Correlation and simple regression

In this section we discuss the population and sample correlation coefficients. These quantities are measures of linear association between two variables. Pearson's r is appropriate when both the independent and dependent variables are measured on approximately an interval level or on a ratio level. The correlation coefficient (sometimes called the "simple correlation coeff" or the "Pearson's r" or the "product-moment correlation coefficient") is calculated as

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

e.g. Can a person's weight as a young adult be predicted from his/her birth wt.?

subject	Birth wt (oz), x	Adult (lb) y
1	112	120
2	118	135
3	118	125
4	118	138
5	101	115
6	93	118
7	128	145
8	118	120
9	140	144
10	108	130
11	118	120
12	143	135
13	112	120
14	127	132
15	112	125
16	134	135

p8.1a

Use Minitab to perform data on Page 8.1

```
MTB > read 'p81dat' c1 c2

MTB > prin c1-c2

 ROW      C1      C2

   1     112     120
   2     118     135
   3     118     125
   4     118     138
   5     101     115
   6      93     118
   7     128     145
   8     118     120
   9     140     144
  10     108     130
  11     118     120
  12     123     135
  13     112     120
  14     127     132
  15     112     125
  16     134     135

MTB > desc c1 c2

                 N      MEAN    MEDIAN    TRMEAN     STDEV    SEMEAN
C1              16    117.50    118.00    117.64     11.73      2.93
C2              16    128.56    127.50    128.36      9.49      2.37

                MIN       MAX        Q1        Q3
C1            93.00    140.00    112.00    126.00
C2           115.00    145.00    120.00    135.00

MTB > let c3=c1-117.5
MTB > let c4=c2-128.56
MTB > let c5=c3*c4
MTB > let c6=c3*c3
MTB > let c7=c4*c4
MTB > prin c1-c7
```

ROW	C1	C2	C3	C4	C5	C6	C7
1	112	120	-5.5	-8.5600	47.080	30.25	73.274
2	118	135	0.5	6.4400	3.220	0.25	41.474
3	118	125	0.5	-3.5600	-1.780	0.25	12.674
4	118	138	0.5	9.4400	4.720	0.25	89.114
5	101	115	-16.5	-13.5600	223.740	272.25	183.874
6	93	118	-24.5	-10.5600	258.720	600.25	111.514
7	128	145	10.5	16.4400	172.620	110.25	270.274
8	118	120	0.5	-8.5600	-4.280	0.25	73.274
9	140	144	22.5	15.4400	347.400	506.25	238.394
10	108	130	-9.5	1.4400	-13.680	90.25	2.074
11	118	120	0.5	-8.5600	-4.280	0.25	73.274
12	123	135	5.5	6.4400	35.420	30.25	41.474
13	112	120	-5.5	-8.5600	47.080	30.25	73.274
14	127	132	9.5	3.4400	32.680	90.25	11.834
15	112	125	-5.5	-3.5600	19.580	30.25	12.674
16	134	135	16.5	6.4400	106.260	272.25	41.474

```
MTB > let k1=sum(c5)
MTB > let k2=sum(c6)
MTB > let k3=sum(c7)
MTB > prin k1,k2,k3
K1        1274.50
K2        2064.00
K3        1349.94
MTB > let k4=k1/sqrt(k2*k3)
MTB > prin k4
K4        0.763534
MTB > corr c1 c2

Correlation of C1 and C2 = 0.764

MTB > regr c2 1 c1

The regression equation is
C2 = 56.0 + 0.617 C1

Predictor       Coef      Stdev    t-ratio       p
Constant       56.01      16.48       3.40   0.004
C1            0.6175     0.1396       4.42   0.001

s = 6.341       R-sq = 58.3%     R-sq(adj) = 55.3%

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1      786.99      786.99      19.57   0.001
Error        14      562.95       40.21
Total        15     1349.94


MTB > stop
```

Hypothesis about the correlation coefficient

The correlation coefficient, r, which we calculate from a sample, is an estimate of a population parameter namely the correlation coefficient in the population that was sampled. This parameter is denoted by ρ. If we wish to ask whether there is, in fact, a correlation between Y and X in the population, we can test:

$H_0: \rho = 0$

$H_A: \rho \neq 0$ Test statistic $t_{calc} = \dfrac{r\sqrt{N-2}}{\sqrt{1-r^2}}$

$\sim t_{N-2}$ d.f.

e.g. Calculate the sample correlation coefficient, and testing $H_0: \rho = 0$. The data are wing and tail lengths among birds of a particular species

Wing length (cm) X	Tail length (cm) Y
10.4	7.4
10.8	7.6
11.1	7.9
10.2	7.2
10.3	7.4
10.2	7.1
10.7	7.4
10.5	7.2
10.8	7.8
11.2	7.7
10.6	7.8
11.4	8.3

8.2a

Pearson Corelation for P8.2 .

```
. use minitab
initiating minitab (MiniTab 7.2) for use
Variable set: MTBMACRO=/pl2a/minitab7.2/macros
acs4:/u5/tfung/stat601.14--> minitab
 MINITAB RELEASE 7.2.3 *** IBM Risc System 6000 Version
 (C) Copyright 1989 Minitab, Inc. - ALL RIGHTS RESERVED
 U.S. FEDERAL GOVERNMENT USERS SEE HELP FGU
 MARCH 31, 1999 *** UNIV OF CALGARY
 STORAGE AVAILABLE 248000
      Type NEWS for information on new features in
      Minitab Release 7.2
> read 'data1' c1 c2
      12 ROWS READ

  ROW     C1     C2

    1   10.4    7.4
    2   10.8    7.6
    3   11.1    7.9
    4   10.2    7.2
   .   .   .

 MTB > desc c1 c2

                 N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
C1              12   10.683   10.650   10.660    0.395    0.114
C2              12    7.567    7.500    7.540    0.350    0.101

               MIN      MAX       Q1       Q3
C1          10.200   11.400   10.325   11.025
C2           7.100    8.300    7.250    7.800

 MTB > let c3=c1-10.683
 MTB > let c4=c2-7.567
 MTB > let c5=c3*c4
 MTB > let c6=c3*c3
 MTB > let c7=c4*c4
 MTB > let k1=sum(c5)
 MTB > let k2=sum(c6)
 MTB > let k3=sum(c7)
 MTB > let k4=sqrt(k2)
 MTB > let k5=sqrt(k3)
 MTB > let k6=k1/(k4*k5)
 MTB > prin k6
 K6        0.870353
 MTB > corr c1 c2

 Correlation of C1 and C2 = 0.870
```

8.3 (P3)

$r = 0.866$

$H_0: \rho = 0$, $H_A: \rho \neq 0$ use $\alpha = 0.05$

$$t_{calc} = \frac{0.866\sqrt{12-2}}{\sqrt{1-(0.866)^2}} = \frac{0.866\sqrt{10}}{\sqrt{0.25}}$$

$$= 5.48$$

$$t_{\frac{\alpha}{2}}(10) = t_{0.025}(10) = 2.228$$

$t_{calc} = 5.48 > 2.228$ reject H_0 at 0.05 level

If one desires to test $H: \rho = \rho_0$ for any $\rho_0 \neq 0$ and $\rho_0 \neq \pm 1$,
$H_A: \rho \neq \rho_0$
we have to use Fisher's z-transformation,

which is $$z(r) = \frac{1}{2}\log_e \frac{1+r}{1-r}$$

Under H_0, $z(r) \sim N\left(z(\rho_0), \frac{1}{N-3}\right)$

The test statistic is then computed as:

$$\zeta = \frac{z(r) - z(\rho_0)}{\left(\frac{1}{\sqrt{N-3}}\right)}$$

If $|\zeta| > \zeta_{\frac{\alpha}{2}}$, then reject H_0 at α level.

