INSTITUTE OF RESEARCH ADVANCES

IRA International Journal of Education and Multidisciplinary Studies
ISSN 2455–2526 Vol. 03 Issue 02 (May, 2016)
Paper DOI: https://dx.doi.org/10.21013/jems.v3.n2.p7

Correlation and Regression using VASSARSTATS

Suma A P[1], K P Suresh[2]
[1]Research scholar, Jain University, Jayanagar 9th Block, Bengaluru, India
[2]Senior Scientist, National Institute of Veterinary Epidemiology and Disease Informatics (NIVEDI), India.

ABSTRACT

In a bivariate or a multivariate data, to understand the association between the variables ***Correlation*** *is the best tool. It gives the degree of relationship between the variables.* ***Regression*** *gives the exact linear relationship between the variables. This article gives details of capabilities of Vassarstats Correlation and Regression and procedure to calculate Correlation coefficient and Regression coefficients with examples. Vassarstats Correlation and Regression can perform Linear Correlation and Regression, Intercorrelations, Multiple Correlation and Regression, Partial Correlation, 0.95 and 0.99 Confidence intervals for population correlation coefficient, Estimating the Population Value of rho, Significance of value of r, Significance of difference between two correlation coefficients, Significance of difference between sample correlation coefficient and hypothetical value of population Correlation coefficient, Rank Order Correlation, Correlation coefficient for a 2*2 contingency table, Point biserial correlation coefficient, Correlation for unordered pairs, and then Simple Logistic Regression.*

Key words: Correlation, Regression, Multiple Correlation, Partial Correlation, Rank Order Correlation, Confidence Interval, Significance of Correlation coefficient.

Introduction

Correlation is a measure which can detect the extent to which two or more variables vary in the same direction or in the opposite direction. Suppose we have two variables X and Y and we are interested in understanding whether there is any association between them, Correlation gives the degree of relationship between the variables. If the variables vary together in the same direction, the Correlation is said to be **Positive**. On the other hand, an increase in one variable results in decrease in the other variable and vice versa, Correlation is said to be **Negative**. The degree of linear relationship is measured by Coefficient of Correlation(r).

Interpretation of Coefficient of Correlation(r):

Table 1 (See Tables Section at the end of the paper**)**

The strength of linear correlation increases as Coefficient of Correlation(r) goes close to +1 or -1 from 0.

There are three types of correlation

- Simple correlation: This is the correlation between two variables.
- Multiple correlation: This is the nothing but the correlation between more than two variables.
- Partial correlation: This is the correlation between any two variables, controlling the effect of other variables.

Regression gives the exact linear relationship between X and Y. The relationship can be written as Y=a + bX, where 'a' is the intercept and 'b'is the slope. So, for a given value of X, Y can be predicted with the help of Regression equation.

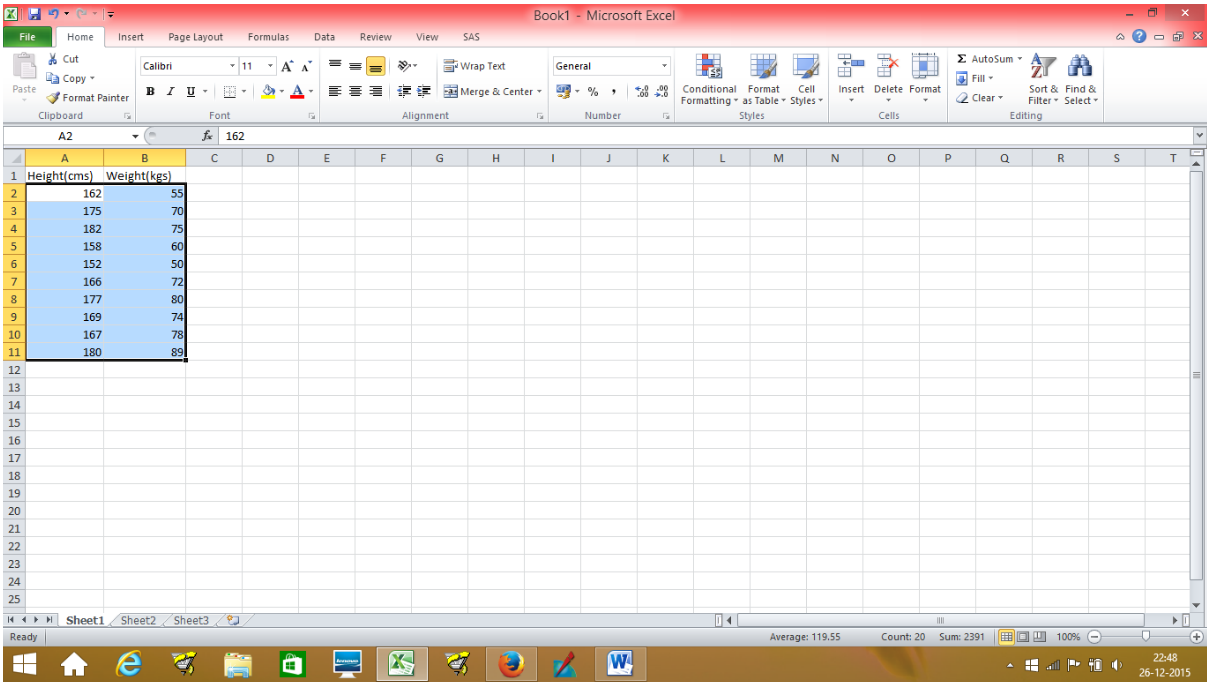
This article gives the method to compute Coefficient of Correlation(r) and Regression using Vassarstats.

