Correlation & Linear Regression

Perhaps the most important analytical tool in the physical sciences is the ability to quantify the validity of a model relating a set of measurable parameters. The idea is as follows:

(1) Rearrange the model in such a way that it becomes a *linear equation* of the form $y = mx + c$

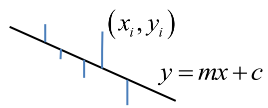
(2) Plot experimental (x,y) data on a graph and determine the **line of best fit** through the data.

(3) Determine *gradient* m and *vertical intercept* c from the line of best fit.

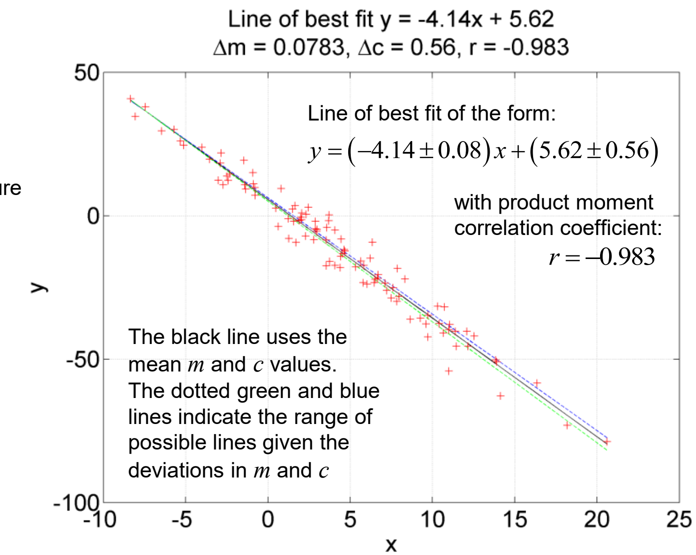
(4) Determine the standard deviation of both gradient m and intercept c, and a quantitative measure of how good the fit is (this is called the **product moment correlation coefficient**).

To determine the line of best fit*, let us sum the *squared* deviations of (x,y) from the line of best fit.

$$\boxed{S = \sum_{i=1}^{N} (y_i - mx_i - c)^2}$$



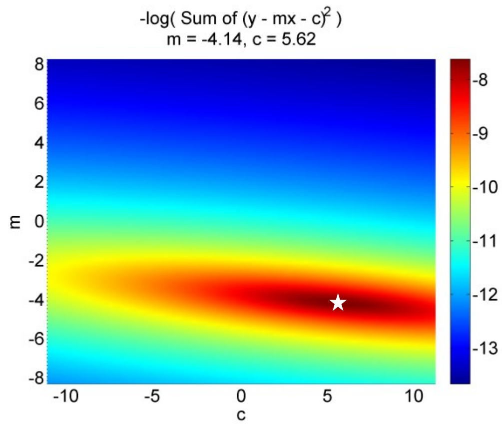
Using the (*negatively correlated)* data on the right, we can plot a surface of S vs m and c values. We can see this has a **minimum** at a particular (m,c) coordinate. (Note for clarity the plots below are of $-\log S$, so the (m,c) coordinate corresponds to the peak, i.e. maximum, instead).

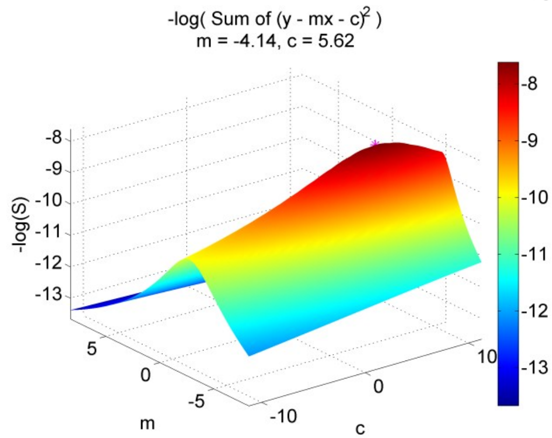


Define the following quantities:

