Correlation analysis

Characterizing the relationship between two quantitative variables X and Y

Ricco Rakotomalala
Ricco.Rakotomalala@univ-lyon2.fr

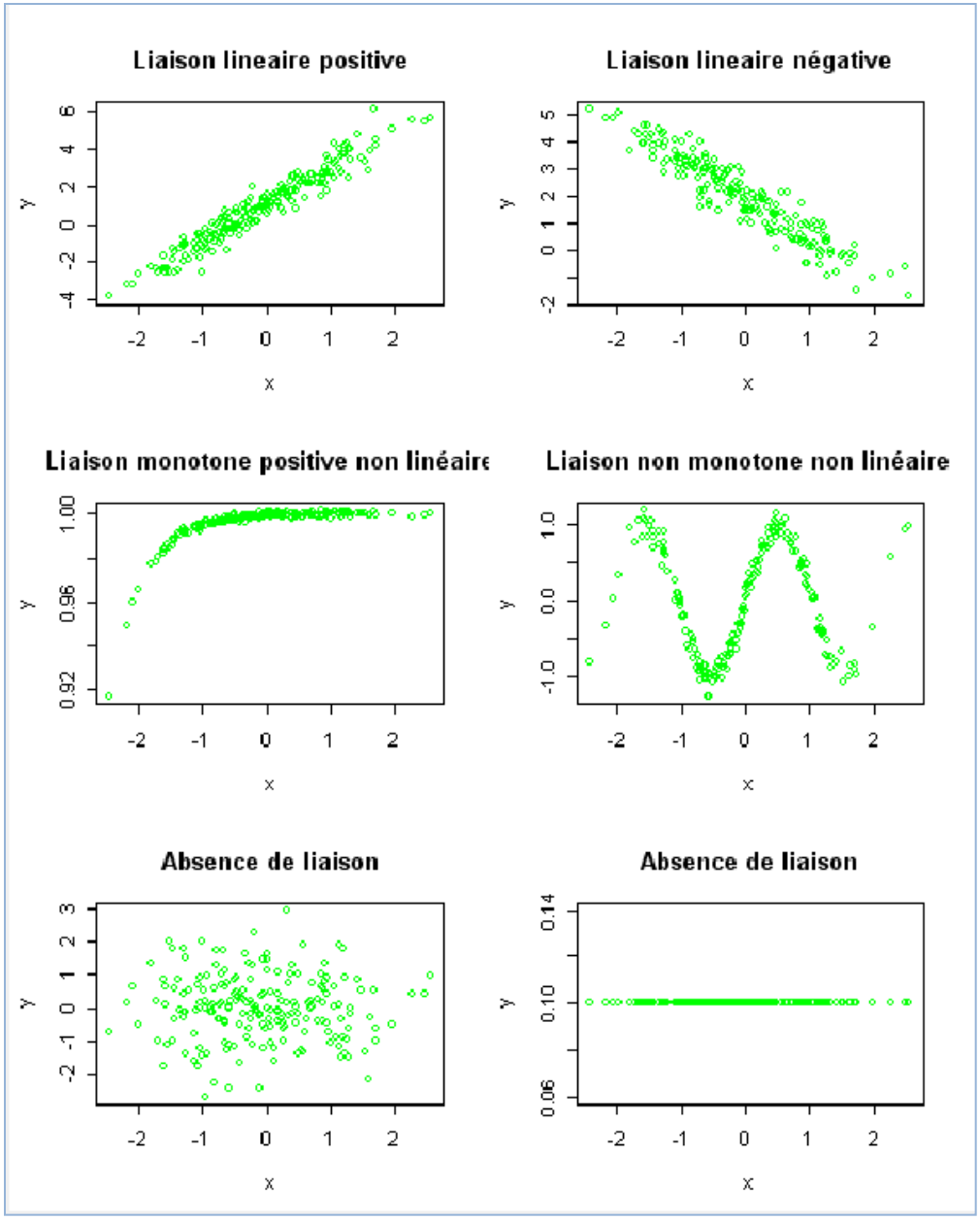
1. INVESTIGATING THE RELATIONSHIP BETWEEN TWO VARIABLES

Problem statement

X and Y are two quantitative variables, we want:

- to determine the existence of the relationship between X and Y ;
- to characterize the nature of the relationship ;
- to measure the strength of the relationship.

Note: the role of the variables is symmetrical, we do not seek to know if one determines the other or not (it will be the purpose of the regression analysis)



Visual inspection of scatterplots

Two viewpoints:

- in terms of variation i.e. when X increases, Y increases or decreases (keywords: linearity, monotonic relationship, positive or negative relationship);
- in terms of value i.e. when X is high, Y is high or low (but high/low compared with what?)