Linear Correlation and Regression

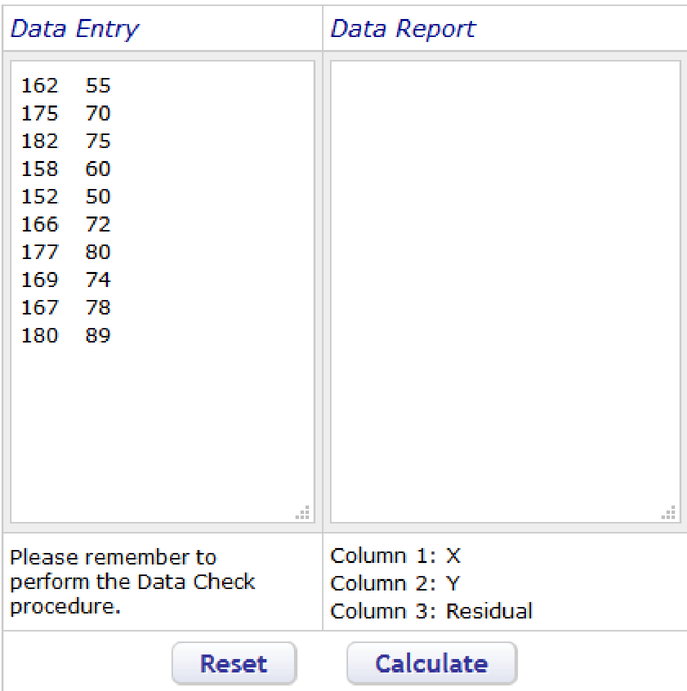
Go to *Vassarstats*, then to *Correlation and Regression* and then to *Linear Correlation and Regression*, then to *Data entry version* if the data is in excel sheet, otherwise go to *Direct entry version.*

Data in excel sheet: If the data is in excel sheet, select the data and copy it. Then the data is to be pasted in the *Data entry box*. Template1 shows the data which is copied. Template2 shows the data which is entered in *Data entry box*.

Template1



Template2



It is of utmost importance that the cursor is next to the last data i.e. 89. Then click ***Calculate***.

The results are shown in Template3.

Template3:

Data Summary

$\sum X =$	1688	$\sum X^2 =$	285796
$\sum Y =$	703	$\sum Y^2 =$	50715
$\sum XY =$	119554		

	X	Y
N	10	
Mean	168.8	70.3
Variance	95.7333	143.7889
Std.Dev.	9.7843	11.9912
Std.Err.	3.0941	3.792

r	r^2	Slope	Y Intercept	Std. Err. of Estimate
0.8406	0.7066	1.030176	-103.593779	6.8894

t	df	P		
4.39	8		one-tailed	0.001159
			two-tailed	0.002318

0.95 and 0.99 Confidence Intervals for rho

	Lower Limit	Upper Limit
0.95	0.449	0.961
0.99	0.245	0.975

0.95 and 0.99 Confidence Intervals for the Slope of the Regression

	Lower Limit	Upper Limit
0.95	0.488	1.5724
0.99	0.2416	1.8188

It can be seen that the 1st column is X and the second column is Y. The Correlation coefficient is 0.8406. That means, X and Y are positively correlated. Coefficient of determination is r^2 = 0.7066. That means, 71% of the variation in Y is explained by X.

The Regression line can be written as Y= -103.593779 + 1.030176X. To estimate the weight of a person with height 185cms, substitute X=185 in the Regression line. It can be seen that Y = 86.989kgs.

Vassarstats also gives the confidence interval for population correlation coefficient(rho) and slope of the regression as shown above.

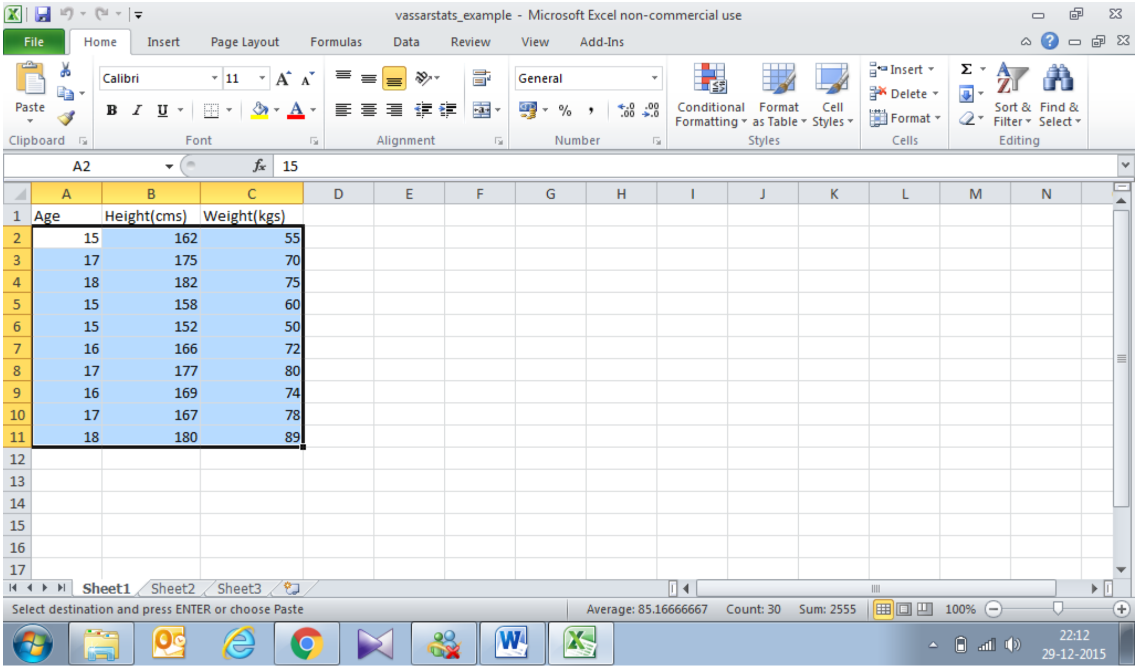
Direct entry method: Data can also be entered directly in the *Data Entry* column. First enter value of X, then a single space and then the value of Y. Then press *Enter* key. Repeat the procedure until the last value of Y is entered. The cursor should be next to the last value of Y. Then click *Calculate*.