$$\boxed{\begin{aligned} &\bar{x} = \tfrac{1}{N}\sum_{i=1}^{N} x_i, \quad \bar{y} = \tfrac{1}{N}\sum_{i=1}^{N} y_i, \\ &\overline{x^2} = \tfrac{1}{N}\sum_{i=1}^{N} x_i^2, \quad \overline{y^2} = \tfrac{1}{N}\sum_{i=1}^{N} y_i^2 \\ &\overline{xy} = \tfrac{1}{N}\sum_{i=1}^{N} x_i y_i \\ &V[x] = \overline{x^2} - \bar{x}^2, \quad V[y] = \overline{y^2} - \bar{y} \\ &\operatorname{cov}[x,y] = \overline{xy} - \overline{xy} \end{aligned}}$$

i.e. variance and *covariance*





The minimum of S can be found by differentiating S with respect to m and c, and setting these expressions equal to zero. Since S is a function of two variables we must use *partial derivatives*.

$$S = \sum_{i=1}^{N} (y_i - mx_i - c)^2$$

$$\frac{\partial S}{\partial m} = 2\sum_{i=1}^{N} (y_i - mx_i - c)(-x_i)$$

$$\therefore \frac{\partial S}{\partial m} = 0 \Rightarrow \sum_{i=1}^{N} x_i (y_i - mx_i - c) = 0$$

$$\therefore \sum_{i=1}^{N} x_i y_i - m\sum_{i=1}^{N} x_i^2 - c\sum_{i=1}^{N} x_i = 0$$

$$S = \sum_{i=1}^{N} (y_i - mx_i - c)^2$$

$$\frac{\partial S}{\partial c} = 2\sum_{i=1}^{N} (y_i - mx_i - c)(-1)$$

$$\therefore \frac{\partial S}{\partial c} = 0 \Rightarrow \sum_{i=1}^{N} (y_i - mx_i - c) = 0$$

$$\therefore \sum_{i=1}^{N} y_i - m\sum_{i=1}^{N} x_i - cN = 0$$

Hence:

$$\sum_{i=1}^{N} x_i y_i - m\sum_{i=1}^{N} x_i^2 - c\sum_{i=1}^{N} x_i = 0 \quad \therefore \overline{xy} - m\overline{x^2} - c\bar{x} = 0$$

$$\sum_{i=1}^{N} y_i - m\sum_{i=1}^{N} x_i - cN = 0 \quad \therefore \bar{y} - m\bar{x} - c = 0$$

Therefore:

$$\boxed{c = \bar{y} - m\bar{x}}$$

$$\therefore \overline{xy} - m\overline{x^2} - (\bar{y} - m\bar{x})\bar{x} = 0$$

$$\therefore m\left(\bar{x}^2 - \overline{x^2}\right) + \overline{xy} - \bar{y}\bar{x} = 0$$

$$\therefore \boxed{m = \frac{\overline{xy} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{\operatorname{cov}[x,y]}{V[x]}}$$

If we repeat the analysis for the line: $x = My + d \Rightarrow M = \dfrac{\operatorname{cov}[x,y]}{V[y]}$

If this was the *same line but rearranged*:

$M = \frac{1}{m} \quad \therefore mM = 1$

$y = mx + c$

Hence define a **product moment correlation coefficient:**

$$\boxed{r = \frac{\operatorname{cov}[x,y]}{\sqrt{V[x]V[y]}}}$$

This will be +1 for a perfect positive correlation and -1 for a perfect negative correlation (i.e. $S = 0$ in both cases).

*We will use the *vertical* deviations. You can alternatively use horizontal deviations or indeed perpendicular deviations from the line of best fit.

