ISSN: 2754-6705

Journal of Mathematical & Computer Applications

SCIENTIFIC Research and Community

Research Article

Open Access

A Statistical Study on the Slope of a Line for Data Related to Height-Weight and COVID Infections-Recoveries

Prerna Verma and Soubhik Chakraborty*

Department of Mathematics, Birla Institute of Technology, Mesra, Ranchi-835215, Jharkhand, India

ABSTRACT

The slope of the line reveals changes in y-coordinate with respect to x-coordinate, represented by the equation "$y=mx+c$", known as the equation of a line. The slope m is of significance in several areas including artificial intelligence and regression analysis where it helps in predicting the response by using one known value of the predictor. Our aim here is to predict the average value of y for a given value of x. A related analysis has been discussed in. To begin the analysis, we take data of two dependent sets. The data analysed relate to height-weight and COVID infections-recoveries. We first initialize the value of k which is the range for our discrete uniform probability distribution U[1,2,3…k] as discussed in the paper. The probability variate is used to generate 30 random numbers using the formula u = 1+integral part of (r*k) where r is a continuous U[0,1] variate. This formula gives 30 independent U[1, 2…k] variates; k is then incremented by 30 after each run. We run the loop for 6 times and for every k we have the values of m (slope of the line) and c (constant) which are calculated by the Least Squares Method discussed in. Using the values of m, c and x we estimate the value of y which is referred to as the y-estimate. Every regression equation gives some error; in order to find the error, we subtract the estimated y from the observed y. To see the change, we use the method of plotting. The graphs of k versus m and k versus c are plotted. The experimental results show that

- In case of height and weight data. the values of m and c do not depend on the discrete uniform variate's range k.
- In case of COVID infections and recoveries data, the regression equation depicts a quadratic pattern for both m and c which depend on k.

***Corresponding author**

Soubhik Chakraborty, Department of Mathematics, Birla Institute of Technology, Mears Ranchi-835215, Jharkhand, India.

Received: February 12, 2023; **Accepted:** February 20, 2023, **Published:** February 28, 2023

Keywords: Slope, Probability Distribution and Error, Statistical Analysis.

Introduction

The slope of the line is a very important characteristic of a straight line and along with the y-intercept, it tells us about the nature of a straight line. To study the nature of straight lines we use Uniform Discrete Probability Distribution as discussed in the paper and we observe how the change in parameter of the probability distribution affects the slope of the straight line and its y intercept [1].

We begin our analysis by taking two sets of data, one set for height and weight of 200 people, which is collected from [2], shown in (Table 1) and one set for infected and recovered from COVID in a month, collected from [3], shown in (Table 2).

Data Used

Data for Height and Weight of 200 People

Table 1: Height and weight of 200 people

Sl. No	Weight (in Kg)	Height (in cm)
1	65.78	112.99
2	71.52	136.49
3	69.40	153.03
4	68.22	142.34
5	67.79	144.30
6	68.70	123.30
7	69.80	141.49
8	70.01	136.46
9	67.90	112.37
10	66.78	120.67
11	66.49	127.45
12	67.62	114.14
13	68.30	125.61
14	67.12	122.46
15	68.28	116.09
16	71.09	140.00
17	66.46	129.50
18	68.65	142.97
19	71.23	137.90
20	67.13	124.04
21	67.83	141.28
22	68.88	143.54
23	63.48	97.90
24	68.42	129.50
25	67.63	141.85

