close to that of the Pearson correlation coefficient. For n> 10, the Spearman rank correlation coefficient can be tested for significance using the t test given earlier.

Summary and Additional Information

In summary, correlation and regression have many similarities and some important differences. Regression is primarily used to build models/equations to predict a key response, Y, from a set of predictor (X) variables. Correlation is primarily used to quickly and concisely summarize the direction and strength of the relationships between a set of 2 or more numeric variables.

The table below summarizes the key similarities and differences between correlation and regression.

Topic	Correlation	Regression
When to use	For a quick and simple summary of the direction and strength of pairwise relationships between two or more numeric variables.	To predict, optimize, or explain a numeric response Y from X, a numeric variable thought to influence Y.
Quantifies direction of relationship	Yes	Yes
Quantifies strength of relationship	Yes	Yes
X and Y interchangeable	Yes	No
Y Random	Yes	Yes
X Random	Yes	No
Prediction and Optimization	No	Yes
Equation	No	Yes
Extension to curvilinear fits	No	Yes
Cause and effect	No	Attempts to establish

Video Links:

Introduction to Correlation & Regression, Part 1

- https://www.youtube.com/watch?v=z7kMeJQWr4Y

Introduction to Correlation and Regression, Part 2

- https://www.youtube.com/watch?v=ZZNN7QXoYWw

Correlation & Regression: Concepts with Illustrative examples

- https://www.youtube.com/watch?v=xTpHD5WLuoA

References

- https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable5.html
- https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/
- https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression#:~:text=Correlation%20describes%20the%20strength%20of,correlation%20between%20B%20and%20A.&text=If%20y%20represents%20the%20dependent,regression%20of%20y%20on%20x.