It is possible to show* that the standard deviations (i.e. 'errors') in m and c are:

$$\Delta m = \frac{s}{\sqrt{N}}\frac{1}{\sqrt{V[x]}}$$

$$\Delta c = \frac{s}{\sqrt{N}}\sqrt{1+\frac{\overline{x}^2}{V[x]}}$$

$$s = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - mx_i - c\right)^2}$$

This is very useful in the physical sciences, as the errors in m and c will often be the uncertainties in model parameters (e.g. the strength of gravity...)

s is the *unbiased estimator* of the standard deviation in the y values from the line of best fit. The N-2 factor is due to two parameters (m and c) being used in the calculation, which are of course derived from the sample data themselves as shown above.

In many situations a **direct proportion** is asserted between y and x. The computation of the line of best fit (which passes through (0,0) follows a similar argument to the one above.

$$S = \sum_{i=1}^{N}\left(y_i - mx_i\right)^2$$

$$\frac{\partial S}{\partial m} = 2\sum_{i=1}^{N}\left(y_i - mx_i\right)\left(-x_i\right)$$

$$\therefore \frac{\partial S}{\partial m} = 0 \Rightarrow \sum_{i=1}^{N} x_i\left(y_i - mx_i\right) = 0$$

$$\therefore \sum_{i=1}^{N} x_i y_i - m\sum_{i=1}^{N} x_i^2 = 0$$

$$\therefore \boxed{m = \frac{\overline{xy}}{\overline{x^2}}}$$

The product moment correlation coefficient is the same as before but the standard deviation in m is slightly different since only *one* parameter is used in the computation of s (i.e. m).

$$\Delta m = \frac{s}{\sqrt{N}}\frac{1}{\sqrt{V[x]}}$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - mx_i\right)^2}$$