26	67.21	129.72
27	70.84	142.42
28	67.49	131.55
29	66.53	108.33
30	65.44	113.89
31	69.52	103.30
32	65.81	120.75
33	67.82	125.79
34	70.60	136.22
35	71.80	140.10
36	69.21	128.75
37	66.80	141.80
38	67.66	121.23
39	67.81	131.35
40	64.05	106.71
41	68.57	124.36
42	65.18	124.86
43	69.66	139.67
44	67.97	137.37
45	65.98	106.45
46	68.67	128.76
47	66.88	145.68
48	67.70	116.82
49	69.82	143.62
50	69.09	134.93
51	69.91	147.02
52	67.33	126.33
53	70.27	125.48
54	69.10	115.71
55	65.38	123.49
56	70.18	147.89
57	70.41	155.90
58	66.54	128.07
59	66.36	119.37
60	67.54	133.81
61	66.50	128.73
62	69.00	137.55
63	68.30	129.76
64	67.01	128.82
65	70.81	135.32
66	68.22	109.61
67	69.06	142.47
68	67.73	132.75
69	67.22	103.53
70	67.37	124.73
71	65.27	129.31
72	70.84	134.02
73	69.92	140.40
74	64.29	102.84
75	68.25	128.52
76	66.36	120.30
77	68.36	138.60
78	65.48	132.96
79	69.72	115.62
80	67.73	122.52
81	68.64	134.63
82	66.78	121.90
83	70.05	155.38
84	66.28	128.94
85	69.20	129.10
86	69.13	139.47
87	67.36	140.89
88	70.09	131.59
89	70.18	121.12
90	68.23	131.51
91	68.13	136.55
92	70.24	141.49
93	71.49	140.61
94	69.20	112.14
95	70.06	133.46
96	70.56	131.80
97	66.29	120.03
98	63.43	123.10
99	66.77	128.14
100	68.89	115.48
101	64.87	102.09
102	67.09	130.35
103	68.35	134.18
104	65.61	98.64
105	67.76	114.56
106	68.02	123.49
107	67.66	123.05
108	66.31	126.48
109	69.44	128.42
110	63.84	127.19
111	67.72	122.06
112	70.05	127.61
113	70.19	131.64
114	65.95	111.90
115	70.01	122.04
116	68.61	128.55
117	68.81	132.68
118	69.76	136.06
119	65.46	115.94
120	68.83	136.90
121	65.80	119.88
122	67.21	109.01
123	69.42	128.27

124	68.94	135.29
125	67.94	106.86
126	65.63	123.29
127	66.50	109.51
128	67.93	119.31
129	68.89	140.24
130	70.24	133.98
131	68.27	132.58
132	71.23	130.70
133	69.10	115.56
134	64.40	123.79
135	71.10	128.14
136	68.22	135.96
137	65.92	116.63
138	67.44	126.82
139	73.90	151.39
140	69.98	130.40
141	69.52	136.21
142	65.18	113.40
143	68.01	125.33
144	68.34	127.58
145	65.18	107.16
146	68.26	116.46
147	68.57	133.84
148	64.50	112.89
149	68.71	130.76
150	68.89	137.76
151	69.54	125.40
152	67.40	138.47
153	66.48	120.82
154	66.01	140.15
155	72.44	136.74
156	64.13	106.11
157	70.98	158.96
158	67.50	108.79
159	72.02	138.78
160	65.31	115.91
161	67.08	146.29
162	64.39	109.88
163	69.37	139.05
164	68.38	119.90
165	65.31	128.31
166	67.14	127.24
167	68.39	115.23
168	66.29	124.80
169	67.19	126.95
170	65.99	111.27
171	69.43	122.61
172	67.97	124.21
173	67.76	124.65
174	65.28	119.52
175	73.83	139.30
176	66.81	104.83
177	66.89	123.04
178	65.74	118.89
179	65.98	121.49
180	66.58	119.25
181	67.11	135.02
182	65.87	116.23
183	66.78	109.17
184	68.74	124.22
185	66.23	141.16
186	65.96	129.15
187	68.58	127.87
188	66.59	120.92
189	66.97	127.65
190	68.08	101.47
191	70.19	144.99
192	65.52	110.95
193	67.46	132.86
194	67.41	146.34
195	69.66	145.59
196	65.80	120.84
197	66.11	115.78
198	68.24	128.30
199	68.02	127.47
200	71.39	127.8

Data for Infected and Recovered Patients from Covid-19 for a Month

Table 2: Infected and Recovered patients from COVID-19 for a month

Sl. No	Infected	Recovered
1	41	2
2	41	6
3	41	7
4	41	7
5	41	12
6	45	12
7	62	16
8	121	21
9	198	25
10	270	25
11	375	25
12	444	28
13	549	31
14	729	34
15	1052	44
16	1423	46

17	2714	49
18	3554	82
19	4586	92
20	5806	118
21	7153	168
22	9074	218
23	11177	295
24	13522	397
25	16678	522
26	19665	633
27	22112	817
28	24953	1218
29	27100	1480
30	29631	1795
31	31728	2222

Computer Program
(Language: C++; system specification CPU)

