UNIT-IV CORRELATION AND REGRESSION

Correlation coefficient:

The quantity r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson.

The correlation coefficient between two variables x and y is given by

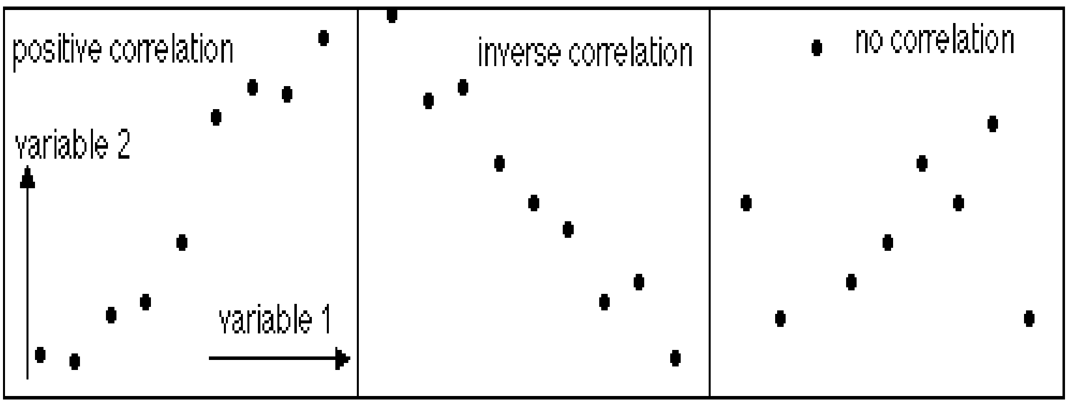
$$\rho \; or \; r = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}$$

where n is the number of pairs of data.

The value of r is such that $-1 < r < +1$. The $+$ and $-$ signs are used for positive linear correlations and negative linear correlations, respectively.

A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, the slope of this line is negative.



Positive correlation:

If x and y have a strong positive linear correlation, r is close to $+1$. An r value of exactly $+1$ indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.

Negative correlation:

If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.

UNIT-IV CORRELATION AND REGRESSION

No correlation:

If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables.

Note: r is a dimensionless quantity; that is, it does not depend on the units employed.

A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Coefficient of Determination r^2 or R^2 :

The coefficient of determination, r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.

The coefficient of determination is the ratio of the explained variation to the total variation.

The coefficient of determination is such that $0 < \boldsymbol{r^2} < 1$, and denotes the strength of the linear association between x and y. The coefficient of determination represents the percent of the data that is the closest to the line of best fit.

For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in *y* can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in *y* remains unexplained.

The coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

Example:

Emotion seems to play a pivotal role in determining popularity of a celebrity. In an exclusive survey made available to them by ad agency, Denstu-India, shows that the top 29 celebrity rankings are hugely impacted by the love/like quotient besides other parameters like performance. As per an article in Economic Times dt. 16th October 2006, the following are the scores for some of the Indian celebrities for the years 2005 and 2006.

UNIT-IV CORRELATION AND REGRESSION

Celebrity	Like Score 2005*	Like Score 2006**
Rahul Dravid	59	53
Amitabh Bachchan	56	51
Sachin Tendulkar	43	50
Aishwarya Rai	56	50
Sania Mirza	21	49
Yuvaraj Singh	61	46
Sushmita Sen	56	46
Virendra Sehwag	64	46
Aamir Khan	57	45
Rani Mukherjee	57	45

Find the correlation coefficient between Like score 2005 and 2006.

Solution:

Here n=10

x	y	xy	x^2	y^2
59	53	3127	3481	2809
56	51	2856	3136	2601
43	50	2150	1849	2500
56	50	2800	3136	2500
21	49	1029	441	2401
61	46	2806	3721	2116
56	46	2576	3136	2116
64	46	2944	4096	2116
57	45	2565	3249	2025
57	45	2565	3249	2025
530	481	25418	29494	23209

$$\sum x_i = 530, \quad \sum y_i = 481, \quad \sum {x_i}^2 = 29494, \quad \sum {y_i}^2 = 23209,$$

$$\sum x_i y_i = 25418$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{530}{10} = 53$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{481}{10} = 48.1$$

$$\rho = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum x_i{}^2 - n\bar{x}^2}\sqrt{\sum y_i{}^2 - n\bar{y}^2}}$$

$$\rho = \frac{25418 - 10 \times 53 \times 48.1}{\sqrt{29494 - 10 \times (53)^2}\sqrt{23209 - 10 \times (48.1)^2}}$$

$$\rho = \frac{-75}{\sqrt{1404}\sqrt{72.9}}$$

$$\rho = -0.23$$

Example:

The following data gives the closing prices of BSE Sensex, and the stock price of an individual company viz. ICICI bank during the 10 trading days during the period from 6th to 21st March 2006.