Intercorrelations

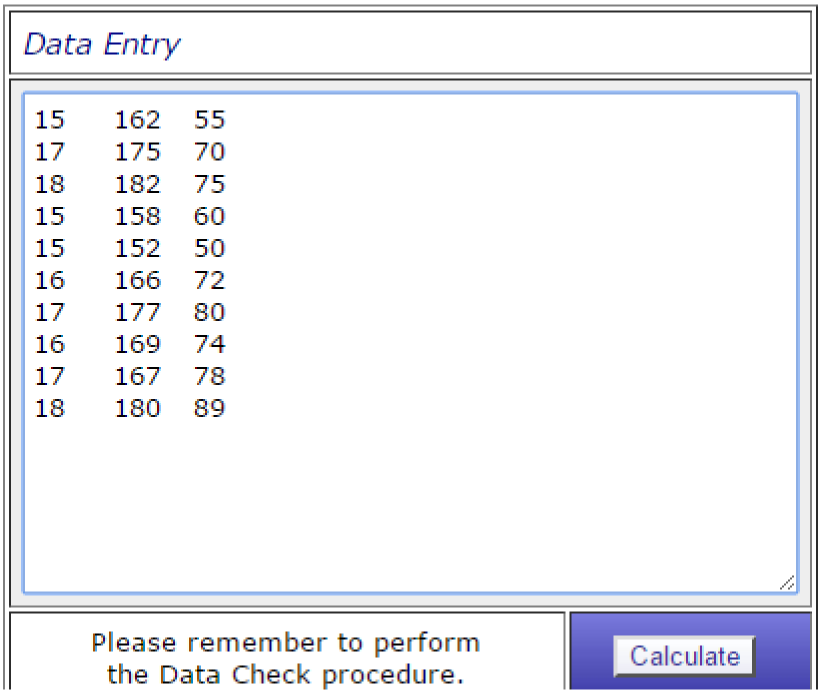
Suppose we have three or more variables and we are interested in the Correlation between any two variables, In *Vassarstats*, go to *Correlation and Regression* and then to *Matrix Intercorrelations.*

Data in excel sheet: If the data is in excel sheet, select*version1*. Then go to excel sheet, select the data and copy it. Then the data is to be pasted in the *Data entry box*. Template3 shows the data which is copied. Template4 shows the data which is entered in the *Data entry box*.

Template3



Template 4



Be sure that the cursor is next to last entry 89.

Now click *Calculate.* The result is shown Template5.

Template 5

VassarStats: Correlation Matrix
Number of variables = 3
Observations per variable =10

	V1	V2	V3
V1	1	0.927	0.867
V2	0.927	1	0.841
V3	0.867	0.841	1

Direct entry method: Data can also be entered directly in the *Data Entry* column. For this, click *version 2*. Enter the number of observations in each variable. A table appears. First enter value of first variable V1, then press tab key, enter the value. Proceed till the last value is entered. Follow the same procedure for the other variables to be entered. Then click *Calculate*. In this method, maximum of five variables can be entered. Template 6 shows the data entered.

Template 6

Data Entry:

	Variables				
count	A	B	C	D	E
1	15	162	55		
2	17	175	70		
3	18	182	75		
4	15	158	60		
5	15	152	50		
6	16	166	72		
7	17	177	80		
8	16	169	74		
9	17	167	78		
10	18	180	89		

Reload Reset Calculate

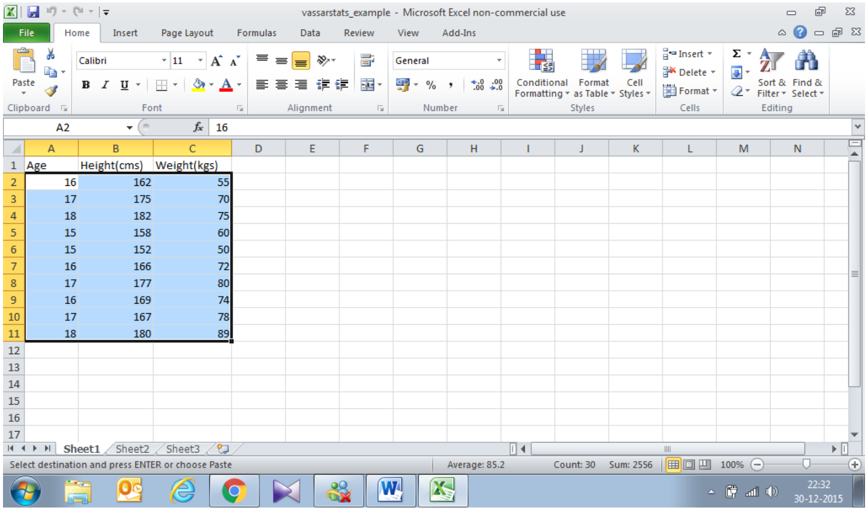
Multiple Correlation and Regression

If we have severalindependent variables $X_1, X_2, ...X_n$to which a dependent variable Y is related simultaneously, multiple correlation coefficient can be computed.

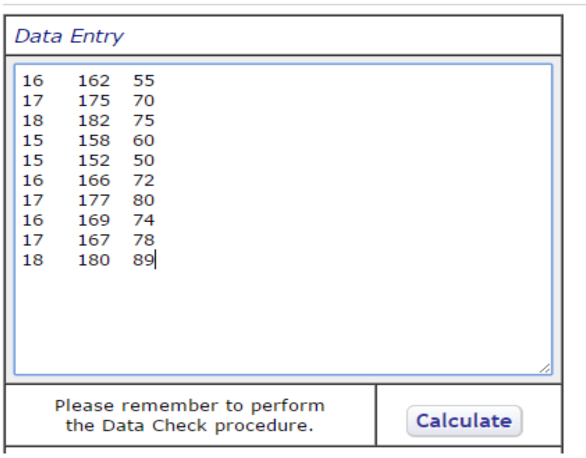
To compute Multiple correlation coefficient and to determine multiple Regression line, In Vassarstats, go to *Correlation and Regression* then go to*Multiple Regression* and then to *Basic Multiple Regression.*This procedure requires data to be in excel sheet.

Data in excel sheet: Go to excel sheet, select the data and copy it. Then paste the data in the *Data entry box*. Template 7 shows the data which is copied. Template 8 shows the data entered in *Data entry box*. The last column should be the dependent variable Y. In this example, it is *weight.*

Template 7



Template 8



It is important that the cursor is next to the last entry 89. Then click C*alculate*. The result is shown in Template 9.