$$r = \frac{\text{cov}[x,y]}{\sqrt{V[x]V[y]}}$$

LINE OF BEST FIT CALCULATOR y = mx + c

Dr Andy French. March 2019

paste as values x,y data here

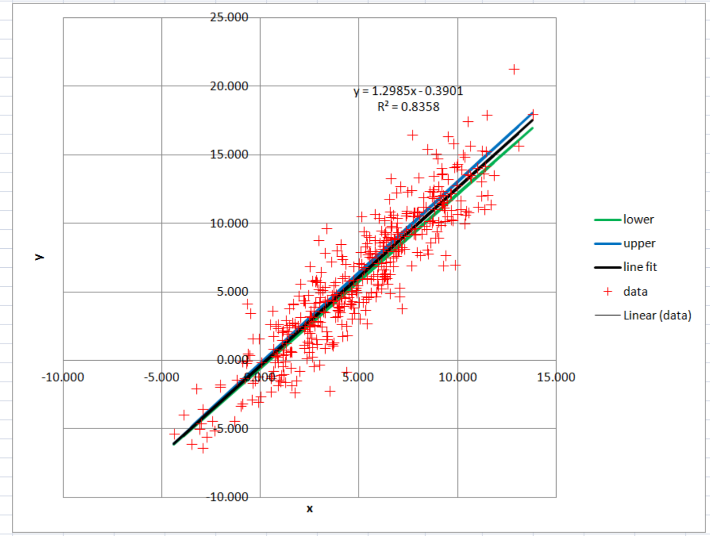
	x	y	x^2	y^2	xy	xfit	yfit	(y-fit	ylow	yupp
6	0.4647	2.6687	0.216	7.122	1.240	0.465	0.213	6.029	0.030	0.397
7	0.8766	2.3997	0.769	5.758	2.104	0.877	0.748	2.727	0.553	0.944
8	-0.698	4.1591	0.487	17.299	-2.903	-0.698	-1.296	29.762	-1.447	-1.145
9	-0.401	1.6333	0.161	2.668	-0.655	-0.401	-0.911	6.475	-1.071	-0.752
10	4.1428	7.4095	17.163	54.901	30.697	4.143	4.990	5.856	4.702	5.277
11	-4.397	-5.303	19.337	28.120	23.318	-4.397	-6.100	0.636	-6.147	-6.053
12	4.6021	3.0838	21.179	9.510	14.192	4.602	5.586	6.261	5.286	5.886
13	0.8422	0.2935	0.709	0.086	0.247	0.842	0.704	0.168	0.509	0.898
14	2.1911	1.1794	4.801	1.391	2.584	2.191	2.455	1.627	2.223	2.687
15	-3.114	-4.958	9.694	24.578	15.436	-3.114	-4.433	0.275	-4.516	-4.350
16	-0.941	-0.048	0.886	0.002	0.045	-0.941	-1.613	2.449	-1.757	-1.468
17	-0.294	-1.475	0.086	2.175	0.433	-0.294	-0.771	0.495	-0.934	-0.609
18	1.1318	-0.398	1.281	0.158	-0.451	1.132	1.080	2.184	0.877	1.282
19	-3.03	-4.546	9.182	20.664	13.774	-3.030	-4.325	0.049	-4.410	-4.239
20	1.2774	1.2494	1.632	1.561	1.596	1.277	1.269	0.000	1.062	1.475
21	-2.068	-1.725	4.275	2.974	3.566	-2.068	-3.075	1.824	-3.188	-2.963
22	0.5142	-1.273	0.264	1.620	-0.655	0.514	0.278	2.404	0.093	0.463
23	1.7104	2.7385	2.926	7.499	4.684	1.710	1.831	0.824	1.612	2.050
24	1.0837	1.489	1.174	2.217	1.614	1.084	1.017	0.223	0.816	1.218
25	2.6187	5.8698	6.858	34.455	15.371	2.619	3.010	8.176	2.766	3.255
26	2.4895	2.6357	6.198	6.947	6.562	2.489	2.843	0.043	2.602	3.083
27	-0.801	-1.254	0.641	1.572	1.004	-0.801	-1.430	0.031	-1.578	-1.282
28	1.0392	-1.065	1.080	1.133	-1.106	1.039	0.959	4.096	0.759	1.159
29	-1.34	-4.419	1.795	19.523	5.920	-1.340	-2.130	5.238	-2.263	-1.997
30	-2.071	-1.926	4.288	3.710	3.988	-2.071	-3.079	1.330	-3.192	-2.967
31	2.4463	2.51	5.984	6.300	6.140	2.446	2.787	0.076	2.547	3.026
32	0.6206	1.7709	0.385	3.136	1.099	0.621	0.416	1.836	0.228	0.604
33	0.5346	-1.128	0.286	1.272	-0.603	0.535	0.304	2.050	0.118	0.490
34	2.4905	6.8804	6.203	47.339	17.136	2.491	2.844	16.293	2.603	3.085
35	1.8813	2.5584	3.539	6.546	4.813	1.881	2.053	0.256	1.829	2.276
36	1.4164	3.8362	2.006	14.717	5.434	1.416	1.449	5.698	1.239	1.660
37	3.2404	3.5318	10.500	12.474	11.444	3.240	3.818	0.082	3.556	4.079
38	2.6233	2.9017	6.882	8.420	7.612	2.623	3.016	0.013	2.772	3.261
39	1.2671	2.9387	1.606	8.636	3.724	1.267	1.255	2.834	1.049	1.462

N	420

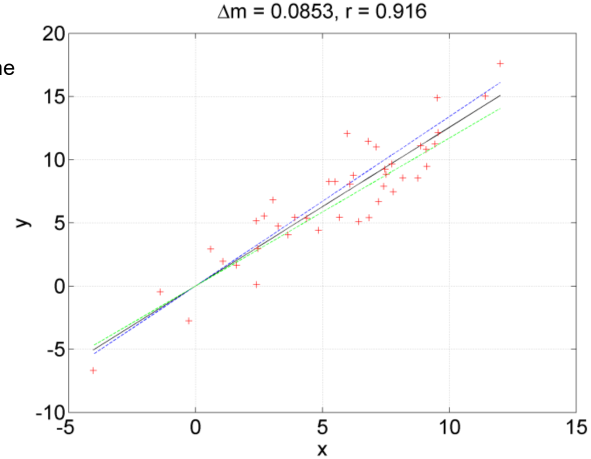
xbar	ybar	x^2 bar	y^2 bar	xy bar
4.961	6.052	36.855	61.327	45.922