Date	SENSEX	ICICI Bank
6-3-2006	10735	613.20
7-3-2006	10725	600.65
8-3-2006	10509	590.55
9-3-2006	10574	601.75
10-3-2006	10765	612.90
13-3-2006	10804	603.10
16-3-2006	10878	607.50
17-3-2006	10860	605.25
20-3-2006	10941	605.40
21-3-2006	10905	597.80

Find the correlation coefficient between SENSEX and ICICI Bank.

Solution:

x	y	xy	x^2	y^2
10735	613.2	6582702	115240225	376014.2
10725	600.65	6441971	115025625	360780.4
10509	590.55	6206090	110439081	348749.3
10574	601.75	6362905	111809476	362103.1
10765	612.9	6597869	115885225	375646.4
10804	603.1	6515892	116726416	363729.6
10878	607.5	6608385	118330884	369056.3
10860	605.25	6573015	117939600	366327.6
10941	605.4	6623681	119705481	366509.2
10905	597.8	6519009	118919025	357364.8
107696	**6038.1**	**65031519**	**1160021038**	**3646281**

$$\sum x_i = 107696, \quad \sum y_i = 6038.1, \quad \sum x_i{}^2 = 1160021038\,, \quad \sum y_i{}^2 = 3646281,$$

$$\sum x_i y_i = 65031519, \mathrm{n} = 10$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{107696}{10} = 10769.6$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{6038.1}{10} = 603.81$$

$$\rho = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum x_i{}^2 - n\bar{x}^2}\sqrt{\sum y_i{}^2 - n\bar{y}^2}}$$

$$\rho = \frac{65031519 - 10 \times 10769.6 \times 603.81}{\sqrt{1160021038 - 10 \times (10769.6)^2}\sqrt{3646281 - 10 \times (603.81)^2}}$$

$$\rho = \frac{3597.24}{\sqrt{178196.4}\sqrt{415.84}}$$

$$\rho = 0.42$$

UNIT-IV CORRELATION AND REGRESSION

Spearman Rank Correlation:

Spearman Rank Correlation is a measure of the strength of the associations between two variables. Spearman's Rank correlation coefficient is a technique which can be used to summarize the strength and direction (negative or positive) of a relationship between two variables.
Spearman Rank Correlation between two variables x and y is given by

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

Always r_s lies between -1 and $+1$.

Method - calculating the coefficient:

- Create a table from your data.
- Rank the two data sets. Ranking is achieved by giving the ranking '1' to the biggest number in a column, '2' to the second biggest value and so on. The smallest value in the column will get the lowest ranking. This should be done for both sets of measurements.
- Tied scores are given the mean (average) rank. For example, the three tied scores of 1 euro in the example below are ranked fifth in order of price, but occupy three positions (fifth, sixth and seventh) in a ranking hierarchy of ten. The mean rank in this case is calculated as (5+6+7) ÷ 3 = 6.
- Calculate the coefficient r_s by the formula given below

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

Example:

As per a study, the following are the ranks of priorities for ten factors taken as 'Job Commitment Drivers' among the executives in Asia Pacific (AP) and India. Calculate the rank correlation between properties of 'Job Commitment Drivers' among executives from India and Asia Pacific.