Template 9

Correlation Matrix

	X1	X2	Y
X1	1	0.936	0.811
X2	0.936	1	0.841
Y	0.811	0.841	1

Regression Coefficients:

The multiple regression equation is of the general form

$Y = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k$

where **a** is a starting-point constant analogous to the intercept in a simple two-variable regression, and $\mathbf{b_1}$, $\mathbf{b_2}$, etc., are the unstandardized regression weights for X_1, X_2, etc., each analogous to the slope in a simple two-variable regression. In the present analysis, **a** = -101.5769 and the values of **b** are as indicated below. The values listed as **B** are the standardized regression weights.

	b	B	B x r_{xy}
X1	2.1538	0.194	0.1573
X2	0.8077	0.659	0.554
Multiple R^2 = 0.7113			
Adjusted Multiple R^2 = 0.6288			

Standard Error of Multiple Estimate	6.4434

Here, column 'b' is important to us.

The Regression equation can be written as $Y = -101.5769 + 2.1538\ X_1 + 0.8077\ X_2$.

Partial Correlation

Suppose we have severalindependent variables X_1, X_2, ...$X_n$and a dependent variable Y and we are interested in the correlation between Y and one X, controlling the effect of other independent variables, the Correlation is said to be Partial Correlation.

Partial Correlation for four intercorrelated variables

By taking the examplein Template 7 and considering the three variables as X, Y, Z respectively, Partial Correlation coefficient can be computed in Vassarstats as follows.

First of all Correlation between XY, YZ and XZ should be computed as explained in Linear Correlation and Regression section. Make a note of Correlation coefficient between XY, YZ and XZ. For the above example in Age(X), Height(y) and Weight (Z), Correlation coefficient between XY, YZ and XZ are 0.9357, 0.8406 and 0.8107 respectively. Then click *For three intercorrelated variables*under *Partial Correlation***.**

Enter the number of observations in each variable under N. Then enter Correlation coefficients between XY, YZ and XZ in the respective boxes. Then click *Calculate*.

Template 10 shows the entries.

Template 10

[Optional] N = 10

Original Correlations

	r	r^2
XY	0.9357	---
XZ	0.8406	---
YZ	0.8107	---

Reset Calculate

Template 11 below shows the result.

	r	r^2	t	P
XY.Z	0.802	0.643	3.55	0.0093
XZ.Y	0.397	0.158	1.14	0.2918
YZ.X	0.126	0.016	0.34	0.7438

P values are non-directional (two-tailed)

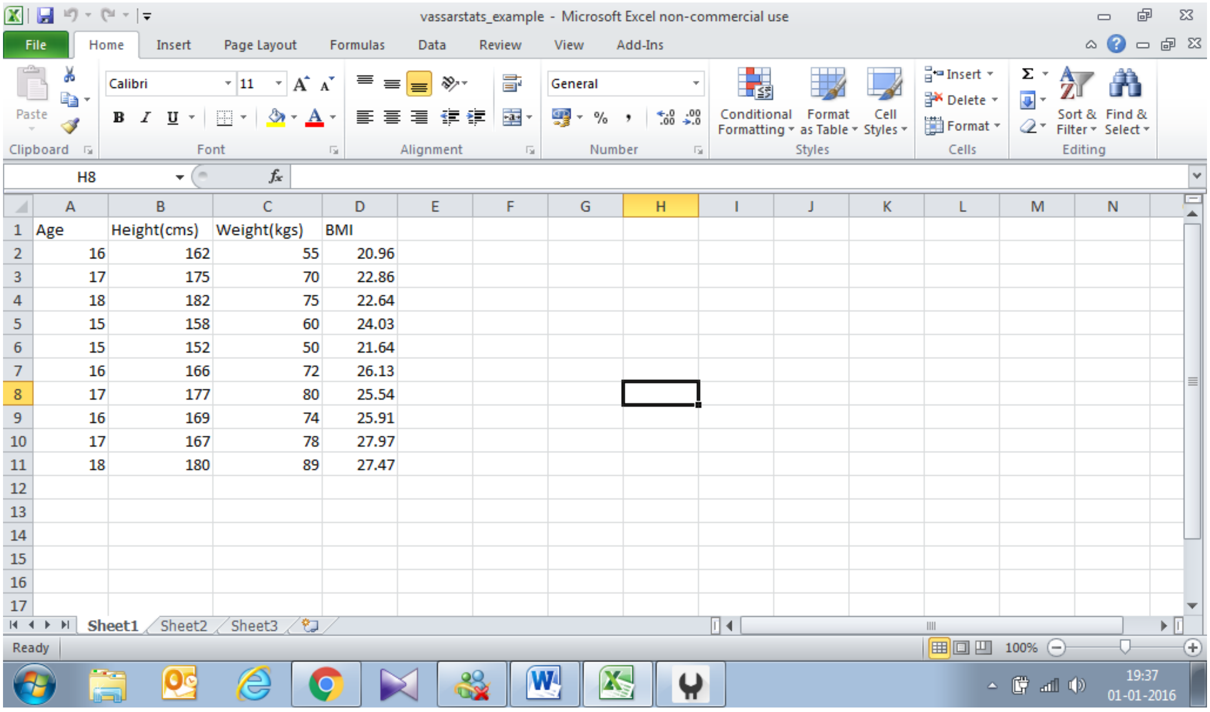
The first column gives the Partial correlation coefficients.

Partial Correlation for four intercorrelated variables:

Suppose we have four variables Age (W), Height(X), Weight(Y) and BMI (Z), and we are interested in Partial correlation between the variables, controlling the effect of one variable in the First order Partial Correlation and controlling the effect of two variables in the Second order Partial Correlation, first of all compute Correlation coefficients between WX, WY, WZ, XY, XZ and YZ as explained under the section *Linear Correlation and Regression*.

Consider the data in the excel sheet in Template 12.

Template 12



Age	Height(cms)	Weight(kgs)	BMI
16	162	55	20.96
17	175	70	22.86
18	182	75	22.64
15	158	60	24.03
15	152	50	21.64
16	166	72	26.13
17	177	80	25.54
16	169	74	25.91
17	167	78	27.97
18	180	89	27.47

The Correlation coefficient between

WX = 0.9357

WY = 0.8107

WZ = 0.3515

XY = 0.8406

XZ = 0.3429

YZ = 0.7936

Now go to *Partial Correlations* and then go to *for four intercorrelated variables.* Enter the above linear Correlations in 'r' column and click *Calculate*.