Notation

Variable: In capital letter (X is a variable)

Value: Observed value for the individual *i* (x_i) or at the date *t* (x_t)

Population: Statistical population Ω^{pop}

Sample: Ω (cases drawn from the population)

Size of the sample: n = card(Ω)

For the calculation: $\Omega = \{(x_i, y_i), i=1,\ldots,n\}$

Two statistics →

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample mean

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sample variance [we use "1/n" instead of "1/(n-1)"]

2. COVARIANCE AND CORRELATION

Covariance: definition, interpretation

E[X] is expected value of X

Definition :

$$COV(X,Y) = E\{(X - E[X])(Y - E[Y])\}$$
$$= E[XY] - E[X]E[Y]$$

Interpretation :

- It measures the tendency of the two variables to be simultaneously above or below their respective expected value.
- The reference is the expected values (mean) of variables.
- It characterizes monotonic and linear relationships.
- It gives indications about the direction of the relationship:

COV(X,Y) > 0, positive association ; COV(X,Y) < 0, negative

- About its strength: more greater is |COV|, more strong is the association
- COV(X,X) = V(X)

Covariance: properties

- **Symmetry**: COV(X,Y) = COV(Y,X)
- **Distributive**: COV(X,Y+Z) = COV(X,Y) + COV(X,Z)
- **Covariance with a constant**: COV(X,a) = 0
- **Covariance with a linear transformation of a variable**: COV(X,a+b.Y) = b.COV(X,Y)
- **Variance of the sum of two variables**: V(X+Y) = V(X) + V(Y) + 2.COV(X,Y)
- **Covariance of two independent variables**: COV(X,Y) = 0

See http://www.math.uah.edu/stat/expect/Covariance.html

Definition domain: $-\infty < COV < +\infty$

→ This is a non-normalized indicator (it depends on the variable's measurement unit)

Covariance: estimation on a sample

Sample covariance: $$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Biased estimator: $$E[S_{xy}] = \frac{n-1}{n} COV(X,Y)$$

Simplified formula: $$S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}.\bar{y}$$

Unbiased estimator: $$\hat{COV}(X,Y) = \frac{n}{n-1} S_{xy}$$

Covariance: estimation under Excel

Numero	Modele	Cylindree	Puissance	XY
1	Daihatsu Cuore	846	32	27072
2	Suzuki Swift 1.0 GLS	993	39	38727
3	Fiat Panda Mambo L	899	29	26071
4	VW Polo 1.4 60	1390	44	61160
5	Opel Corsa 1.2i Eco	1195	33	39435
6	Subaru Vivio 4WD	658	32	21056
7	Toyota Corolla	1331	55	73205
8	Opel Astra 1.6i 16V	1597	74	118178
9	Peugeot 306 XS 108	1761	74	130314
10	Renault Safrane 2.2. V	2165	101	218665
11	Seat Ibiza 2.0 GTI	1983	85	168555
12	VW Golt 2.0 GTI	1984	85	168640
13	Citroen ZX Volcane	1998	89	177822
14	Fiat Tempra 1.6 Liberty	1580	65	102700
15	Fort Escort 1.4i PT	1390	54	75060
16	Honda Civic Joker 1.4	1396	66	92136
17	Volvo 850 2.5	2435	106	258110
18	Ford Fiesta 1.2 Zetec	1242	55	68310
19	Hyundai Sonata 3000	2972	107	318004
20	Lancia K 3.0 LS	2958	150	443700
21	Mazda Hachtback V	2497	122	304634
22	Mitsubishi Galant	1998	66	131868
23	Opel Omega 2.5i V6	2496	125	312000
24	Peugeot 806 2.0	1998	89	177822
25	Nissan Primera 2.0	1997	92	183724
26	Seat Alhambra 2.0	1984	85	168640
27	Toyota Previa salon	2438	97	236486
28	Volvo 960 Kombi aut	2473	125	309125