Job Commitment Drivers	Favourable Rank	
	India	Asia Pacific
Job satisfaction	1	1
Work environment	2	2
Team work	3	4
Communication	4	3
Performance Management	5	5
Innovation	6	6
Leadership	7	9
Training and development	8	7
Supervision	9	8
Compensation/Benefits	10	10

Solution:

Here $n = 10$

r_1	r_2	$d_i = r_1 - r_2$	d_i^2
1	1	0	0
2	2	0	0
3	4	-1	1
4	3	1	1
5	5	0	0
6	6	0	0
7	9	-2	4
8	7	1	1
9	8	1	1
10	10	0	0
Sum			8

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

$$= 1 - \frac{6(8)}{10(10^2 - 1)} = 1 - 0.048$$

$$r_s = 0.952$$

Example:

Calculate rank correlation coefficient between the two series X and Y, given below:

X	70	65	71	62	58	69	78	64
Y	91	76	65	83	90	64	55	48

Solution:

$Here\ n = 8$

X	$Rank\ of\ X$ r_1	y	$Rank\ of\ y$ r_2	$d = r_1 - r_2$	d^2
70	6	91	8	-2	4
65	4	76	5	-1	1
71	7	65	4	3	9
62	2	83	6	-4	16
58	1	90	7	-6	36
69	5	64	3	2	4
78	8	55	2	6	36
64	3	48	1	2	4
Sum					110

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

$$= 1 - \frac{6 \times 110}{8(8^2 - 1)}$$

$$= 1 - \frac{660}{504}$$

$$= 1 - 1.3095$$

$$r_s = -0.3095$$

Regression analysis:

A statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

Types of Regression:

The two basic types of regression are linear regression and multiple regression.

- Linear regression uses one independent variable to explain and/or predict the outcome of Y.
- Multiple regression uses two or more independent variables to predict the outcome.

The general form of each type of regression is:

Linear Regression: $Y = a + bX$

Multiple Regression: $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_tX_t$

Where

Y= the variable that we are trying to predict

X= the variable that we are using to predict Y

a= the intercept

b= the slope

In multiple regression the separate variables are differentiated by using subscripted numbers.

Regression takes a group of random variables, thought to be predicting Y, and tries to find a mathematical relationship between them. This relationship is typically in the form of a straight line (linear regression) that best approximates all the individual data points. Regression is often used to determine how much specific factors such as the price of a commodity, interest rates, particular industries or sectors influence the price movement of an asset.

Finding the regression line using method of least squares:

The regression line y on x is given by $y = a + bx$

Where

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

UNIT-IV CORRELATION AND REGRESSION

The regression line x on y is given by $x = c + dy$

Where

$$d = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum y_i^2 - n\bar{y}^2}$$

$$c = \bar{x} - d\bar{y}$$

$$\bar{x} = \frac{\sum x_i}{n}, \qquad \bar{y} = \frac{\sum y_i}{n} \ and$$

n is the number of pairs of data.

Standard error of estimator (Regression Line):

The square root of the residual variance is called the standard error of regression line.

The residual variance for the regression line $y = a + bx$ is given by

$$\sigma_e^2 = \frac{\sum(y_i - \widehat{y_i})^2}{n} \ where \ \widehat{y_i} = a + bx_i$$

Example:

A tyre manufacturing company is interested in removing pollutants from the exhaust at the factory, and cost is a concern. The company has collected data from other companies concerning the amount of money spent on environmental measures and the resulting amount of dangerous pollutants released (as a percentage of total emissions)

Money spent (Rupees in lakhs)	8.4	10.2	16.5	21.7	9.4	8.3	11.5	18.4	16.7	19.3	28.4	4.7	12.3
Percentage of dangerous pollutants	35.9	31.8	24.7	25.2	36.8	35.8	33.4	25.4	31.4	27.4	15.8	31.5	28.9

a) Compute the regression equation.

b) Predict the percentage of dangerous pollutants released when Rs. 20,000 is spent on control measures.

c) Find the standard error of the estimate(regression line).

Solution:

Let x and y represents money spent and percentage of dangerous pollutants respectively.

Here n= 13

x	y	xy	x^2	$\hat{y} = a + bx$	$y - \hat{y}$	$(y - \hat{y})^2$
8.4	35.9	301.56	70.56	30.3472	5.5528	30.83359
10.2	31.8	324.36	104.04	30.1006	1.6994	2.88796
16.5	24.7	407.55	272.25	29.2375	-4.5375	20.58891
21.7	25.2	546.84	470.89	28.5251	-3.3251	11.05629
9.4	36.8	345.92	88.36	30.2102	6.5898	43.42546
8.3	35.8	297.14	68.89	30.3609	5.4391	29.58381
11.5	33.4	384.1	132.25	29.9225	3.4775	12.09301
18.4	25.4	467.36	338.56	28.9772	-3.5772	12.79636
16.7	31.4	524.38	278.89	29.2101	2.1899	4.795662
19.3	27.4	528.82	372.49	28.8539	-1.4539	2.113825
28.4	15.8	448.72	806.56	27.6072	-11.8072	139.41
4.7	31.5	148.05	22.09	30.8541	0.6459	0.417187
12.3	28.9	355.47	151.29	29.8129	-0.9129	0.833386
185.8	384	5080.27	3177.12			310.8354

