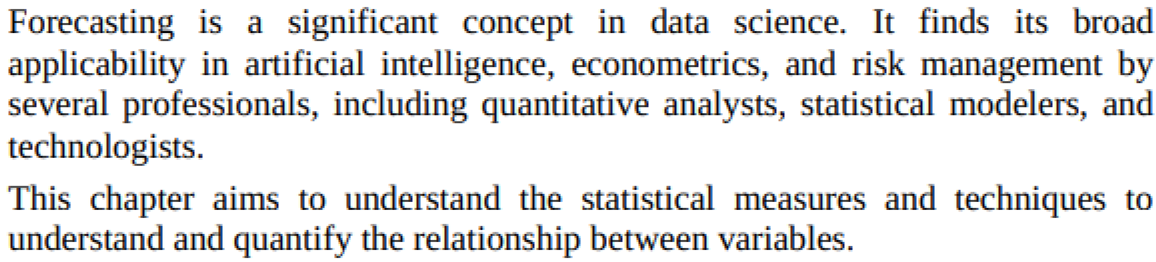
Week 8 Regression (LEC)

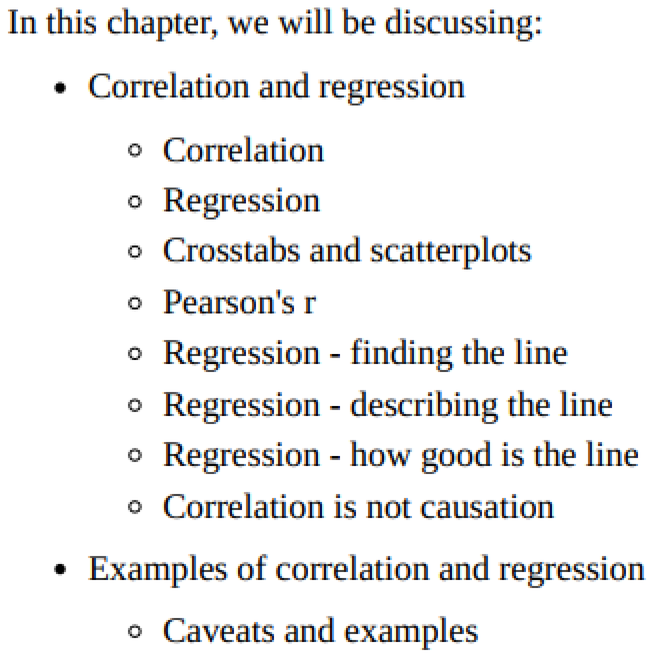
Introduction



Forecasting is a significant concept in data science. It finds its broad applicability in artificial intelligence, econometrics, and risk management by several professionals, including quantitative analysts, statistical modelers, and technologists.

This chapter aims to understand the statistical measures and techniques to understand and quantify the relationship between variables.

Structure



In this chapter, we will be discussing:

- Correlation and regression
  - Correlation
  - Regression
  - Crosstabs and scatterplots
  - Pearson's r
  - Regression - finding the line
  - Regression - describing the line
  - Regression - how good is the line
  - Correlation is not causation
- Examples of correlation and regression
  - Caveats and examples

Correlation and regression

Correlation and linear regression are the most commonly used statistical measurements for exploring the association between two quantitative variables

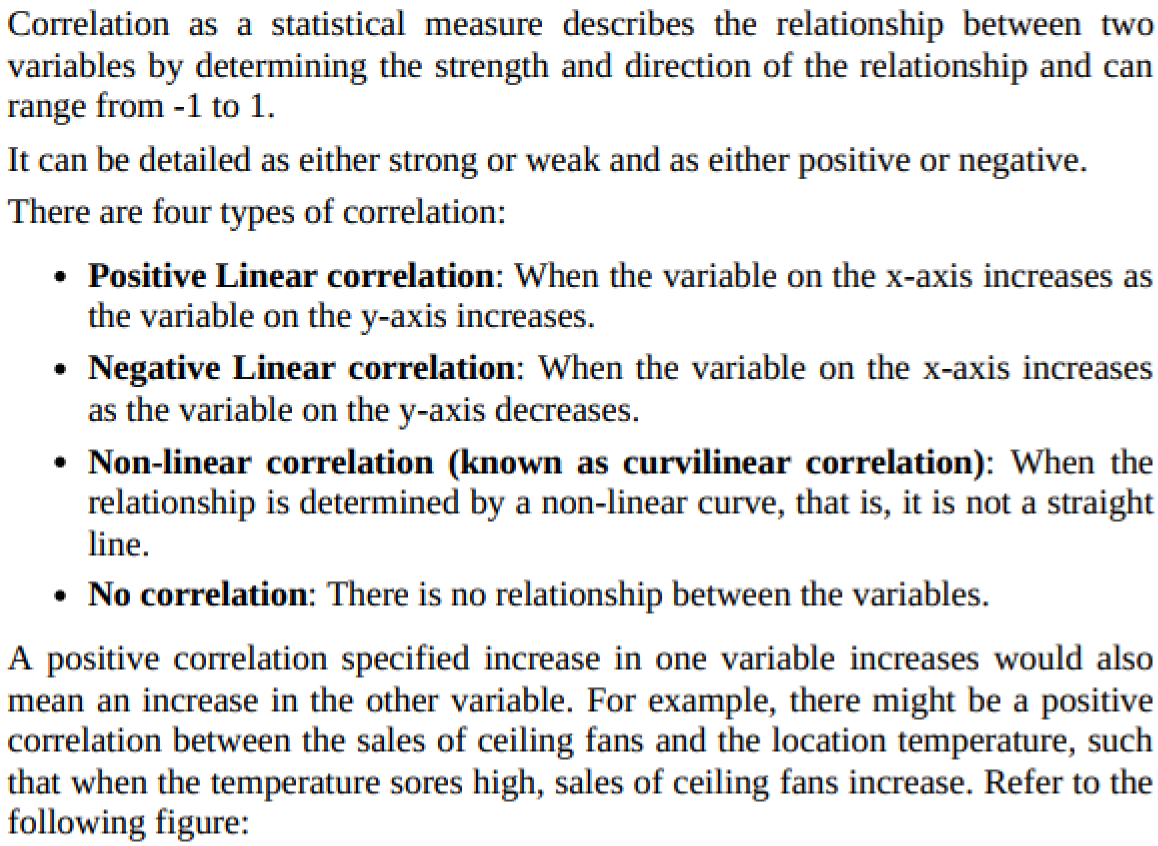
assumed to have a linear relationship.

While correlation quantifies the linear relationship between a pair of variables, regression indicates the relationship as an equation.

For example, suppose a lady owns a luxury house; then it is assumed that she must be financially well. Correlation and regression are used to numerically quantify this relationship.

A large number of business applications exist that use regression analysis, especially in finance, where it is used by analysts to understand the markets for stocks and hedge funds analysis to see changes in interest rates and how the bond price will charge. It can also be used to see the impact on employee productivity based on various pieces of training imparted, evaluate trends and make estimates in businesses, and more.

Correlation



Correlation as a statistical measure describes the relationship between two variables by determining the strength and direction of the relationship and can range from -1 to 1.

It can be detailed as either strong or weak and as either positive or negative.

There are four types of correlation:

- **Positive Linear correlation**: When the variable on the x-axis increases as the variable on the y-axis increases.
- **Negative Linear correlation**: When the variable on the x-axis increases as the variable on the y-axis decreases.
- **Non-linear correlation (known as curvilinear correlation)**: When the relationship is determined by a non-linear curve, that is, it is not a straight line.
- **No correlation**: There is no relationship between the variables.