∴ $100(1-\alpha)\%$ confidence interval for $z(\rho)$:

$$-z_{\alpha/2} < \frac{z(r) - z(\rho)}{\sigma_{z(r)}} < z_{\alpha/2}$$

$$\Rightarrow z(r) - z_{\alpha/2}\,\sigma_{z(r)} < z(\rho) < z(r) + z_{\alpha/2}\,\sigma_{z(r)}$$

$$\Rightarrow z(r) - z_{\alpha/2}\frac{1}{\sqrt{N-3}} < z(\rho) < z(r) + z_{\alpha/2}\frac{1}{\sqrt{N-3}} \quad (1)$$

Use the inverse Fisher transformation:

$$r = \frac{e^{2z(r)} - 1}{e^{2z(r)} + 1}$$, the confidence interval for ρ are obtained.

e.g. (a) Use the data 'wing & tail lengths' example

$H_0: \rho = 0.75$, $r = 0.866$

$H_A: \rho \neq 0.75$ $N = 12$ $\alpha = 0.05$

(b) construct 95% C.I. of ρ.

Solution: $z(r) = z(.866) = \frac{1}{2}\log_e \frac{1+0.866}{1-0.866}$

$$= \frac{1}{2}\log_e \frac{1.866}{0.134} = 1.3169$$

$$z(\rho_0) = z(0.75) = \frac{1}{2}\log_e \frac{1+0.75}{1-0.75} = \frac{1}{2}\log_e \frac{1.75}{0.25}$$

$$= 0.9729$$

$$z_{calc} = \frac{z(.866) - z(.75)}{\left(\frac{1}{\sqrt{9}}\right)} = (1.3169 - .9729) \times 3$$

$$= 1.032$$

$z_{0.025} = 1.96$, $z_{calc} = 1.032 < 1.96$

DO NOT reject H_0 at 0.05 level.

(b) $Z(r) = 1.3169$, $N = 12$, $z_{0.025} = 1.96$

from eqt ①, we have

$$1.3169 - \frac{1.96}{3} \leq Z(\rho) \leq 1.3169 + \frac{1.96}{3}$$

$$\Rightarrow \quad 0.664 \leq Z(\rho) \leq 1.970$$

$$\Rightarrow \quad 0.664 \leq \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \leq 1.970$$

$\therefore$ 95% C.I. for $Z(\rho) = (0.664, 1.970)$

$$\Rightarrow \text{95\% C.I. for } \rho = \left(\frac{e^{2\times 0.664} - 1}{e^{2\times 0.664} + 1}, \frac{e^{2\times 1.97} - 1}{e^{2\times 1.97} + 1}\right)$$

$$= \left(\frac{e^{1.328} - 1}{e^{1.328} + 1}, \frac{e^{3.94} - 1}{e^{3.94} + 1}\right) = \left(\frac{3.773 - 1}{3.773 + 1}, \frac{51.419 - 1}{51.419 + 1}\right)$$

$$= (0.581, 0.962)$$

Comparing Two Correlation Coefficients:

Hypothesis about two correlation coefficients may be approached by the use of

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}}$$

e.g. For a sample of 15 bird wing & tail length a correlation coeff. of 0.843 was calculated. A sample of 18 such measurements from a second bird species yielded a correlation coefficient of 0.78 Let us test the equality of two population coefficients.

$H_0: \rho_1 = \rho_2$
$H_A: \rho_1 \neq \rho_2$ $\alpha = 0.05$

$r_1 = .843$
$n_1 = 15$

$$z_1 = z(r_1) = \frac{1}{2}\log_e \frac{1+.843}{1-0.843} = \frac{1}{2}\log_e \frac{1.843}{0.157} = 1.2314$$

$r_2 = 0.784$

$n_2 = 18$

$$z_2 = z(r_2) = \frac{1}{2}\log_e \frac{1+.784}{1-.784} = \frac{1}{2}\log_e \frac{1.784}{0.216} = 1.0557$$

$$Z = \frac{1.2314 - 1.0557}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0.1758}{0.3873} = 0.454$$

$z_{\alpha/2} = 1.96$ $0.454 < 1.9$

DO NOT reject H_0 at 0.05 level

Spearman Rank Correlation

Spearman's Rank correlation is a nonparametric analog of Pearson's r. A correlation coefficient that makes use of the rank-order information in the observations, and not their actual values, was proposed by the British psychologist Charles Spearman in 1904. It is typically used when the dependent variable is measured on the ordinal scale, or when one or both variables being correlated is severely skewed or has an outlier. It is also preferred by some researchers when the number of case is less than 30.