a) $y = a + bx$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{185.8}{13} = 14.29$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{384}{13} = 29.54$$

$$b = \frac{\sum x_i y_i - n\,\bar{x}\,\bar{y}}{\sum x_i^2 - n\,\bar{x}^2}$$

$$= \frac{5080.27 - 13 \times 14.29 \times 29.54}{3177.12 - 13 \times 14.29^2} = \frac{-407.65}{2972.92} = -0.137$$

$$a = \bar{y} - b\bar{x} = 29.54 - (-0.137) \times 14.29 = 31.498$$

$$y = 31.498 - 0.137\,x$$

b) When Rs. 20,000 is spent on control then the percentage of dangerous pollutants released is

$$y = 31.498 - 0.137 \times 0.2$$

$$y = 31.471$$

c)

$$\sigma_e^2 = \frac{\sum(y_i - \hat{y_i})^2}{n} \; where \; \hat{y_i} = a + bx_i$$

$$\sigma_e^2 = \frac{310.8354}{13} = 23.91$$

$$Standard\ error = \sqrt{\sigma_e^2} = \sqrt{23.91} = 4.89$$

$$Standard\ error = 4.89$$

Example:

A national level organization wishes to prepare a manpower plan based on the ever-growing sales offices in the country. Find the regression coefficient of Manpower on Sales Offices for the following data:

Year	Manpower	Sales Offices
2001	370	22
2002	386	25
2003	443	28
2004	499	31
2005	528	33
2006	616	38

Solution:

Let x and y represents sales offices and manpower respectively.

Here $n = 6$,

x	y	xy	x^2
22	370	8140	484
25	386	9650	625
28	443	12404	784
31	499	15469	961
33	528	17424	1089
38	616	23408	1444
177	2842	86495	5387

$$\sum x_i = 177, \sum y_i = 2842, \sum x_i y_i = 86495, \sum x_i^2 = 5387$$

The regression line of Y on X is given by $Y = a + bX$

$$where\ b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{177}{6} = 29.5$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{2842}{6} = 473.67$$

$$b = \frac{86495 - 6 \times 29.5 \times 473.67}{5387 - 6 \times (29.5)^2} = 16.04$$

$$a = \bar{y} - b\bar{x} = 473.67 - 16.04 \times 29.5 = 0.49$$

The regression line of manpower on sales offices is given by

$$y = 0.49 + 16.04\,x$$

Example:

The quantity of a raw material purchased by a company at the specified prices during the 12 months of 1992 is given

MONTH	PRICE/KG	QUANTITY (KG)
Jan	96	250
Feb	110	200
Mar	100	250
Aprl	90	280
May	86	300
June	92	300
July	112	220
Aug	112	220
Sep	108	200
Oct	116	210
Nov	86	300
Dec	92	250

a) Find the regression equation based on the above data

b) Can you estimate the appropriate quantity likely to be purchased if the price shoot upon Rs 124/kg?

c) Hence or otherwise obtain the coefficient of correlation between the price prevailing and the quantity demanded

Solution:

Let x and y represents price and quantity of a raw material purchased by the company respectively.

x	y	xy	x^2	y^2
96	250	24000	9216	62500
110	200	22000	12100	40000
100	250	25000	10000	62500
90	280	25200	8100	78400
86	300	25800	7396	90000
92	300	27600	8464	90000
112	220	24640	12544	48400
112	220	24640	12544	48400
108	200	21600	11664	40000
116	210	24360	13456	44100
86	300	25800	7396	90000
92	250	23000	8464	62500
1200	**2980**	**293640**	**121344**	**756800**

$$\sum x_i = 1200, \sum y_i = 2980, \sum x_i y_i = 293640, \sum {x_i}^2 = 121344, \sum {y_i}^2 = 756800,$$
$$n = 12$$

a) The regression line of Y on X is given by $Y = a + bX$

Where

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum {x_i}^2 - n\bar{x}^2}$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1200}{12} = 100$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{2980}{12} = 248.33$$