A positive correlation specified increase in one variable increases would also mean an increase in the other variable. For example, there might be a positive correlation between the sales of ceiling fans and the location temperature, such that when the temperature sores high, sales of ceiling fans increase. Refer to the following figure:

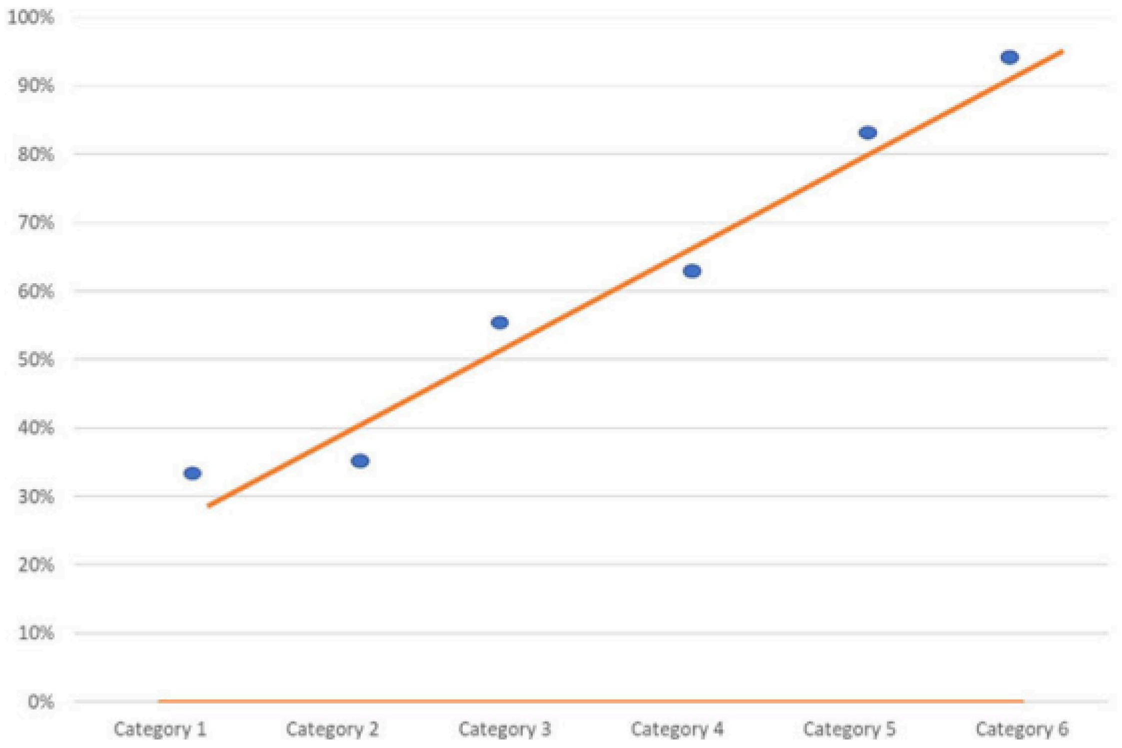


Figure 8.1: *Positive correlation*

A negative correlation signifies a decrease in one variable with an increase in the other variable. For example, there might be a negative correlation between the sales of heat radiators and local temperature, such that when the temperature drops, sales of heat radiators increase. Refer to the following figure:

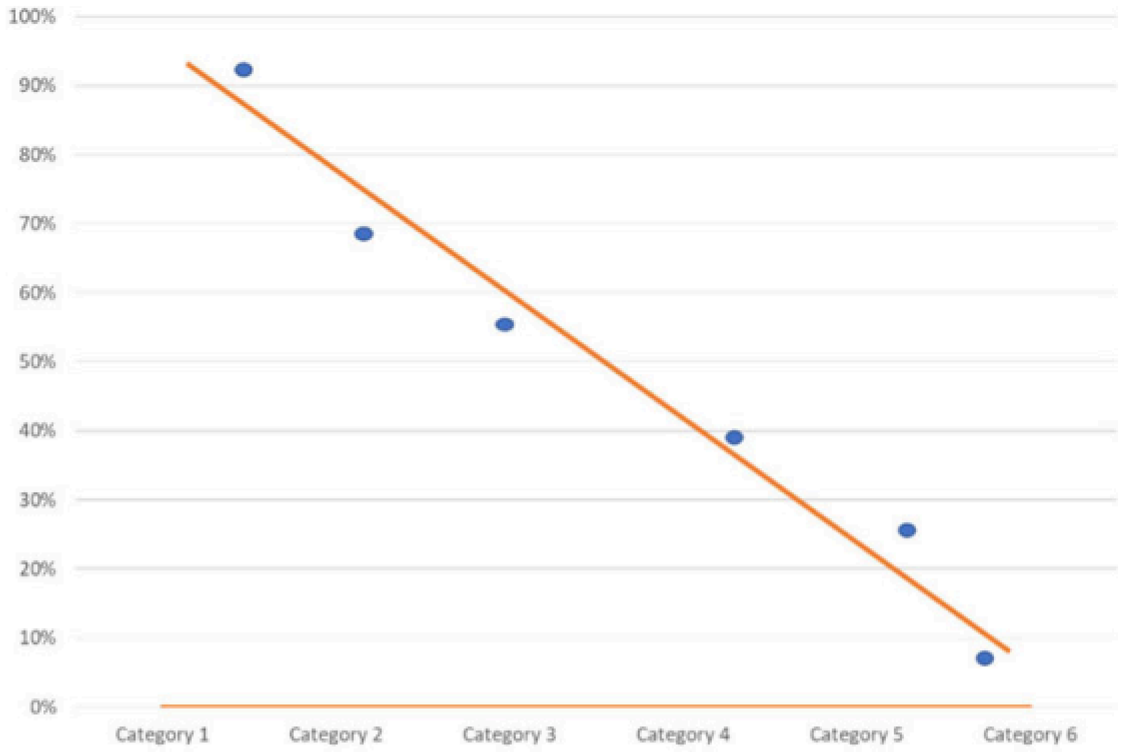


Figure 8.2: *Negative correlation*

Non-linear correlation (known as curvilinear correlation) is where there is a non-linear correlation between the variables. For example, the relationship between the restaurant's daily output of meals and the number of cooks is nonlinear. Refer to the following figure:

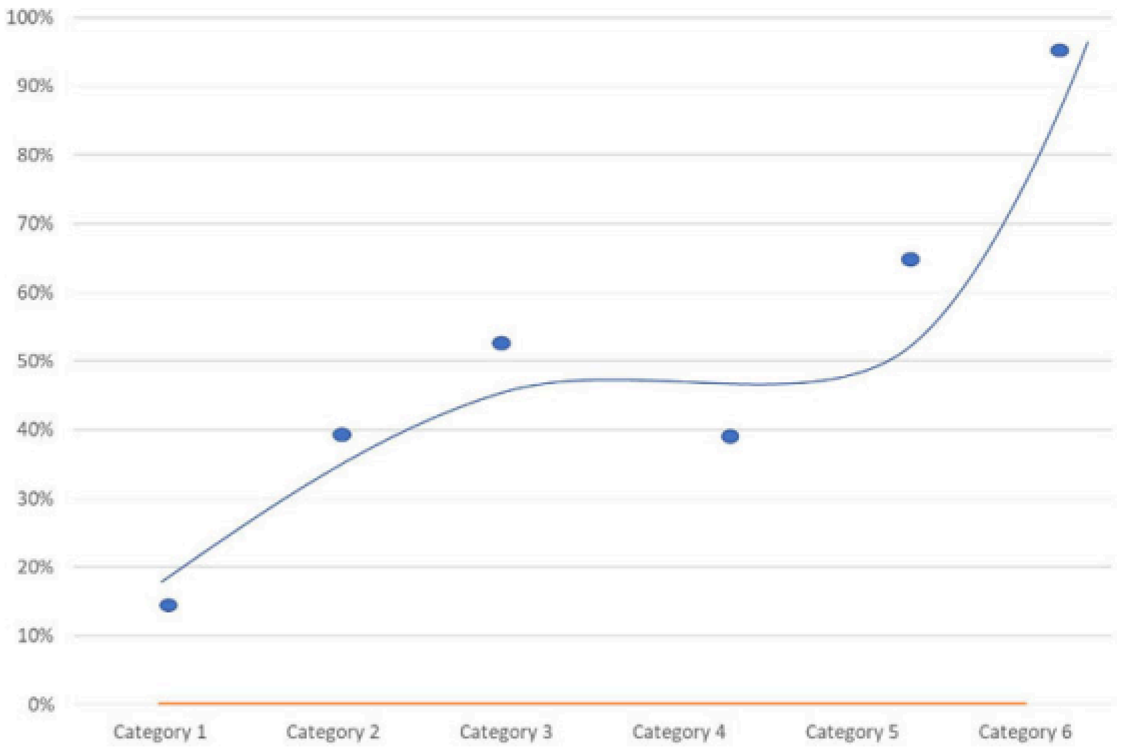


Figure 8.3: *non-linear correlation*

A correlation of 0 indicates that there is no relationship between the variables. For example, the number of sand particles on the road and the number of vehicles passing on the road on a daily basis have no correlation. Refer to the following figure:

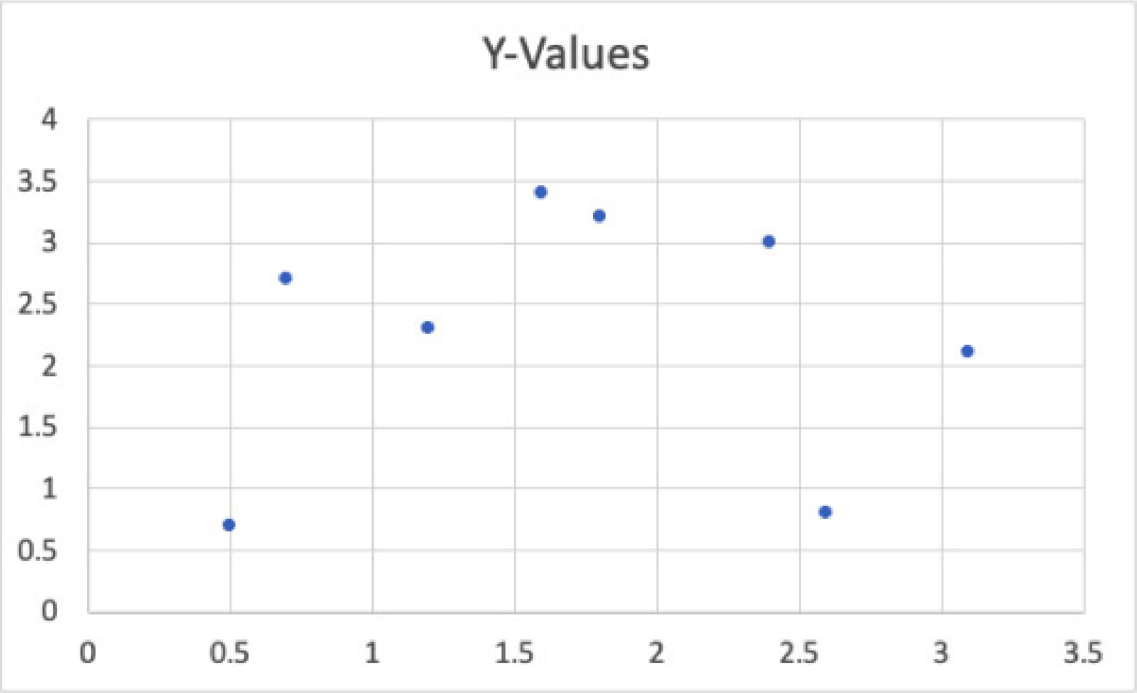
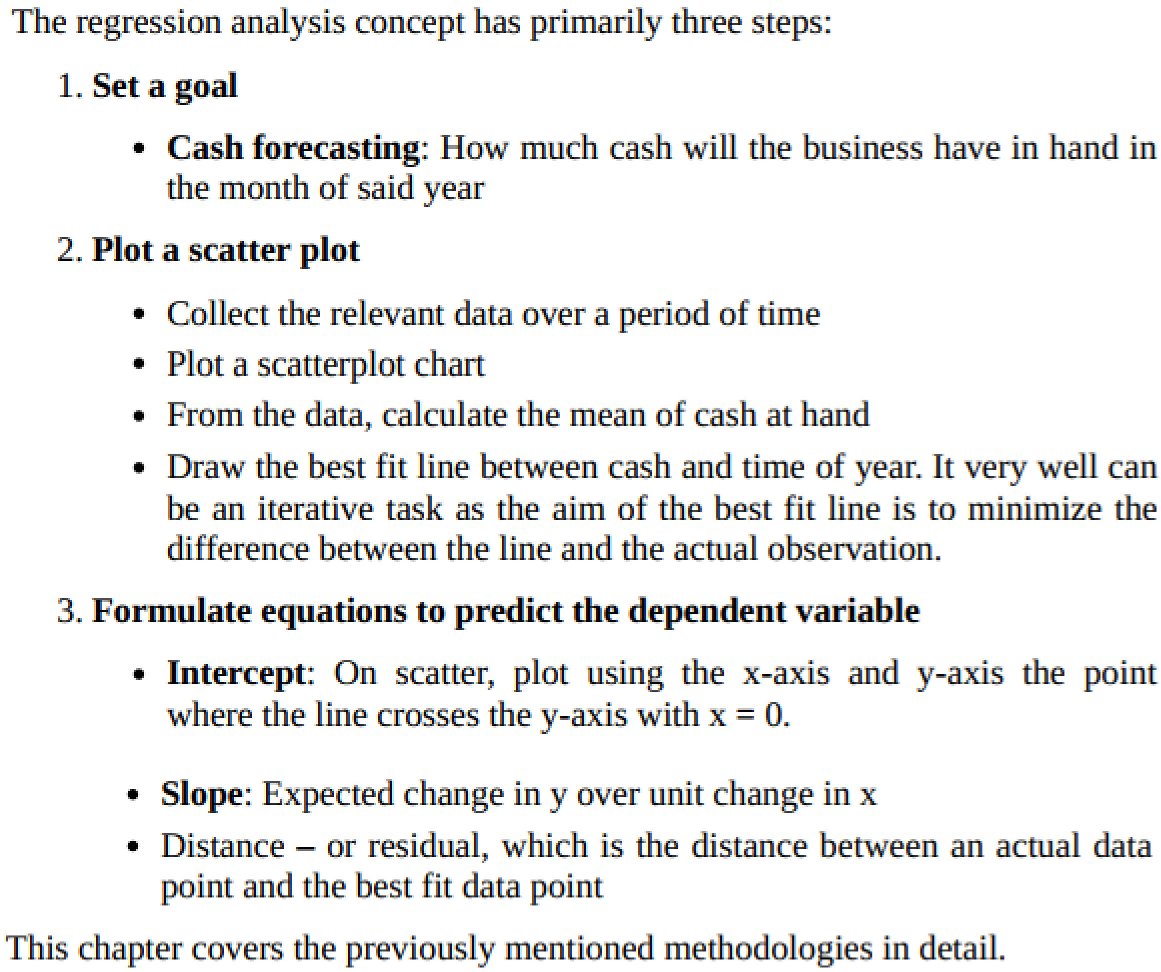


Figure 8.4: *No correlation*

Regression



The regression analysis concept has primarily three steps:

1. **Set a goal**
   - **Cash forecasting**: How much cash will the business have in hand in the month of said year
2. **Plot a scatter plot**
   - Collect the relevant data over a period of time
   - Plot a scatterplot chart
   - From the data, calculate the mean of cash at hand
   - Draw the best fit line between cash and time of year. It very well can be an iterative task as the aim of the best fit line is to minimize the difference between the line and the actual observation.
3. **Formulate equations to predict the dependent variable**
   - **Intercept**: On scatter, plot using the x-axis and y-axis the point where the line crosses the y-axis with x = 0.
   - **Slope**: Expected change in y over unit change in x
   - Distance – or residual, which is the distance between an actual data point and the best fit data point

This chapter covers the previously mentioned methodologies in detail.