The following driver code is used for the computational analysis part

Driver Code
Data for Height and Weight of 200 People

```
int valu[30] = {0};
int k = 30;

while(k<200){
for (int i = 0; i <= 30; i++)
{
int ran = rand() ;
int maxr = RAND_MAX;
float r = ran / float(maxr);
double a, b;

a = r * k;
modf(a, &b);
int u = 1 + b;

if (lin_search(valu, i, u))
{
valu[i] = u;
}

else
--i;
}
int mu = sizeof(valu) / sizeof(valu[0]);
quickSort(valu, 0, mu - 1);
int counter = 0;
for (int i = 0; i < 30; i++)
{
int flag = 1;
for (int j = 0; j < 30; j++)
{
if (j == i)
continue;

if (valu[i] == valu[j])
{
flag = 0;
break;
}
}

if (flag)
counter++;
}
int s=0 ;
double x[30] ;
double y[30] ;
for( int z = 0 ; z < 30 ; z++ )
{

s = valu[z]  ;
y[z]=arr[s][0] ;
x[z]=arr[s][1] ;
}
```

The above code randomly generates 30 values using the rand function and these values are used as index to point to the elements of the dataset we have taken [2].

Data for Infected and Recovery of Patients from Covid-19 in a Month

```
int k = 7;

while(k<31){
for (int i = 0; i <= 7; i++)
{
int ran = rand() ;
int maxr = RAND_MAX;
float r = ran / float(maxr);
double a, b;

a = r * k;
modf(a, &b);
int u = 1 + b;

if (lin_search(valu, i, u))
{
valu[i] = u;
}

else
--i;
}

int counter = 0;

for (int i = 0; i < 30; i++)
{
int flag = 1;
for (int j = 0; j < 30; j++)
{
 if (j == i)
continue;

if (valu[i] == valu[j])
{
flag = 0;
break;
```

```
}
}

if (flag)
counter++;
}

int mu = sizeof(valu) / sizeof(valu[0]);
quickSort(valu, 0, mu - 1);

int s=0 ;
double x[7] ;
double y[7] ;


for( int z = 0 ; z < 7 ; z++ )
{

s = valu[z]  ;

y[z]=arr[s][1] ;
x[z]=arr[s][0] ;

}
double sum_x=0, sum_y=0, sum_xx=0, sum_yy=0, sum_xy=0;
double d=0, e=0, f=0 ;
for ( int w = 0 ; w < 7 ; w ++)
{
sum_x+=x[w] ;
sum_y+=y[w];
sum_xx+=(x[w]*x[w]);
sum_yy+=(y[w]*y[w]);
sum_xy+=(x[w]*y[w]);
}
double lambda ;
lambda=((sum_xx*sum_y) - (sum_x*sum_xy))/((7*sum_xx)-
(sum_x*sum_x)) ;
cout/*<<"The value of constant is "*/<<lambda<<endl ;
double gamma ;
gamma=((7*sum_xy)-(sum_x*sum_y))/((7*sum_xx)-
(sum_x*sum_x)) ;
cout/*<<"The value of slope is "*/<<gamma<<endl ;


int calculated_Y ;
int sum_calculated_Y=0 ;

for(int i = 0 ; i <7 ; i++){
calculated_Y=gamma*x[i]+lambda;
cout<<"The value of x used is "<<x[i]<<endl ;
cout<<"The calculated Y is "<<calculated_Y<<endl ;
cout<<"Observed Y is "<< y[i]<<endl ;
}
k+=7;
}
```

The above mentioned code randomly generates 7 values using the rand function and these values are used as index to point to the elements of the dataset we have taken [3].

Statistical Analysis

The index found using the above driver codes **D1** & **D2** is used to select random samples from the sample space and is used to find the straight line for each run. The value of slope (m) and constant (c) is found using the property of least squares discussed in [4].

Least Squares Method

The least squares method, as discussed in, is a statistical procedure to find the best fit for a set of data points by minimising the sum of the offsets or residuals of points from the plotted curve [4]. The formula of the above definitions is given by

Equation 1:

$$y=c+mx+e$$

Here c is y-intercept of straight line and m is the slope of straight line and e is the error or residual which is minimized by this method

$$Constant\ (c) = \frac{\Sigma X^2 \Sigma Y - \Sigma X \Sigma XY}{n\Sigma X^2 - (\Sigma X)^2}$$

$$Slope\ (m) = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

Using **Equation 1** we first analyse the data from Table 1 and store our findings in table 3.