Vx	Vy	Cov[x,y]	s
12.242	24.699	15.897	2.016

r	r^2		m	c		dm	dc
0.914	0.836		1.299	-0.390		0.028	0.171



Line of best fit y = 1.26x
Δm = 0.0853, r = 0.916



Line of best fit (or '**linear regression**') analysis can be clearly demonstrated using a computer spreadsheet package such as Microsoft Excel.

In the above example, the gradient and vertical intercept values are manually computed, and compared to the built-in *trendline* function.

* http://mathworld.wolfram.com/LeastSquaresFitting.html

Hypothesis testing of correlation using Student's t-test

We can define a **null hypothesis** H_0 that $\{x,y\}$ data, with N data points, with a particular product-moment correlation coefficient r is ***uncorrelated***.

$$\boxed{r = \frac{\mathrm{cov}[x,y]}{\sqrt{V[x]V[y]}}}$$

To **assess whether the null hypothesis is rejected** (i.e. *the data is correlated*, or 'not uncorrelated') to a *significance* s
We can apply a **1-tail t-test** to determine a critical t value t_*, and then use the formula above to determine the critical r value.

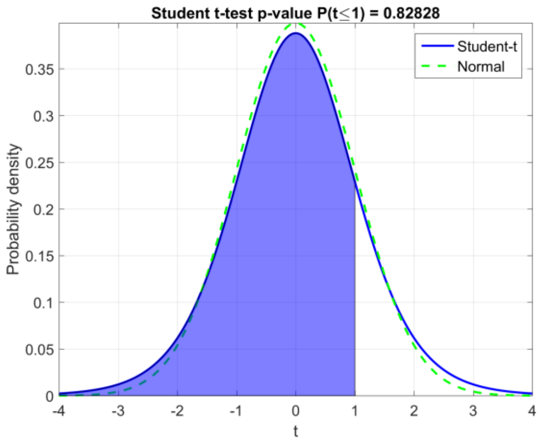
H_0 (i.e. $\{x,y\}$ data *uncorrelated*) is ***rejected*** (to significance s) if the modulus of the r for the data set value **is greater than the critical value of *r***

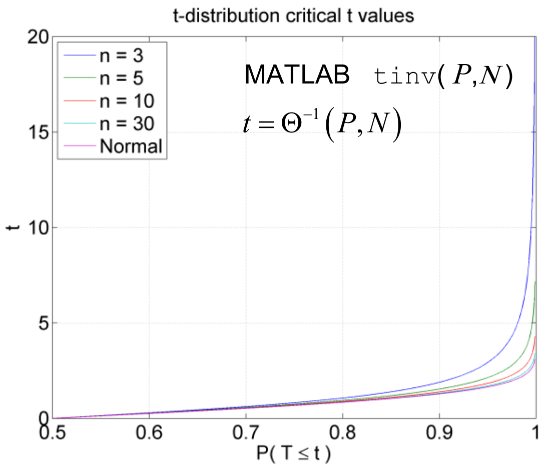
$$\boxed{r_* = \frac{t_*}{\sqrt{t_*^2 + N - 2}}}$$

i.e. H_0 rejection implies ***r*** is > 0 ***or*** < 0.

If instead we go for ***r* does not equal 0**
Then we need a *two tail test*. In this case use 0.5s.