Crosstabs and scatterplots

The crosstab and scatterplots are used to describe patterns across multiple variables, expressing how one variable may change (or correlate) with another.

Crosstabs

Cross-tabulation is a statistical technique that is used to analyze categorical data. Categorical data is data or variables that can be classified into different mutually exclusive categories. An example of categorical data is hair color.

Cross tabs (also known as contingency tables) display the relationship between two or more variables in the form of a table. They are used to determine the possible association between the variables. Crosstab is a great tool for presenting multidimensional data.

Example contingency table

Table 8.1 describes a crosstab that shows the number of persons in various age groups across genders:

	Number of persons	
Age group	**Male**	**Female**
18-22	30	32
23-32	45	40
33-42	30	28
43-52	20	25
53-62	29	25
62+	30	35

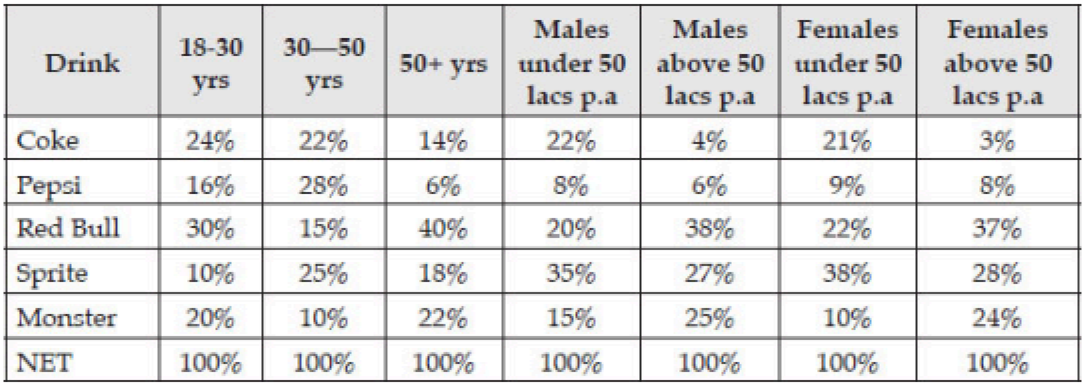
Table 8.1: *Crosstab*

For example, there are 30 people in the age group 18-22 who are male and 32 people in the same age group who are female.

Crosstabs are useful in not only understanding the relationship between different variables but also possibly identifying patterns or trends in the data and probabilities within data sets.

Crosstabs with more than two variables

Let us consider crosstabs that contain more than two variables. For example, Table 8.2 shows four variables. The rows represent one categorical variable that identifies drink preference, and the columns represent age and income within gender.



Drink	18-30 yrs	30—50 yrs	50+ yrs	Males under 50 lacs p.a	Males above 50 lacs p.a	Females under 50 lacs p.a	Females above 50 lacs p.a
Coke	24%	22%	14%	22%	4%	21%	3%
Pepsi	16%	28%	6%	8%	6%	9%	8%
Red Bull	30%	15%	40%	20%	38%	22%	37%
Sprite	10%	25%	18%	35%	27%	38%	28%
Monster	20%	10%	22%	15%	25%	10%	24%
NET	100%	100%	100%	100%	100%	100%	100%

Table 8.2: *Crosstab with more than two variables*

Crosstabs are routinely created with many more variables. For example, each row and each column may represent a different variable.

Scatterplots

A scatterplot (also known as a scatter graph or a scatter chart) is a graphical representation of two variables with values plotted as dots on a graph.

The position of each dot on the horizontal and vertical axis indicates values for an individual data point. That is, each observation is represented by a dot on the graph, with the position of the dot being determined by the values of the two variables.

Scatterplots are used to observe the relationship between two variables and identify any patterns or trends or see if there is any overlap between two sets of data. Refer to the following figure:

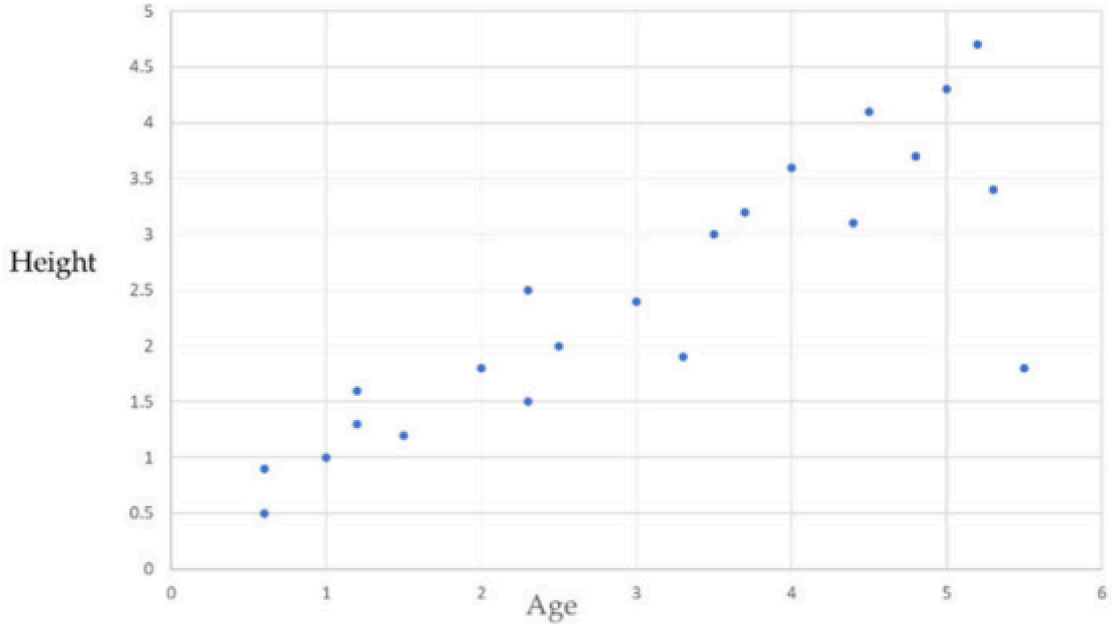


Figure 8.5: *Scatterplot of age and height*

Let’s say the scatterplot in graph 1 is a visualization of the relationship between two variables: X and Y. Each dot on the graph represents a single observation, with the position of the dot determined by the values of X and Y for that observation.

In graph 1, the height and age of the kids are plotted. X-axis represents the age in years while the X-axis represents the height in feet. Each dot represents a single kid. Each dot’s horizontal position indicates that kid’s age (in years), and the vertical position indicates the kid’s height (in feet).

From the plot, we can observe a general positive correlation between a kid’s age and a kid’s height, as the dots tend to fall along a line that ascends from left to right. Another point to observe is an outlier dot, a kid who is much older than the others but appears fairly short for their age, which might warrant further investigation.