Table 3: Values of c and m for different trials

Trial Number	Value of c	Value of m
1	58.7554	0.0730783
2	58.3416	0.0776913
3	59.8408	0.0682531
4	51.8077	0.126691
5	59.1357	0.0700093
6	61.8425	0.0483243

For each run, the value of k is incremented by 30 and x and y intercepts of the straight line are stored.

It is clear from table 3 that m and c are not changing markedly except in trial 4 in which some change is noticed.

Similarly, using **equation 1** we analyse the data from table 2 and store the data in table 4.

In this case too we increment the value of k by 7 and we observe the changes in value of c and value of m. Figures 1 and 2 depict a quadratic pattern for both c and m as a function of k.

Table 4: Values of c and m for different trials

Trial Number	Value of c	Value of m
7	2.57029	0.160735
14	14.8739	0.0314869
21	17.2678	0.0166628
28	-31.3181	0.0522608

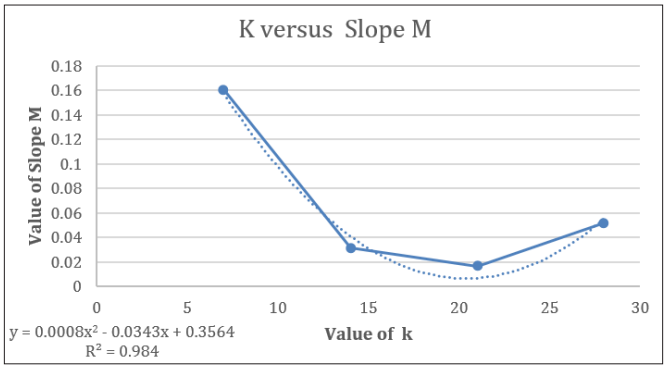

Figure 1: k versus slope m (written as K and M in the graph respectively; in the regression equation y stands for m and x stands for k)

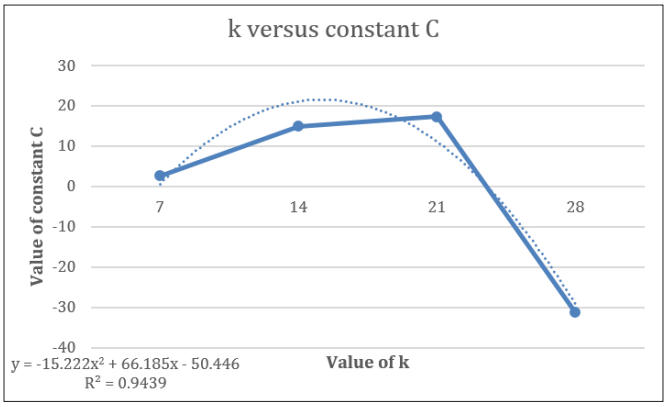

Figure 2: k versus Constant C (in the regression equation, y stands for c written as C in the graph and x stands for k)

Prediction from Data

Using above statistical analysis we try to predict average weight and number of recoveries. We compare these predictions to the actual data which is taken from that we have to check the accuracy of data obtained [2,3]. Table 5 shows the observed and estimated value of y for Height-Weight table and Table 6 shows the observed and estimated value of y for Infected- Recovery Data.

Table 5: Observed and calculated y for height-weight data

Sl No.	Observed Value of y	Estimated Value of y
1	7	37
2	25	23
3	28	20
4	82	112
5	295	436
6	633	797
7	1218	1022

Using the table 5, we calculate the error and the mean error is 0.00001 and mean error percentage is 4.92×10^{-6}% [5].