Template 13 below shows the result.

Template 13

[Optional] N =

Original Correlations

	r	r^2	t	P
WX	0.9357	0.876	---	---
WY	0.8107	0.657	---	---
WZ	0.3515	0.124	---	---
XY	0.8406	0.707	---	---
XZ	0.3429	0.118	---	---
YZ	0.7936	0.63	---	---

Reset Calculate

You get first and second order Partial Correlations. The result is shown below.

1st Order Partial Correlations

	r	r^2	t	P
WX.Y	0.802	0.643	---	---
WX.Z	0.927	0.859	---	---
WY.X	0.126	0.016	---	---
WY.Z	0.934	0.872	---	---
WZ.X	0.092	0.008	---	---
WZ.Y	-0.819	0.671	---	---
XY.W	0.397	0.158	---	---
XY.Z	0.995	0.99	---	---
XZ.W	0.042	0.002	---	---
XZ.Y	-0.984	0.968	---	---
YZ.W	0.928	0.861	---	---
YZ.X	0.993	0.986	---	---

2nd Order Partial Correlations

	r	r^2	t	P
WX.YZ	-0.038	0.001	---	---
WY.XZ	0.295	0.087	---	---
WZ.XY	-0.283	0.08	---	---
XY.WZ	0.962	0.925	---	---
XZ.WY	-0.955	0.912	---	---
YZ.WX	0.994	0.988	---	---

If you want statistic t and p value, then enter the number of observations in each variable against N. In this example N=10.

0.95 and 0.99 Confidence intervals for population correlation coefficient rho

Vassarstats can give you 0.95 and 0.99 confidence intervals for rho directly. For this, InVassarstats, go to *Correlation and Regression*, go to 0.95 *and 0.99 Confidence intervals for r,* enter the value of**r** and **n** in the respective place and then click *calculate*.

Template 14 below shows r and n entered.

r =	0.8406	Reset
n =	10	Calculate

The Result is shown in Template 15.

Template 15

0.95 and 0.99 Confidence Intervals of rho

	Lower Limit	Upper Limit
0.95	0.449	0.961
0.99	0.245	0.975

Estimating the Population Value of rho on the Basis of Several Observed Sample values of r And Test for the Heterogeneity of several Values of r

Population Correlation coefficient *rho* can be estimated with the help of several samples drawn from the same population. Also, test for heterogeneity of several samples of values can be done in Vassarstats.

For this, In Vassarstats, go to *Correlation and Regression, go to Estimating the Population Value of rho on the Basis of Several Observed Sample values of r*, enter the values of n and r and then click *calculate*. Note that, minimum of two values and maximum of twelve values of **r** can be entered. Template 16 shows the data entry.

Template 16

Data Entry

	n	r
1	10	0.85
2	12	0.77
3	8	0.45
4	10	0.67
5	11	0.77
6	12	0.56
7		
8		
9		
10		
11		
12		

Reset

Calculate

The result is shown in Template 17.

It can be concluded that several values of **r** are heterogeneous if Chi Square is significant.

Template 17

Chi-Square	df	P
2.69	5	0.7476474631790
Estimated rho	Estimated Confidence Intervals	
0.711	Lower Limit	Upper Limit
.95 CI	0.534	0.827
.99 CI	0.465	0.854

Significance of value of r

To find out whether r is significant or not, In Vassarstats, go to *Correlation and Regression*, go to *Thesignificance of an observed value of r,* enter the values of **N** and **r** and then click *Calculate*.

N =	10	r =	0.8406
Reset	Calculate		

The following result is obtained.

t	df
4.389	8

Probability	
directional	0.0011595
non-directional	0.002319

A significant value of probability indicates that **r** is significant. Directional is one sided probability and non-directional is two sided probability.

Significance of difference between two correlation coefficients

Suppose we are interested in testing whether the difference between Correlation coefficients of two independent samples is significant, In Vassarstats, go to *Correlation and Regression*, go to Significance *of difference between two correlation coefficients*and then enter the values of **r** and **n** in the respective places as shown below and then click *Calculate*.

As an example, if the Correlation coefficient between Marks scored by students of two sections in Computer Applications and Statistics are respectively, 0.7896 and 0.8654, then

Sample A		Sample B		
r_a =	0.7896	r_b =	0.8654	Reset
n_a =	10	n_b =	12	Calculate

The p value is as follows.

P	one-tailed	0.3156
	two-tailed	0.6312

A significant p value indicates that the difference between Correlation coefficients is significant.

Significance of difference between sample correlation coefficient and hypothetical value of population Correlation coefficient

To test whether there is any significant difference between the sample Correlation coefficient **r** and hypothetical value of correlation coefficient of the population rho from which the sample is drawn, In Vassarstats, go to *Correlation and Regression*, go to *An observed value of r and hypothetical value of rho* and enter values of r, n and rho in the respective places and then click *Calculate*.

Observed for Sample		Hypothetical for Population		
r =	0.8406	rho =	0.95	Reset
n =	10			Calculate

The result is as below.

P	one-tailed	0.053699
	two-tailed	0.107398

A significant p value indicates that there is significant difference between sample Correlation coefficient and population correlation coefficient.

Rank Order Correlation

If the data is qualitative in nature, Spearman's Coefficient of Rank Correlation is used to compute Correlation coefficient.

Ranked data**:** If the data is already ranked, In Vassarstats, go to *Correlation and Regression, then to Rank Order Correlation*, enter the value of **n**, enter the Ranks for

X and Y in the data entry field and then click *Calculate from ranks*. Shown in template 18.

Template 18

Data Entry

	Ranks for		Raw Data for	
pairs	X	Y	X	Y
1	5	3		
2	4	2		
3	1	4		
4	3	5		
5	8	6		
6	10	9		
7	9	7		
8	6	8		
9	2	1		
10	7	10		
Reset	Calculate from Ranks		Calculate from Raw Data	

Data Import

Import Raw Data

See to that the cursor is next to the last entry 10.

The result is as below.

n	r_s	t	df
10	0.7333	3.05	8

P		
	one-tailed	0.0079105
	two-tailed	0.015821

Raw data**:** If the data is not ranked, it can be entered directly in the *Raw data for X and Y columns* or can be imported from the excel sheet and can be pasted in the *Data Import* field.