William "Student"
Sealy Gosset 1876-1937
Worked for Guinness and was educated at Winchester College





Student t-distribution

$$\boxed{p(t) = \frac{\Gamma\left(\frac{1}{2}(v+1)\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{1}{2}v\right)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}} \qquad v = N-1}$$

P-value of t-distribution

$$\boxed{\begin{aligned} &P(t \le t_*) = \int_{-\infty}^{t_*} p(t)\,dt = \Theta(t_*) \\ &P = 1 - s \\ &t_* = \Theta^{-1}(P,N) \end{aligned}}$$

$$\boxed{\Gamma(x) = \int_0^\infty z^{x-1} e^{-z}\,dz}$$

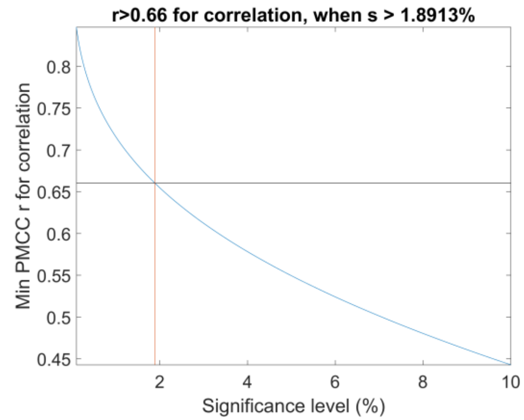
Gamma (Special) function

```
function t = tinv_calc(PHI,N)
v = N-1; % Degrees of freedom
         % of t-distribution
i = find(PHI<0.5); PHI(i) = 1 - PHI(i);
y = betainc( 1, v/2,0.5 ) - (2*PHI-1);
x = betaincinv(y, v/2, 0.5 );
t = sqrt( v./x - v ); t(i) = -t(i);
```

MATLAB `tinv(` P,N `)`

$$\boxed{\beta(x,z,w) = \frac{\Gamma(z+w)}{\Gamma(z)\Gamma(w)}\int_0^x k^{z-1}(1-k)^{w-1}\,dk}$$

Incomplete beta-function



Example 1: For a dataset of $N=10$ $\{x,y\}$ pairs, determine the minimum r value to reject the null hypothesis of no correlation as a function of significance $100s$ (%).

What is the smallest significance such that $r > 0.66$ implies the data is correlated?

Note two tails since H_0 means $r \neq 0$

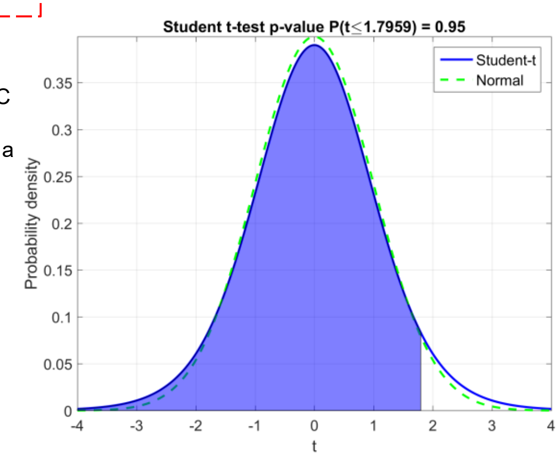
Example 2: For a dataset of $N=12$ $\{x,y\}$ pairs, what is the critical PMCC value such that a null hypothesis of uncorrelated data can be rejected to a significance of 10%?