If we assign ranks (1, 2, 3 ... n) to the observational units with respect to the values of each variable separately and then compute the Pearson correlation of the ranks, the result is Spearman's rank correlation coefficient:

$$r_s = r_{ranks}$$

Spearman's r_s is a Pearson correlation coefficient calculated on ranks (r_{ranks}) instead of on raw scores (r_{xy}). It follows that Spearman's r_s is an index of the linear relationships of the ranks of the variables, and that its value is limited to the interval from -1 to 1, inclusive.

In assigning ranks to the n values of the two variables, it is immaterial whether the ranks (1, 2, ... n) are assigned from largest to smallest or from smallest to largest, as long as the direction of ranking

is the same for both variables. However, it is conventional, and generally less conducive to confusion, to assign the rank of 1 to the smallest observation and successively larger rank numbers to successively larger observations. If there are tied observations, they should be assigned the mean of the ranks they would have received if they had not been exactly tied.

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

where d_i is the difference between the two ranks assigned to the two variables for the ith observational unit.

e.g. Owners of show jumping competition horses usually insure their horses, in view of the unpredictable of injury and disease. To what extent does the amount of insurance depend on the # of competitions won by horses in recent yrs?

Horse	# winning (x)	x-rank	insured value y	y-rank	d_i	d_i^2
1	19	5	7.5	5	0	0
2	33	9	10	7	2	4
3	35	10	15	9	1	1
4	20	6.5	25	10	-3.5	12.25
5	4	1	3.5	1	0	0
6	16	4	5	3.5	0.5	.25
7	15	3	5	3.5	-0.5	.25
8	12	2	4	2	0	0
9	20	6.5	10	7	-.5	.25
10	29	8	10	7	1	1
					0	19

8.86

$$r_s = 1 - \frac{6 \times 19}{10(10^2-1)} = 0.883$$

The null hypothesis when using Spearman's Correlation is that there is NO linear relationship between the two sets of ranks. i.e. that the population correlation (ρ_s) is 0. Table B-8 (textbook) of Appendix B, with a sample size of 10, the critical value for $\alpha = 0.05$ (for a two-sided test) is 0.648, $0.883 > 0.648$ reject H_0 at 0.05 level. Note that the table of critical values for Spearman correlation can only used when $N \leq 30$ and $N \geq 5$. When there are more than 30 cases, a reasonably good approximation of the significance test can be obtained by using the table of critical value for Pearson's r (Table B-6 with $N-2$ degree of freedom.

8.8c

```
execute.
list variables=id x rx y ry.
```

List:spearman.spo

id	x	Rx	y	Ry
1.00	19.00	5.000	7.50	5.000
2.00	33.00	9.000	10.00	7.000
3.00	35.00	10.000	15.00	9.000
4.00	20.00	6.500	25.00	10.000
5.00	4.00	1.000	3.50	1.000
6.00	16.00	4.000	5.00	3.500
7.00	15.00	3.000	5.00	3.500
8.00	12.00	2.000	4.00	2.000
9.00	20.00	6.500	10.00	7.000
10.00	29.00	8.000	10.00	7.000

Number of cases read: 10 Number of cases listed: 10

```
correlation rx with ry.
```

Correlations

Correlations

		Ry RANK of y
Rx RANK of x	Pearson Correlation	.883
	Sig. (2-tailed)	.001
	N	10

```
nonpar corr x with y.
```