Direct entry of data**:** Enter the values of X and Y as shown Template 19. Then click *Calculate fromdata.*

Template 19

Data Entry

	Ranks for		Raw Data for	
pairs	X	Y	X	Y
1			56	67
2			78	80
3			75	70
4			75	70
5			86	88
6			77	78
7			79	76
8			89	95
9			85	80
10			80	76
Reset	Calculate from Ranks		Calculate from Raw Data	

Data Import

Import Raw Data

The result is shown in Template 20

Template 20

Data Entry

	Ranks for		Raw Data for	
pairs	X	Y	X	Y
1	1	1	56	67
2	5	7.5	78	80
3	2.5	2.5	75	70
4	2.5	2.5	75	70
5	9	9	86	88
6	4	6	77	78
7	6	4.5	79	76
8	10	10	89	95
9	8	7.5	85	80
10	7	4.5	80	76
Reset	Calculate from Ranks		Calculate from Raw Data	

Data Import

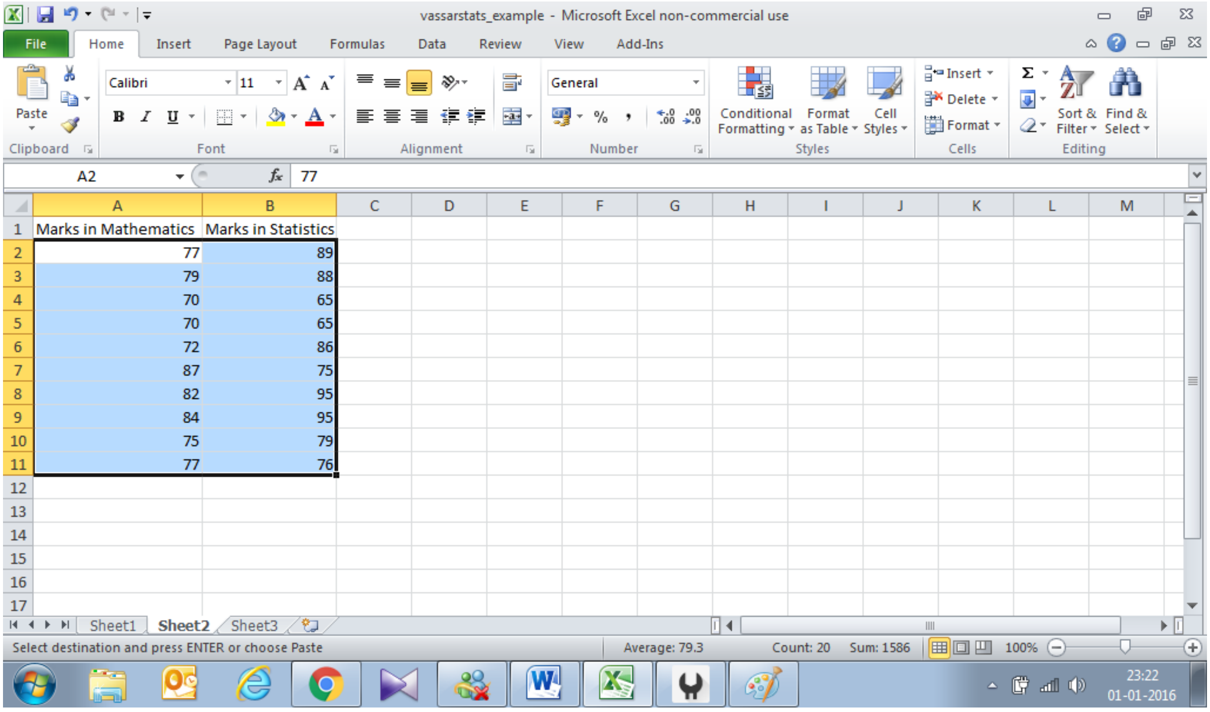
Import Raw Data

n	r_s	t	df
10	0.8835	5.33	8

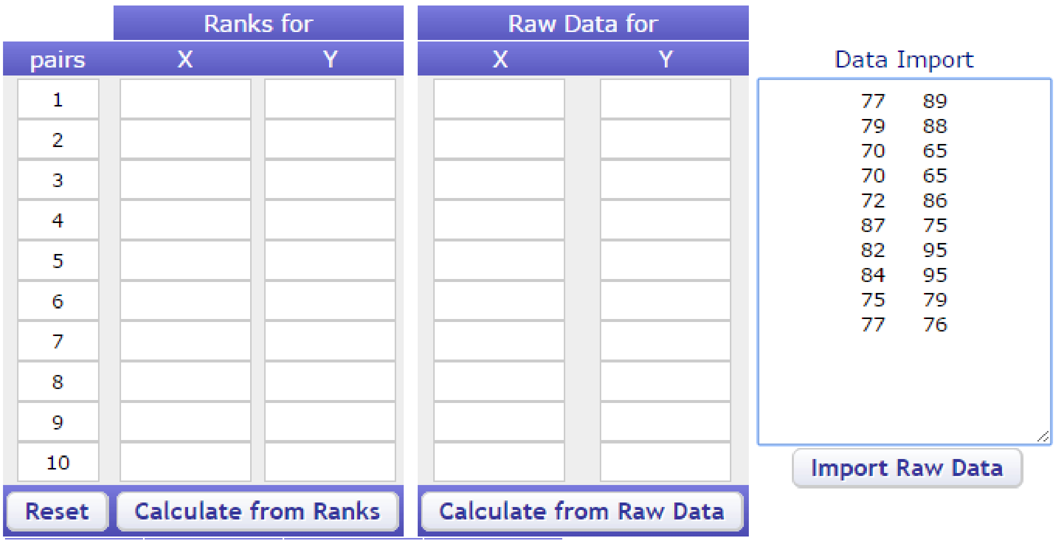
P		
	one-tailed	0.0003515
	two-tailed	0.000703

Data in excel sheet*:* If the data is in excel sheet, copy the data and paste it in the *Data Import* field. Then click *Import Raw Data*. Template 21 shows the data in excel sheet which is copied.