$$t_* = \Theta^{-1}(1-0.05,12) = 1.7959$$

$$\therefore r_* = \frac{t_*}{\sqrt{t_*^2 + N - 2}}$$

$$\therefore r_* = \frac{1.7959}{\sqrt{1.7959^2 + 12 - 2}}$$

$$\therefore \boxed{r_* = 0.4938}$$



CRITICAL VALUES OF THE PRODUCT MOMENT CORRELATION COEFFICIENT

	One tail significance (%)				
Sample size N	10%	5%	2.50%	1%	0.50%
4	0.8000	0.9000	0.9500	0.9800	0.9900
5	0.6870	0.8054	0.8783	0.9343	0.9587
6	0.6084	0.7293	0.8114	0.8822	0.9172
7	0.5509	0.6694	0.7545	0.8329	0.8745
8	0.5067	0.6215	0.7067	0.7887	0.8343
9	0.4716	0.5822	0.6664	0.7498	0.7977
10	0.4428	0.5494	0.6319	0.7155	0.7646
11	0.4187	0.5214	0.6021	0.6851	0.7348
12	0.3981	0.4973	0.5760	0.6581	0.7079
13	0.3802	0.4762	0.5529	0.6339	0.6835
14	0.3646	0.4575	0.5324	0.6120	0.6614
15	0.3507	0.4409	0.5140	0.5923	0.6411
16	0.3383	0.4259	0.4973	0.5742	0.6226
17	0.3271	0.4124	0.4821	0.5577	0.6055
18	0.3170	0.4000	0.4683	0.5425	0.5897
19	0.3077	0.3887	0.4555	0.5285	0.5751
20	0.2992	0.3783	0.4438	0.5155	0.5614
21	0.2914	0.3687	0.4329	0.5034	0.5487
22	0.2841	0.3598	0.4227	0.4921	0.5368
23	0.2774	0.3515	0.4132	0.4815	0.5256
24	0.2711	0.3438	0.4044	0.4716	0.5151
25	0.2653	0.3365	0.3961	0.4622	0.5052
26	0.2598	0.3297	0.3882	0.4534	0.4958
27	0.2546	0.3233	0.3809	0.4451	0.4869
28	0.2497	0.3172	0.3739	0.4372	0.4785
29	0.2451	0.3115	0.3673	0.4297	0.4705
30	0.2407	0.3061	0.3610	0.4226	0.4629
31	0.2366	0.3009	0.3550	0.4158	0.4556
32	0.2327	0.2960	0.3494	0.4093	0.4487
33	0.2289	0.2913	0.3440	0.4032	0.4421

	One tail significance (%)				
Sample size N	10%	5%	2.50%	1%	0.50%
34	0.2254	0.2869	0.3388	0.3972	0.4357
35	0.2220	0.2826	0.3338	0.3916	0.4296
36	0.2187	0.2785	0.3291	0.3862	0.4238
37	0.2156	0.2746	0.3246	0.3810	0.4182
38	0.2126	0.2709	0.3202	0.3760	0.4128
39	0.2097	0.2673	0.3160	0.3712	0.4076
40	0.2070	0.2638	0.3120	0.3665	0.4026
41	0.2043	0.2605	0.3081	0.3621	0.3978
42	0.2018	0.2573	0.3044	0.3578	0.3932
43	0.1993	0.2542	0.3008	0.3536	0.3887
44	0.1970	0.2512	0.2973	0.3496	0.3843
45	0.1947	0.2483	0.2940	0.3457	0.3801
46	0.1925	0.2455	0.2907	0.3420	0.3761
47	0.1903	0.2429	0.2876	0.3384	0.3721
48	0.1883	0.2403	0.2845	0.3348	0.3683
49	0.1863	0.2377	0.2816	0.3314	0.3646
50	0.1843	0.2353	0.2787	0.3281	0.3610
60	0.1678	0.2144	0.2542	0.2997	0.3301
70	0.1550	0.1982	0.2352	0.2776	0.3060
80	0.1448	0.1852	0.2199	0.2597	0.2864
90	0.1364	0.1745	0.2072	0.2449	0.2702
100	0.1292	0.1654	0.1966	0.2324	0.2565

If modulus of PMCC r is greater than these values, then **null hypothesis of no correlation** is ***rejected***. i.e. 'data is potentially correlated.'

* http://mathworld.wolfram.com/LeastSquaresFitting.html