Table 6: Observed and estimated y for Infected-Recovery data

Sl No.	Observed Value of y	Estimated Value of y
1	71.52	68.5586
2	67.79	69.4069
3	68.65	69.2625
4	68.88	69.3244
5	63.48	64.3671
6	65.44	66.1039
7	69.52	64.9537
8	67.82	67.3964
9	69.66	68.904
10	67.97	68.6542
11	69.09	68.3892
12	69.91	69.7024
13	70.18	69.7969
14	67.01	67.7255
15	67.73	68.1524
16	64.29	64.9037
17	70.24	69.1017
18	71.49	69.0061
19	70.56	68.0492
20	63.43	67.1043
21	66.77	67.6517
22	67.66	67.0988
23	65.46	66.3266
24	68.83	68.6032
25	64.4	67.1792
26	65.92	66.4015
27	65.31	67.6702
28	66.81	65.1198
29	66.59	66.8675
30	66.97	67.5985

Using the table 6, we calculate the error and the mean error is 41 and mean error percentage is 1.79% discussed in [5].

Discussion

We notice a that both values of both m and c do not change much with changing values of k. Even on increasing the value of k we do not see any behavioural changes.

But, in case of the second data which is much more interesting, we observe a quadratic relation between k and m and c. There is an interesting observation that the value of m and c attains a maximum value and then there is a sharp decline in the values. This could be due to the fact that both recovery and infection values hugely increases in a very short period.

Using the plots, we are also able to predict regression curves as explained in for all of these data which can be used to predict the relation between the response variable and the predictor for a fixed period [6].

Results

Using random variate from discrete probability distribution discussed in the paper we are able to observe the average change in the value of x and y intercept of a straight line and we are also able to predict a regression curve found in that can be used to find the average value of m and c [1,6]. We also are able to check the accuracy of the predicted dependent variable with respect to the observed dependent variable. We also found the error percentage which was in an acceptable range which further strengthens the fact that data obtained were feasible [5].

Conclusion

We conclude,

- In case of height and weight data. the values of m and c do not depend on the discrete uniform variate's range k.
- In case of COVID infections and recoveries data, the regression equation depicts a quadratic pattern for both m and c which depend on k.

In summary this paper should tell its reader the fact that when x and y intercepts of any straight line has its values derived from uniform discrete probability distribution then there is a chance that there will be repeated values of m and c for different values of k [3]. One of the major goals was to establish a relation between the changing parameter and M & C and having achieved this, we close this paper [7].

Remark: The reader is encouraged to answer questions such as "What should be the average wight of a person with a given height?" by fitting a regression line of weight (y) on height (x) using the data given in table 1. Regression line of y on x is given as $y - y_{mean} = (r\sigma_y/\sigma_x)(x - x_{mean})$ where r is the correlation coefficient between x and y = covariance(x,y)/SD(x)SD(y) where covariance(x,y)=$\Sigma xy/n - x_{mean}y_{mean}$ and $SD(x) = \sigma_x = +\sqrt{\{\Sigma x^2/n - (x_{mean})^2\}}$; $x_{mean}=\Sigma x/n$;
$SD(y) = \sigma_y = +\sqrt{\{\Sigma y^2/n - (y_{mean})^2\}}$; $y_{mean}=\Sigma y/n$. Similarly by fitting regression line of x on y, given by x -xmean= (rσx/σy)(y-ymean), one can answer questions such as "what should be the average height of a person for a given weight?" which is again left as an exercise to the reader.

Ethical Declaration

The authors declare that this research did not receive any funding and that they do not have any conflict of interest.

References

1. Sheldon Ross (2018) A First Course in Probability, Tenth edition, Pearson, Boston, USA https://www.amazon.in/First-Course-Probability-Sheldon-Ross/dp/0134753119.
2. http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights accessed on 15th May 2022.
3. https://figshare.com/articles/journal_contribution/Data-based_analysis_modelling_and_forecasting_of_the_COVID-19_outbreak/12055098 accessed on 15th May 2022.
4. https://www.oreilly.com/library/view/budgeting-basics and/9780470389683/9780470389683_least-squares_method.html (accessed on 20th May 2022)
5. https://www.webassign.net/question_assets/unccolphysmechl1/measurements/manual.html (accessed on 30th May 2022)
6. Norman Draper, Harry Smith (1998) Applied Regression Analysis, John Wiley & Sons https://onlinelibrary.wiley.com/doi/book/10.1002/9781118625590.
7. Subhomoy Haldar, Soubhik Chakraborty (2021) On the Probability of Real Roots in a Quadratic Equation with Coefficients as i.i.d $\boldsymbol{\mathcal{U}}(-\boldsymbol{\theta}, \boldsymbol{\theta})$ Variates, Journal of Applied Mathematics and Computation 5: 48-55.