Nonparametric Correlations

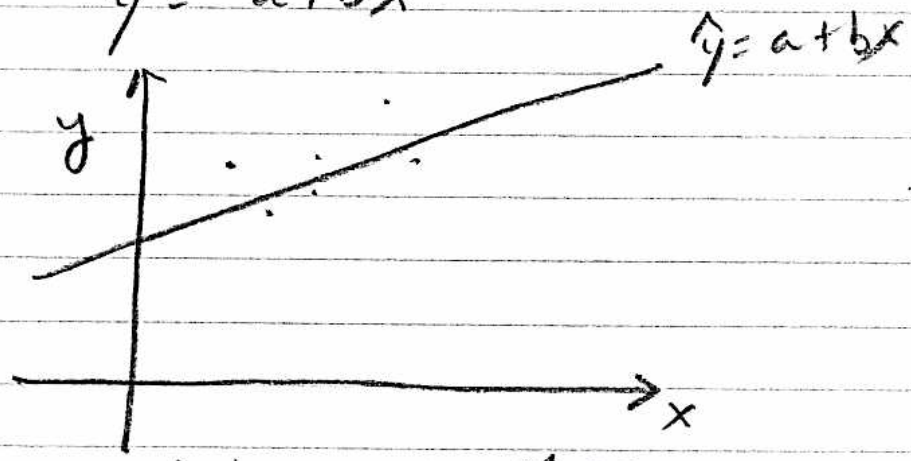
Correlations

			y
Spearman's rho	x	Correlation Coefficient	.883
		Sig. (2-tailed)	.001
		N	10

Simple regression

In research studies, relationships between variables are often linear, but they are never perfect. When a correlation between two variables is not perfect, the statistical technique of regression can be used to identify the straight line that runs through the data points with the best possible fit, as determined by statistical criterion known as least squares. The regression equation is the formula for the best-fitting straight line to characterize the linear relationships between X and y. The basic linear equation is:

$$\hat{y} = a + bX$$



Find a, & b such that

$\sum (y_i - \hat{y}_i)^2$ is minimized.

By Calculus, $b = \dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$

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

e.g.

Mid term	$x_i - \bar{x}$	$(x_i-\bar{x})^2$	Final	$y_i - \bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
2	-3	9	3	-3	9
6	1	1	7	1	1
5	0	0	6	0	0
9	4	16	8	2	8
7	2	4	9	3	6
9	4	16	10	4	16
3	-2	4	4	-2	4
4	-1	1	6	0	0
1	-4	16	2	-4	16
4	-1	1	5	-1	1
$\bar{x}=5.0$	0	68	$\bar{y}=6.0$	0	61

$$b = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2} = \frac{61}{68} = 0.897$$

$$a = \bar{y} - b\bar{x} = 6.0 - (0.897)(5.0) = 1.515$$

$$\hat{y} = 1.515 + 0.897x$$

$$SST = \sum_{i=1}^{n} (y_i-\bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} (y_i-\hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$= SS_{error} + SS_{Reg}$$

SS_{error} : Sum of square due to error/residual

SS_{Reg} : Sum of square due to regression

8.11

P11

$$R^2 = \frac{SS_{reg}}{SS_T}$$

R^2 is defined as the coefficient of determination which is the proportion of the total variation in y that can be accounted for by the linear relationship with x.

The coefficient of determination takes on values from 0 (or 0%) when x and y are NOT linearly related, to 1 (or 100%) when there is a perfect straight-line relationship between x and y in the data.

SPSS syntax for regression:

regression variables = y x / dependent = y / enter x.

Discussion of computer printout in Polit's book p247.

8.11a

This example from Polit's book Page 247:

```
-> get file='d:\stat601.14\p247.sav'.

-> regression variables=final midterm/
->         descriptive=MEAN STDDEV CORR COV/
->         dependent=final/enter midterm.
```

* * * * M U L T I P L E R E G R E S S I O N * * * *

Listwise Deletion of Missing Data

	Mean	Std Dev	Label
FINAL	6.000	2.582	
MIDTERM	5.000	2.749	

N of Cases = 10

Correlation, Covariance:

	FINAL	MIDTERM
FINAL	1.000	.955
	6.667	6.778
MIDTERM	.955	1.000
	6.778	7.556

* * * * M U L T I P L E R E G R E S S I O N * * * *

Equation Number 1 Dependent Variable.. FINAL

Descriptive Statistics are printed on Page 2

Block Number 1. Method: Enter MIDTERM

Variable(s) Entered on Step Number
1.. MIDTERM

Multiple R	.95499
R Square	.91201
Adjusted R Square	.90101
Standard Error	.81236

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	1	54.72059	54.72059
Residual	8	5.27941	.65993

8.11b

F = 82.91922 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
MIDTERM	.897059	.098513	.954992	9.106	.0000
(Constant)	1.514706	.555529		2.727	.0260

End Block Number 1 All requested variables entered.