Pearson's r

Pearson's r, also known as Pearson's correlation coefficient, is a measure of the strength and trend of the linear association between two variables. In other words, it determines if there is any linear component in the relationship between the two variables.

Pearson's correlation coefficient, as denoted by r, with values of r ranging from -1 to 1. The values of r closer to 1 indicate a strong positive correlation, values closer to -1 indicate a strong negative correlation, while values equal to zero indicate no linear correlation or association between the variables. Thus, a correlation coefficient of 0.83 indicates a stronger positive correlation than a value of 0.56. Similarly, a correlation coefficient of -0.38 indicates a stronger negative correlation than a correlation coefficient of -0.11

For values of the correlation coefficient:

- *A +1* value determines a perfect positive relationship between the variables
- *A -1* value determines a perfect negative relationship between the variables
- *A 0* value determines no relationship exists between the variables

$$r = \frac{\Sigma(\,x_i - \overline{x}\,)(\,y_i - \overline{y}\,)}{\sqrt{\Sigma\,(\,x_i - \overline{x}\,)^2\,\Sigma\,(\,y_i - \overline{y}\,)^2}}$$

Where,

r = correlation coefficient

x_i = values of the x-variable in a sample

y_i = mean of the values of the x-variable

$\tilde{x}$ = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Example of Pearson's r

Let's see the steps to calculate the Pearson correlation coefficient:

1. n represents the number of observations, let n = 3.
2. List the variables x and y, as in *Table 8.3*:

x	y
2	1
3	5
4	2

Table 8.3: *variables x and y*

3. Find product of x * y in 3rd column, as in *Table 8.4*:

x	y	x * y
2	1	2
3	5	15
4	2	8

Table 8.4: *Product of x * y*

4. Find x^2 and y^2 and mention in 4th and 5th columns, as in *Table 8.5*:

x	y	x * y	x^2	y^2
2	1	2	4	1
3	5	15	9	25
4	2	8	6	4

Table 8.5: *x^2 and y^2*

5. Find the sum of x, y, x * y, x^2, and y^2 variables and mention it in the last row (as marked in bold in the figure):

x	y	x * y	x^2	y^2
2	1	2	4	1
3	5	15	9	25
4	2	8	16	4
9	8	25	29	30

Table 8.6: *Sum of values x, y, x * y and x^2 and y^2*

6. Insert the values in the formula mentioned and solve it:

$$r = \left(\frac{(3 \times 25 - 9 \times 8)}{\sqrt{(3 \times 29 - 9^2)\,(3 \times 30 - 8^2)}}\right) = 0.019$$

Disadvantages

- The fact that the correlation coefficient between the variables X and Y or Y and X is the same, Pearson's r proves insufficient to differentiate between dependent and independent variables. For example, a person who exercises loses weight. But a person having lost weight need not necessarily be one who exercises.
- Doesn't provide information about the slope of the line. Rather only states the existence of a relationship between the two variables, if any.
- In the case of homogenous data, it may be likely to be misinterpreted.
- It takes much time to arrive at results as compared to other methods.

Important points

- Pearson's r is independent of units of the two variables.
- Pearson's r is symmetric between variables, which means the value of r between X and Y or Y and X remains the same.

Regression - Finding the line

To model the relationship between variables, the regression analysis technique aims to find the line (or a curve) that best fits the data in a way that can be used to make predictions about the dependent variable based on the independent variables.

There are six different types of regression analysis, including linear regression, logistic regression, and polynomial regression. This chapter is focused on linear regression.

In linear regression, the goal is to find the straight line, also called the regression line, that completely fits the data such that the overall distance from the line to the observation points outlined on the graph is the smallest.

The best-fitting line (or regression line) can be expressed with the formula $y = mx + b$, where m is the slope of the line and b is the y-intercept.

The equation $y = mx + b$ is a formula used for any straight line. However, note that in the case of regression, which is a statistical method, the observation points do not lie in a straight line. That is, if a linear pattern exists, the line is a model around which the data lie.

Now let's understand the equation $y = mx + b$ or, in other words, the way to describe the regression line:

Slope

Here m is the slope of the line, which defines the change of y over the change in x.

For example, a slope value of m = 5/9 means as the x-value moves (in the right direction) by 9 units, the y-value moves in the upward direction by 5 units.

Intercept

The y-intercept is the value on the y-axis where the regression line crosses. Note that the value of x-coordinates at the point where a line meets the y-axis is always 0.

For example, in the equation $y= 5x – 9$, the line crosses the y-axis at the value $b= –9$. The coordinates of this point are (0, –9).

Regression - Describing the line

The best–fitting regression line need not necessarily be found by hits and trials from the multiple options through eyeballing a line on the scatterplot.

Slope

The best-fitting line has a definite slope and a y-intercept that can be calculated using specific formulas, which is described as follows:

$$m = r\left(\frac{S_y}{S_x}\right)$$

Where,

m is the slope of the line

The standard deviation of the x values (denoted s_x)

The standard deviation of the y values (denoted s_y)

The correlation between X and Y (denoted r)

Note:

The slope can be negative, indicating the line is going downhill. For example, an increase in police patrolling resulted in a decrease in the number of crimes committed.

The correlation and the slope of the best–fitting line are not the same. While correlation is a unitless number, slope attaches units to the correlation.

y-intercept

The formula to calculate the y-intercept of the best fitting is as follows:

$$b = \overline{y} - m\overline{x}$$

where b is the y-intercept

where the mean of x-values as denoted by $\overline{x}$

and mean of y-values as denoted by $\overline{y}$

and m is the slope of the line

So, the slope of the line is always to be calculated first before calculating the y-intercept.

Where there are more than one independent variables, that is, in multiple regression models, the equation of the line would be more complex, with additional elements for each independent variable.

Let us revisit the correlation and regression formulas.

Pearson's correlation coefficient r is calculated as:

$$r = \frac{\Sigma(\, x_i - \overline{x}\,)(\, y_i - \overline{y}\,)}{\sqrt{\Sigma\,(\, x_i - \overline{x}\,)^2\, \Sigma\,(\, y_i - \overline{y}\,)^2}}$$

Ordinary Least Squares (OLS) Linear Regression is calculated as:

The straight-line equation is:

$$\mathrm{y} = \alpha + \beta X$$

OLS is represented by β, where,

$$\beta = \frac{\sum_1^n (\, \boldsymbol{x_i} - \overline{\boldsymbol{x}}\,)(\, \boldsymbol{y_i} - \overline{\boldsymbol{y}}\,)}{\sum_1^n (\, \boldsymbol{x_i} - \overline{\boldsymbol{x}}\,)^2}$$

$$\beta = r_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\alpha = \bar{y} - \beta \bar{x}$$

Where,

x is the independent variables

x- is the average of independent variables

y is the dependent variables

y- is the average of dependent variables

σ_x is the standard deviation of

σ_y is the standard deviation of

n is the number of data points in the data sets

Residual

Residual, or noise or error terms, indicate deviation of dependent values (Y values) from each expected value. In other words, Residual describes how good our model is against the actual value.

Ordinary Least Square helps us formulate the most suitable regression model. Using **Ordinary Least** (**OLS**) methodology and rules thus help us find the best-fit line or estimate the unknown parameters in a linear regression model.