Template 21



Paste the data in *Data Import* field as shown below.



Now click *Import Raw Data* and click *Calculate from data.*

Data Entry

	Ranks for		Raw Data for	
pairs	X	Y	X	Y
1			77	89
2			79	88
3			70	65
4			70	65
5			72	86
6			87	75
7			82	95
8			84	95
9			75	79
10			77	76
Reset	Calculate from Ranks		Calculate from Raw Data	

Data Import

77	89
79	88
70	65
70	65
72	86
87	75
82	95
84	95
75	79
77	76

Import Raw Data

The result is shown below.

Data Entry

	Ranks for		Raw Data for	
pairs	X	Y	X	Y
1	5.5	8	77	89
2	7	7	79	88
3	1.5	1.5	70	65
4	1.5	1.5	70	65
5	3	6	72	86
6	10	3	87	75
7	8	9.5	82	95
8	9	9.5	84	95
9	4	5	75	79
10	5.5	4	77	76
Reset	Calculate from Ranks		Calculate from Raw Data	

Data Import

77	89
79	88
70	65
70	65
72	86
87	75
82	95
84	95
75	79
77	76

Import Raw Data

n	r_s	t	df
10	0.5706	1.96	8

P		
	one-tailed	0.042829
	two-tailed	0.085658

For a 2*2 contingency table

Vassarstats computes Phi Coefficient of Association, Chi-Square Test of Association and Fisher Exact Probability Test for a 2*2 contingency table.

Suppose we have two categorical variables SMOKING(X) and LITERACY(Y), each with two classes like SMOKERS(X=1), NON SMOKERS(X=0) and LITERATES(Y=1), ILLITERATES(Y=0)respectively and we want to test whether there is any association between SMOKING and LITERACY, In Vassarstats, go to *Correlation and Regression*, then go to *Phi Coefficient of Association.* The data should be entered in the *Data Entry* field as shown below.

Data Entry

		X		
		0	1	Totals
Y	1	25	36	
	0	45	50	
Totals				

Calculate Reset

Now click *Calculate*. The result is shown below.

	Chi-Square	
Phi	Yates	Pearson
+0.06	0.38	0.61
P	0.537603	0.434788

Chi-square is calculated only if all expected cell frequencies are equal to or greater than 5. The Yates value is corrected for continuity; the Pearson value is not. Both probability estimates are non-directional.

Fisher Exact Probability Test:

P	one-tailed	0.2688063980598922
	two-tailed	0.5100082737968189

Point biserial correlation coefficient

Point biserial correlation is the correlation between a dichotomous variable and a non-dichotomous variable. The dichotomous variable(X) is coded as 0 and 1 according to absence or presence of an event. The values of Y variable correspond to X=0 and X=1.

As an example, Heights(Y) of males(X=0) and females(X=1) belonging to ages between 20 and 30 years.

Direct entry method**:** Data can be entered directly in the *data entry* field as shown below.

See to that the cursor is next to the last entry. Then click *Calculate*.

Data Entry

	Items Coded as	
	X=0	X=1
Values of Y	167 166 172 182 185 177 162 160 178 180	152 155 157 165 169 177 168 172 170 156
	Reset	Calculate

The result is as follows.

Data Summary	X=0	X=1	Total
n	10	10	20
ΣY	1729	1641	3370
ΣY^2	299635	269937	569572
SS_Y	690.9	648.9	1727
$mean_Y$	172.9	164.1	168.5

r_{pb}	t	df
-0.47	-2.28	18

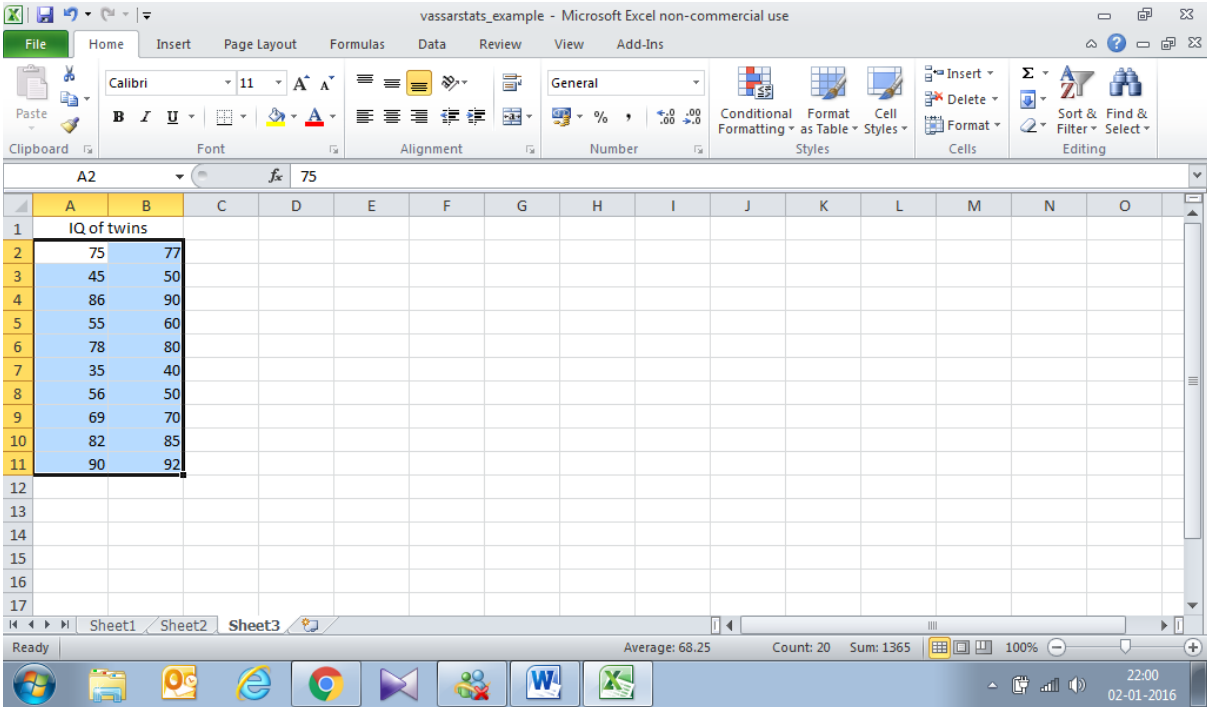
P		
P	one-tailed	0.0175075
	two-tailed	0.035015

Data in excel sheet: If the data is in excel sheet, copy the data corresponding to X=0 and then paste in the respective column. Then do the same procedure for X =1. Then click *calculate*.