$$b = \frac{293640 - 12 \times 100 \times 248.33}{121344 - 12 \times (100)^2} = -3.24$$

$$a = \bar{y} - b\bar{x} = 248.33 - (-3.24) \times 100$$

$$a = 572.33$$

The regression line of Price/kg on quantity in kg is given by

$$y = 572.33 - 3.24\, x \ldots (1)$$

b) Given $x = 124/kg$ substituting this in (1) we get

$$y = 572.33 - 3.24\,(124)$$

$$y = 170.57$$

If the price is Rs 124/kg, then the appropriate quantity likely to be purchased is approximately 171 kg.

c) To find Correlation coefficient:

$$\rho = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum {x_i}^2 - n\bar{x}^2}\sqrt{\sum {y_i}^2 - n\bar{y}^2}}$$

$$= \frac{293640 - 12 \times 100 \times 248.33}{\sqrt{121344 - 12 \times (100)^2}\sqrt{756800 - 12 \times (248.33)^2}}$$

$$\rho = -0.92$$

Example:

Find the regression analysis for given data. An industry a data for his electricity supplied to his industries and agriculture. It gives the data for the demand for electric motors in a certain region of the country for 6 years. The data is given below

Electricity supply	Demand for electric motors
20	16
25	20
31	24
37	30
42	35
43	37

Solution:

Let x and y represents electric supply and demand for electric motors for an industry respectively.

x	y	xy	x^2	y^2
20	16	320	400	256
25	20	500	625	400
31	24	744	961	576
37	30	1110	1369	900
42	35	1470	1764	1225
43	37	1591	1849	1369
198	**162**	**5735**	**6968**	**4726**

$$\sum x_i = 198, \sum y_i = 162, \sum x_i y_i = 5735, \sum {x_i}^2 = 6968, \sum {y_i}^2 = 4726, n = 6$$

The regression line of Y on X is given by

$$Y = a + bX$$

Where

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum {x_i}^2 - n\bar{x}^2}$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{198}{6} = 33$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{162}{6} = 27$$

$$b = \frac{5735 - 6 \times 33 \times 27}{6968 - 6 \times (33)^2}$$

$$b = 0.9$$

$$a = \bar{y} - b\bar{x} = 27 - (0.9) \times 33$$

$$a = -2.7$$

The regression line of electric supply on demand for electric motors is given by

$$y = -2.7 + 0.9\,x$$

The regression line of X on Y is given by

$$X = c + dY$$

Where

$$d = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum {y_i}^2 - n\bar{y}^2}$$

$$d = \frac{5735 - 6 \times 33 \times 27}{4726 - 6 \times (27)^2}$$

$$d = 1.11$$

$$c = \bar{x} - b\bar{y} = 33 - (1.11) \times 27$$

$$c = 3.03$$

The regression line of demand for electric motors on electric supply is given by

$$x = 3.03 + 1.11y$$

Applications of linear regression in Business:

Linear regression is used in business to predict events, manage product quality and analyze a variety of data types for decision-making.

- ***Trend Line Analysis***

  Linear regression is used in the creation of trend lines, which uses past data to predict future performance or "trends." Usually, trend lines are used in business to show the movement of financial or product attributes over time. Stock prices, oil prices, or product specifications can all be analyzed using trend lines.

- ***Risk Analysis for Investments***

  The capital asset pricing model was developed using linear regression analysis, and a common measure of the volatility of a stock or investment is its beta--which is determined using linear regression. Linear regression and its use is key in assessing the risk associated with most investment vehicles.

- ***Sales or Market Forecasts***

  Multivariate (having more than two variables) linear regression is a sophisticated method for forecasting sales volumes, or market movement to create comprehensive plans for growth. This method is more accurate than trend analysis, as trend analysis only looks at how one variable changes with respect to another, where this method looks at how one variable will change when several other variables are modified.

- ***Total Quality Control***

  Quality control methods make frequent use of linear regression to analyze key product specifications and other measurable parameters of product or organizational quality (such as number of customer complaints over time, etc).

- ***Linear Regression in Human Resources***

  Linear regression methods are also used to predict the demographics and types of future work forces for large companies. This helps the companies to prepare for the needs of the work force through development of good hiring plans and training plans for the existing employees.