Common OLS assumptions are as follows:

- **Errors**
  - Are independent of x
  - Have a constant variance
  - Their mean is 0
  - Are uncorrelated with each other
  - Have a normal distribution
- Large outliers are not considered observation points in the data
- Y and X variables have a linear relationship
- No appropriate independent variables have been omitted from the Model

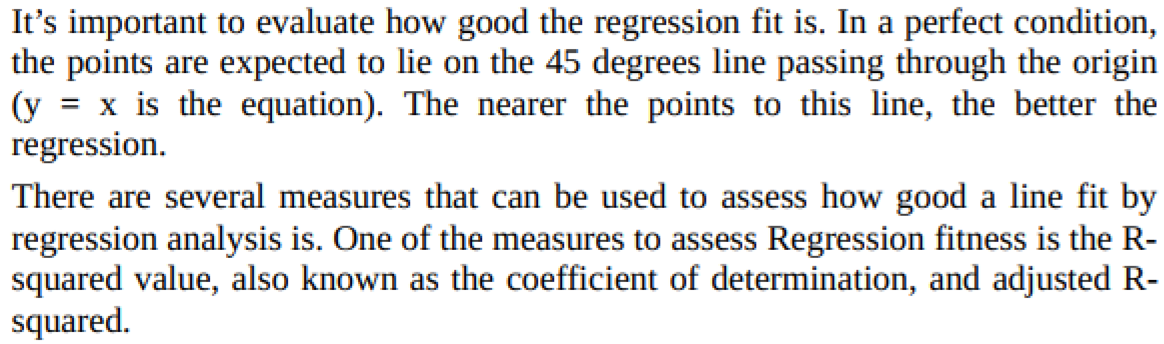
The best fit line is the line that minimizes the sum of squared differences between actual and forecasted results. Smaller the value, the better the regression model. **Mean Squared Error** (**MSE**) is the average of the minimum sum of squared difference. MSE is calculated as follows:

- **Explained Sum Of Squares** (**ESS**): Squaring(Y Estimated — Mean value of Y) and then sum all of the values.
- **Sum Of Squared Residuals** (**SSR**): Squaring(Y Estimated — Actual Y Value) And then sum all values.
- **Total Sum of Squares**, (**TSS**): ESS + SSR or the total sum of squares is the sum of all squares of (y estimated — residual) + sum of squares of (y estimated — mean y).

Note that Squaring values take care of negative values.

Now that the best-fit line has been described using formulas and we get the size of residuals. Let us also consider the line described in terms of its goodness of fit. How good is the line?

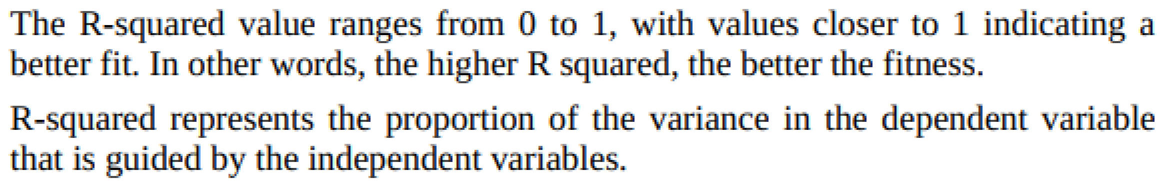
Regression - How good is the line



It’s important to evaluate how good the regression fit is. In a perfect condition, the points are expected to lie on the 45 degrees line passing through the origin (y = x is the equation). The nearer the points to this line, the better the regression.

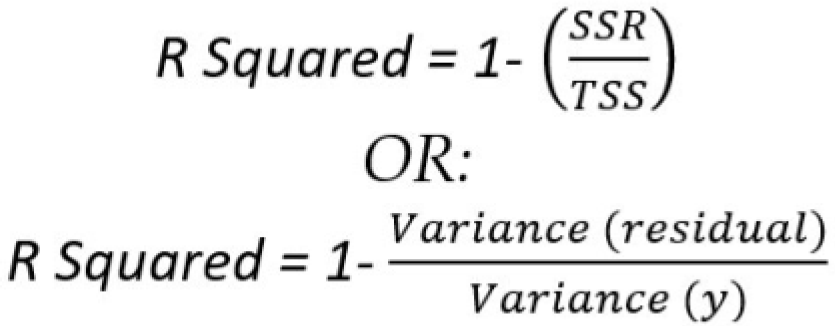
There are several measures that can be used to assess how good a line fit by regression analysis is. One of the measures to assess Regression fitness is the R-squared value, also known as the coefficient of determination, and adjusted R-squared.

R-Squared: goodness of fit



The R-squared value ranges from 0 to 1, with values closer to 1 indicating a better fit. In other words, the higher R squared, the better the fitness.

R-squared represents the proportion of the variance in the dependent variable that is guided by the independent variables.



$$R\ Squared = 1 - \left(\frac{SSR}{TSS}\right)$$

OR:

$$R\ Squared = 1 - \frac{Variance\ (residual)}{Variance\ (y)}$$

Adjusted R Squared

R squared lacks taking into account the number of variables that give the degree of determination. Hence R Squared by itself is not good enough. Therefore, adjusted R squared is calculated to measure the quality of the regression model.

$$\textit{Adjusted R Squared} = \frac{\frac{SSR}{(n-k)}}{SST\ (n-1)}$$

OR:

$$\textit{Adjusted R Squared} = 1 - \left[\left[\frac{n-1}{n-k-1}\right] \times [1 - R^2]\,\right]$$

Where,

n = number of observations,

k = number of independent variables

PS: Adjusted R square is always lower than the R-squared.

Another measure of how good the line is the standard error of the estimate. The standard error of the estimate is a measure of the amount of error or uncertainty in the predictions made by the model. A smaller standard error of the estimate indicates a better fit.

Please note that no single measure can fully represent the goodness of fit of a regression model. Rather, a combination of these measures should be used to get a holistic view of how well the model fits the data.

Correlation is not causation

In statistical terminology, correlation refers to a relationship between or interdependence of variables, such that they tend to vary together in a predictable way if the correlation is non-zero.

For example, there may be a positive correlation between the number of hours an athlete practices and their ranks, meaning that as the number of hours spent practicing increases, their ranks tend to improve.

Causation, on the other hand, means the action of causing something. Please note that correlation does not necessarily imply causation. In other words, if two variables are correlated, it does not mean that one variable is causing the other. Their relationship may be attributed to other factors, or the correlation may be due to chance.

For example, there may be a correlation between the number of cold drinks purchased and the amount of electricity bill in a given month. This does not mean that increase in cold drinks purchases caused higher electricity bills, but rather that a third factor may influence both variables, such as the temperature. In this case, the hot weather may be causing both an increase in cold drink purchases and an increase in the number of hours of running air conditioning, causing an increase in the electricity bill, leading to a correlation between the two variables. Refer to the following figure:

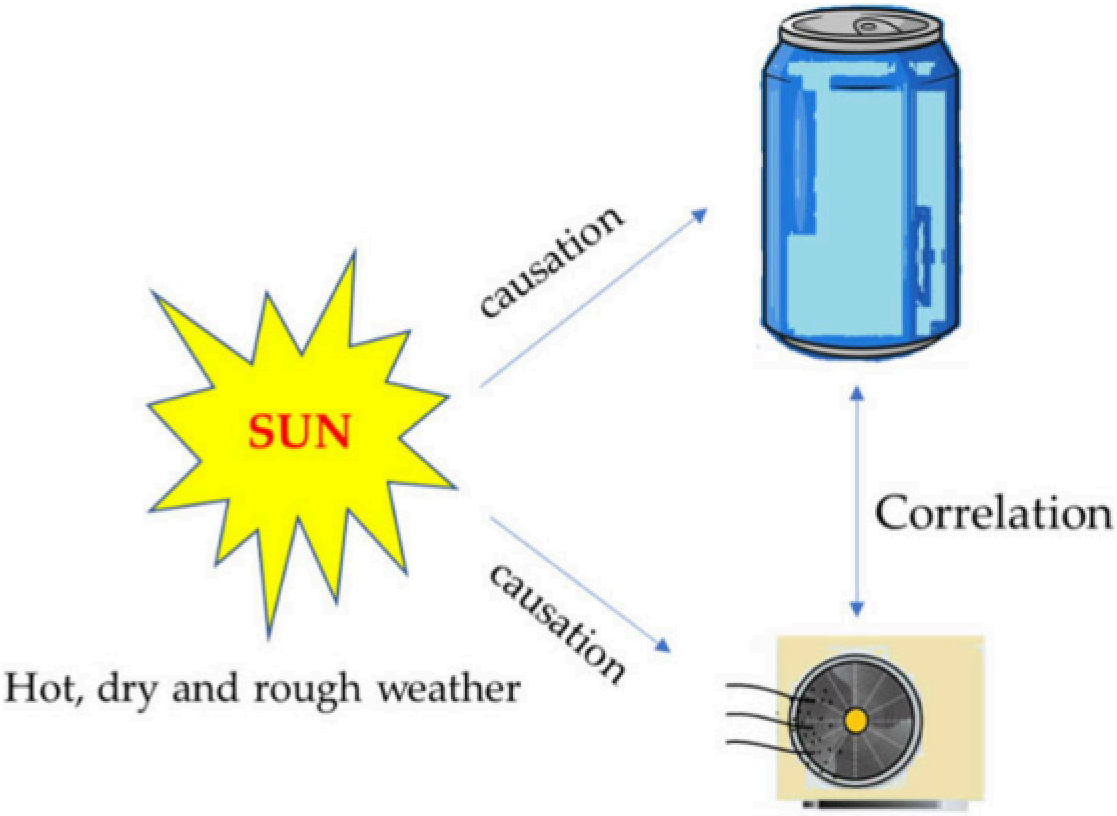


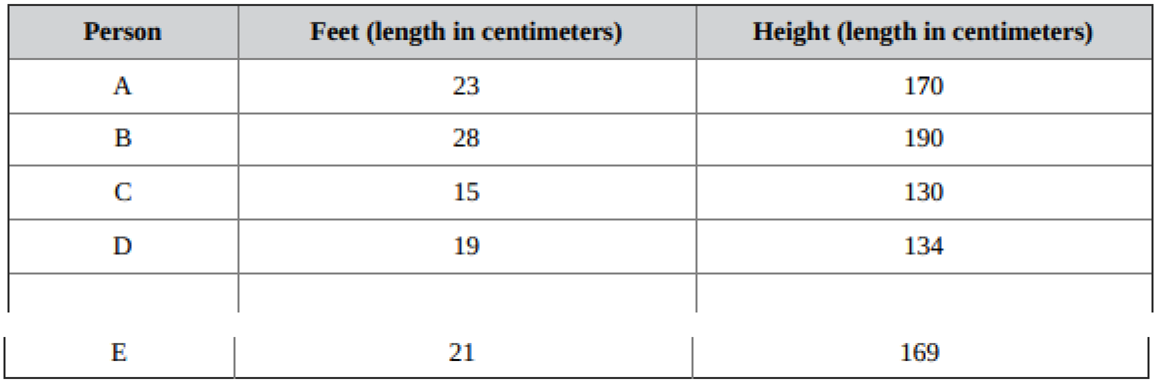
Figure 8.6: *Correlation is not causation*

Examples of correlation and regression

Let us consider examples of correlation and regression.

Example 1

Goal: We have two data sets regarding a person. One describes the length of feet, while the other describes the height. We need to determine the correlation between the length of feet and the height of a person. *Table 8.7* captures the feet and height of multiple persons:

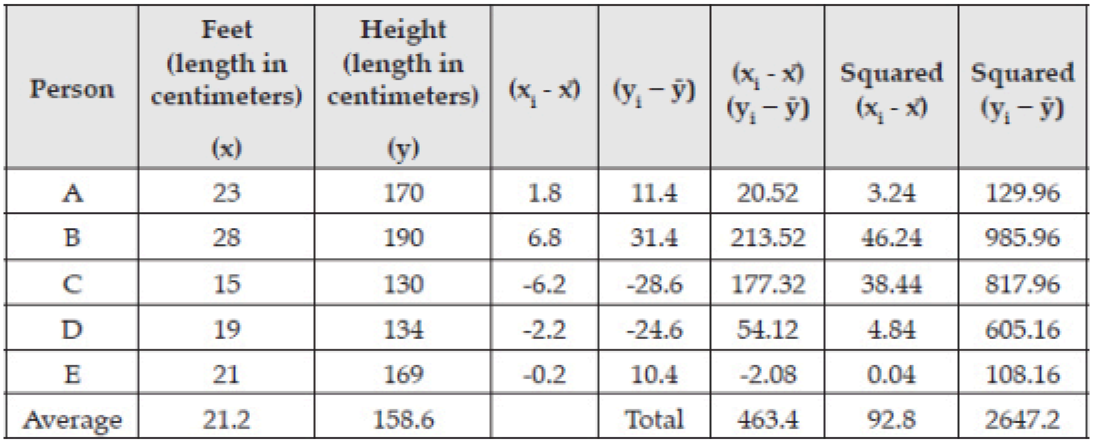


Person	Feet (length in centimeters)	Height (length in centimeters)
A	23	170
B	28	190
C	15	130
D	19	134
E	21	169

Table 8.7: *Feet and height of multiple persons*

Solution

Find Pearson's r as in *Table 8.8*:



Person	Feet (length in centimeters) (x)	Height (length in centimeters) (y)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})$ $(y_i - \bar{y})$	Squared $(x_i - \bar{x})$	Squared $(y_i - \bar{y})$
A	23	170	1.8	11.4	20.52	3.24	129.96
B	28	190	6.8	31.4	213.52	46.24	985.96
C	15	130	-6.2	-28.6	177.32	38.44	817.96
D	19	134	-2.2	-24.6	54.12	4.84	605.16
E	21	169	-0.2	10.4	-2.08	0.04	108.16
Average	21.2	158.6		Total	463.4	92.8	2647.2

Table 8.8: *Find Pearson's r*

Using formula

$$10. \quad r = \frac{\Sigma\,(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2\,(y_i - \bar{y})^2}} = 0.93$$

Answer: The data set has a high positive correlation, and the length of feet and height are strongly correlated.

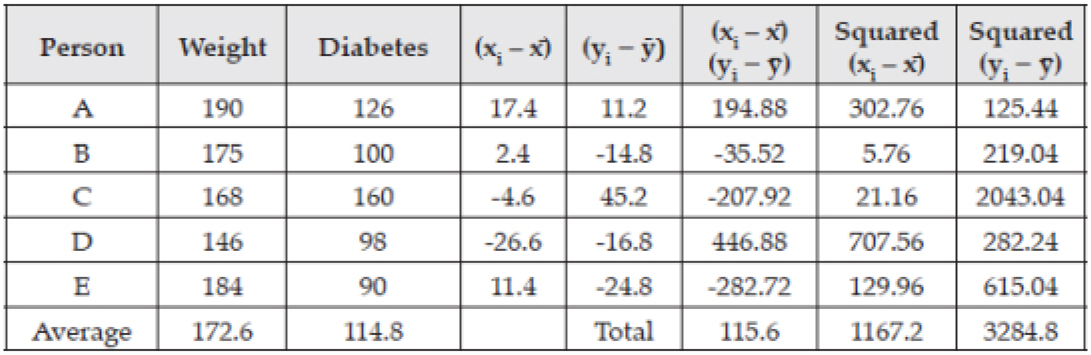
Example 2

Find the equation of the regression line for the following data as in *Table 8.9*:

Person	Weight	Diabetes
A	190	126
B	175	100
C	168	160
D	146	98
E	184	90

Table 8.9: *Weight and Diabetes of multiple persons*

Refer to the following *Table 8.10*:



Person	Weight	Diabetes	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})$ $(y_i - \bar{y})$	Squared $(x_i - \bar{x})$	Squared $(y_i - \bar{y})$
A	190	126	17.4	11.2	194.88	302.76	125.44
B	175	100	2.4	-14.8	-35.52	5.76	219.04
C	168	160	-4.6	45.2	-207.92	21.16	2043.04
D	146	98	-26.6	-16.8	446.88	707.56	282.24
E	184	90	11.4	-24.8	-282.72	129.96	615.04
Average	172.6	114.8		Total	115.6	1167.2	3284.8

Table 8.10: *Pearson's r*

Using formula

$$r = \frac{\Sigma\,(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^{\,2}\,(y_i - \bar{y})^{\,2}}} = 0.059$$

Answer: The data set has a positive correlation, and that weight and diabetes are weekly correlated.

Now the regression line can be calculated as follows:

$$\sigma x = \sqrt{\frac{\Sigma\,(x_i - \bar{x})^{\,2}}{n-1}} = 17.08$$

$$\sigma y = \sqrt{\frac{\Sigma\,(y_i - \bar{y})^{\,2}}{n-1}} = 28.65$$

$$r = 0.059$$

$$\beta = rxy\left(\frac{\sigma x}{\sigma y}\right) = 0.059 \times \left(\frac{17.08}{28.65}\right) = 0.035$$

$$\alpha = \bar{y} - \beta\bar{x} = 114.8 - 0.035 \times 172.6 = 108.75$$

Equation of regression line: y = 108.75 + 0.035x

Notes on Correlation and Regression

- **Both are statistical measurements that are used to quantify the strength of the linear relationship between two variables.**
- **While correlation determines if a linear relationship exists between two variables, regression describes the interconnection between the two.**
- **Pearson's correlation coefficient and the ordinary least squares method are used to perform correlation and regression analysis.**

Let us also take down the difference between correlation and regression as in *Table 8.11*:

Correlation	Regression
Determines whether variables are related or not	Describes how a dependent variable changes with a change in the independent variable
Tries to establish positive linear, negative linear or non-linear, or zero relationships between variables	It can be Linear Regression, Logistic Regression, Ridge Regression, Lasso Regression, Polynomial Regression, or Bayesian Linear Regression.
Variables can be used interchangeably; that is, it's symmetric	Variables cannot be interchanged
Determines the strength of the relationship using positive or negative numerical values	It describes the impact of change on a dependent variable basis change in an independent variable
Pearson's coefficient r is one of the measures of correlation.	The least-squares method is one of the techniques to determine the regression line.
A scatterplot displays the strength, direction, and form of the relationship, while a correlation coefficient measures the strength of that relationship	Regression lines, or the best fit lines, on scatterplots show the overall trend of a set of data

Table 8.11: *Difference between correlation and regression*

Caveats and examples

There are several things to keep in mind when working with regression and correlation:

- Correlation does not imply causation. Just because two variables are correlated does not mean that one causes the other. There could be a third variable that is causing both of them.

  For example, when RAM is 100% in use, the phone hangs as well as the camera doesn't work. There is a correlation between the camera doesn't work and the event that the phone hangs; however, one doesn't cause the other. There is a third variable, that is, RAM and its usage, that is causing the two. Refer to the following figure:

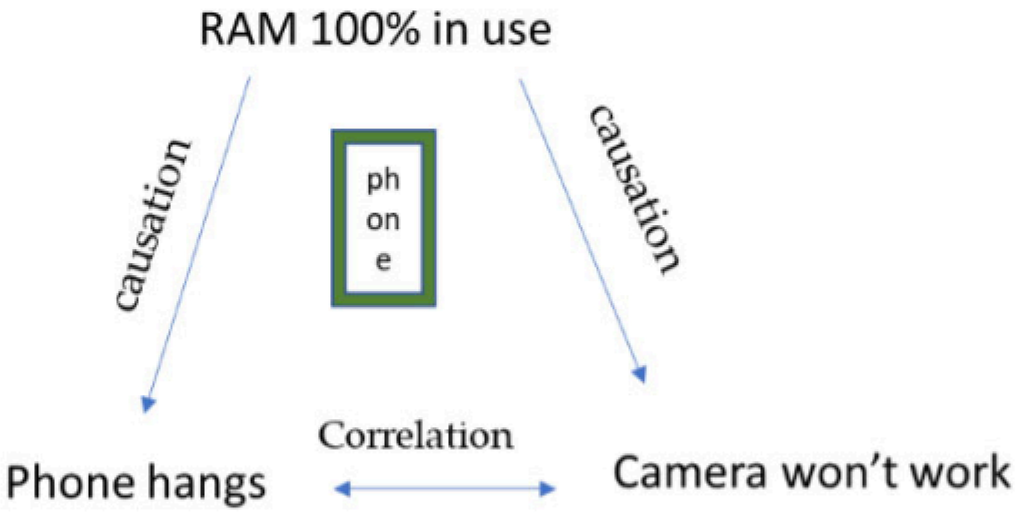


Figure 8.7: *Phone, Camera, and RAM*

- Linear regression is based on the assumption that the relationship between the variables is linear. If the relationship is non-linear, linear regression may not be the most appropriate method.

  For example, a linear regression model would assume that the relationship between a person's height based on their age is linear. However, children grow in height at a faster rate than adults, so the relationship between age and height may be non-linear. In this case, a non-linear model might be more appropriate.

- Outliers can have a big impact on the results of linear regression. A single outlier can significantly change the slope and intercept of the regression line.

  For example, let's say we are using linear regression to predict the price of a house based on its size. Most of the houses in our dataset range in size from 1,000 to 3,000 square feet and have a price ranging from $100,000 to $300,000. However, there is one house that is 10,000 square feet and has a price of $1 million. This outlier will have a big impact on the regression line and may not be representative of the true relationship between size and price for the majority of houses in the dataset.

- The regression line is only an approximation of the true relationship between the variables. It is not a guarantee of the actual relationship.

  For example, let's say we use linear regression to predict a person's weight based on their height. The regression line represents an increase of 1 pound in weight for every 0.5 inches increase in height. However, some people may be heavier or lighter than a forecast based on their height. As such, the linear regression tells us only an estimate, and the actual relationship between height and weight may not be the same.

- Regression and correlation are highly sensitive to the sample size. A larger sample size is generally considered more reliable, but a small sample size that is representative of the population still yields useful results.

  For example, for using linear regression to predict the percentage of college students based on the study hours per day, with a data sample of 10 students, our regression line may not be very reliable, as it is based on small sample size. However, if we collect a larger sample size, say data for 900 students, our regression line will be more reliable

- The regression line has its boundary and is only meaningful for the range of values that were used to fit the model. It may not be suitable for forecasting values outside of this range.

  For example, let the linear regression model predict a person's income based on their years of experience. If the model is based on data for people with 0-15 years of experience, the model may not be reliable for predicting the income of a person with 25 years of experience. The regression line is only meaningful for the range of values used to fit the model (in this case, 0-15 years of experience).

Conclusion

In this chapter, we learned about correlation and regression, and other related terms. We have acquired knowledge on how to relate data with regression and correlation. We also studied a few examples of the application of these statistical methods.

In the next chapter, we will study classification and clustering. These are the methodologies to categorize data into one or more classes based on the features.