Correlation for unordered pairs

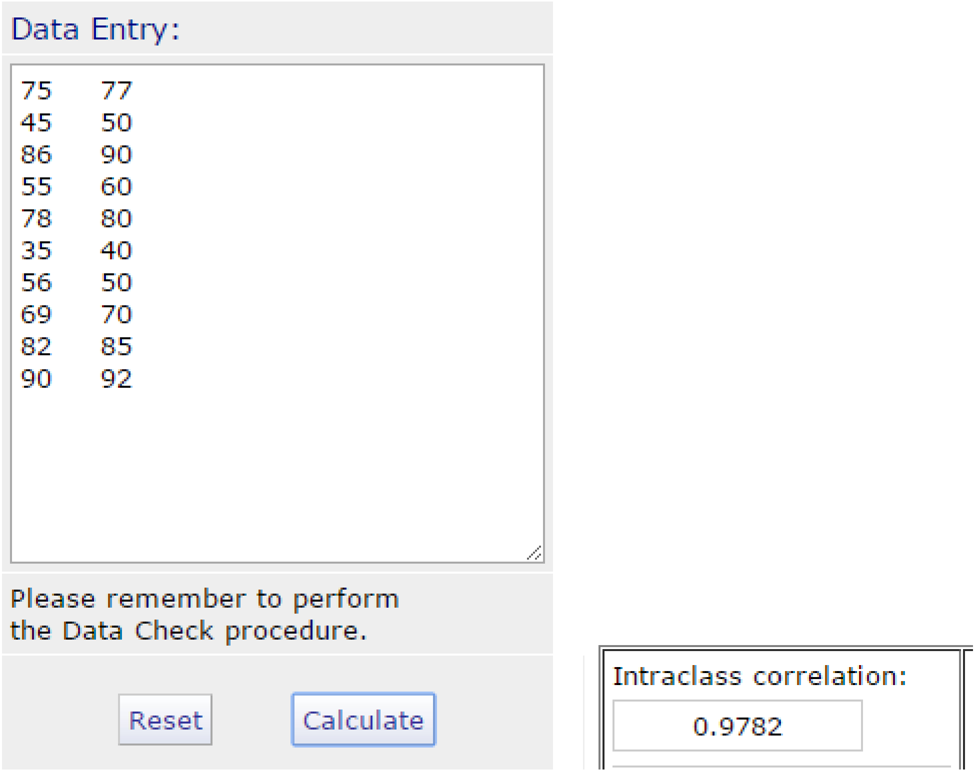
Unordered pairs of observations means the observations under variable X and variable Y may be interchanged and so we can have different pairs of observations. If we have ten pairs of observations under X and Y, we can have 2^{10} unique combinations of X and Y. Pearson's product moment coefficient of correlation can be computed to each combination. The average of these correlation coefficients will be very close to *Interclass correlation coefficient*.

As an example, consider IQ of ten pairs of twins as given below.



IQ of twins	
75	77
45	50
86	90
55	60
78	80
35	40
56	50
69	70
82	85
90	92

Data in excel sheet**:** If the data is in excel sheet, copy the data and paste it in the *Data entry* field. See to that the cursor is next to the last entry. Then click *calculate.*



Direct entry method**:** Data can be entered directly in the *data entry* field as explained under Linear Correlation section. Then Inter class correlation can be calculated.

Simple Logistic Regression

Logistic Regression is performed when the dependent variable is binary in nature. Vassarstats can perform Simple Logistic Regression.

As an example, consider gestational age of infants (in weeks)(X) and whether the baby was breast feeding(Y)or not at the time of discharge from the hospital. Y is coded as Y=1 for ‘yes’ and Y=0 for ‘no’

Enter the data in data entry field and the click *calculate1*.

Data Entry:

X	Instances of Y Coded as 0	Instances of Y Coded as 1
28	4	3
32	5	4
30	2	5
31	3	2
29	3	4

Reset Calculate 1

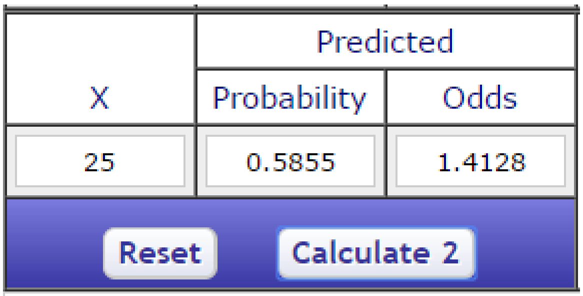
The following result is shown.

For weighted linear regression of log odds on X:

intercept:	1.7181
slope:	-0.0549
exp(slope):	0.9466
R^2:	0.0281

X	Probabilities		Odds	
	Observed	Predicted	Observed	Predicted
28	0.4286	0.5451	0.75	1.1983
32	0.4444	0.4903	0.8	0.962
30	0.7143	0.5178	2.5	1.0737
31	0.4	0.504	0.6667	1.0163
29	0.5714	0.5315	1.3333	1.1343

To get predicted probability, enter the value of X in respective cell and click *Calculate2*. The result is shown below.



X	Predicted	
	Probability	Odds
25	0.5855	1.4128

Conclusion

From this article, it is very clear that Vassarstats is a very simple and useful tool for measuring Correlation and computing Regression coefficients. Infact, Vassarstats can perform almost thirteen concepts of Correlation and Regression analysis.

References

1. Vassarstats.net
2. Y H Chan. Biostatistics 104: Correlational Analysis.Singapore Med J 2003; 44(12): 614-19.
3. Y H Chan. Biostatistics 201: Linear Regression Analysis.Singapore Med J 2004; 45(2): 55-61.

TABLES SECTION

r	meaning
+1	Perfect positive correlation
-1	Perfect negative correlation
0	No correlation
0 < r < 1	Positive correlation
-1 < r < 0	Negative correlation

Table 1