n		Moyenne		Somme
28		1809.07	77.71	4451219

Cov.Empirique	**18381.4133**
Cov.Non-Biaisé	**19062.2063**

Cov.Excel	**18381.4133**

Excel function

Pearson product-moment correlation coefficient

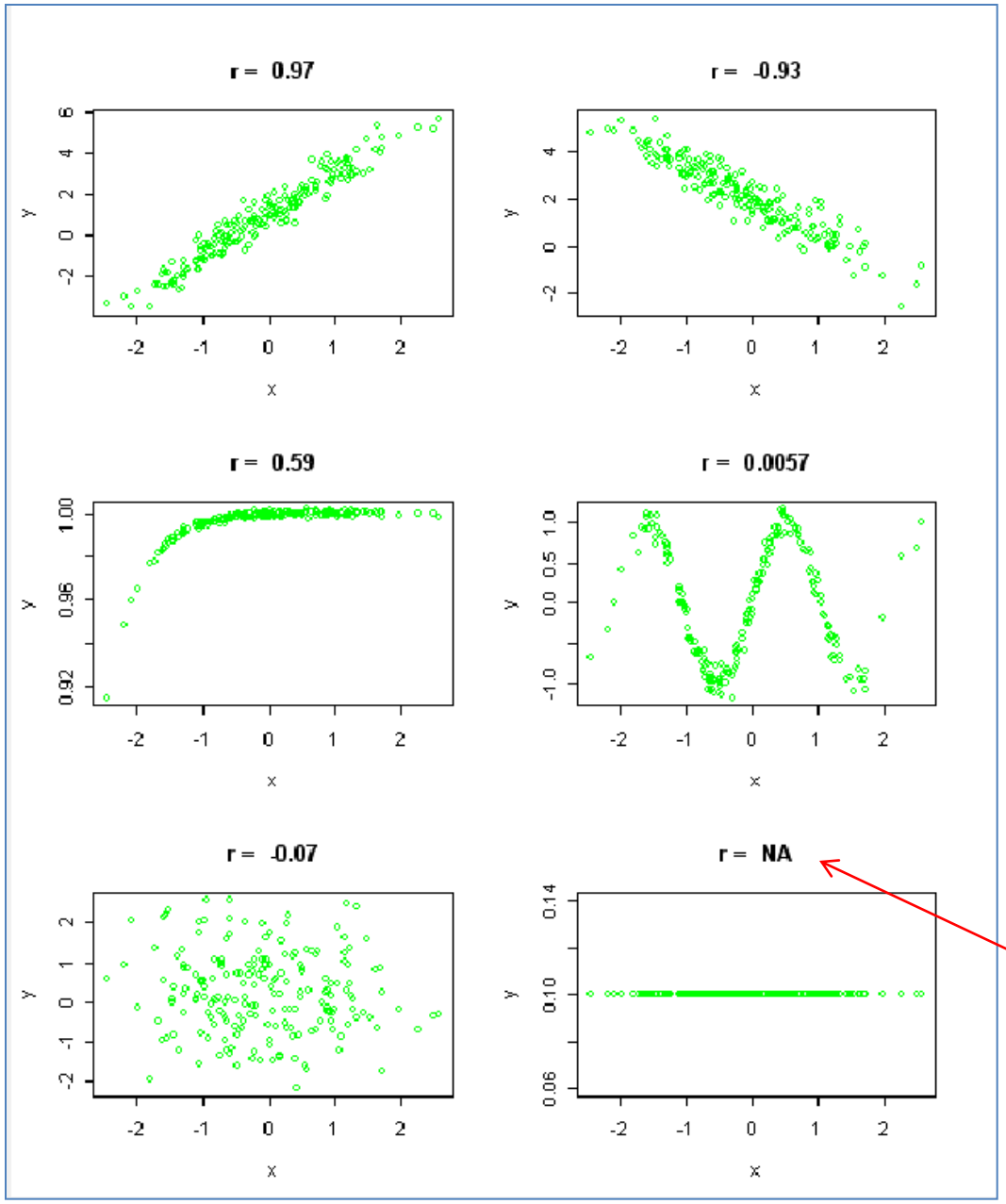
Normalized version of the covariance (divided by the product of standard deviations)

Definition: $$r_{xy} = \frac{COV(X,Y)}{\sqrt{V(X).V(Y)}} = \frac{COV(X,Y)}{\sigma_x.\sigma_y}$$

Normalized measure: $$-1 \leq r_{xy} \leq +1$$

Some remarks:

- It measures the linear (monotonic) relationship between 2 variables
- (X,Y) independents $\Rightarrow$ r = 0 (the reverse is false in general)
- Correlation of a variable with itself: $r_{xx} = 1$
- Correlation = Covariance for standardized variables = Expectation of the product of the standardized variables



Correlation: visual inspection and correlation coefficient r

Important points: monotonic, linear…

why the calculation has failed here?

Correlation: estimation on a sample

Sample correlation:

$$\hat{r} = \frac{S_{xy}}{s_x . s_y}$$

Biased estimator:

$$E[\hat{r}] = r - \frac{r(1-r^2)}{2n}$$

Asymptotically unbiased

Unbiased estimator:

$$\hat{r}_{aj} = \sqrt{1 - \frac{n-1}{n-2}\left(1-\hat{r}^2\right)}$$

Not used in practice, the bias is negligible when n increases.

Correlation: calculation with Excel

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

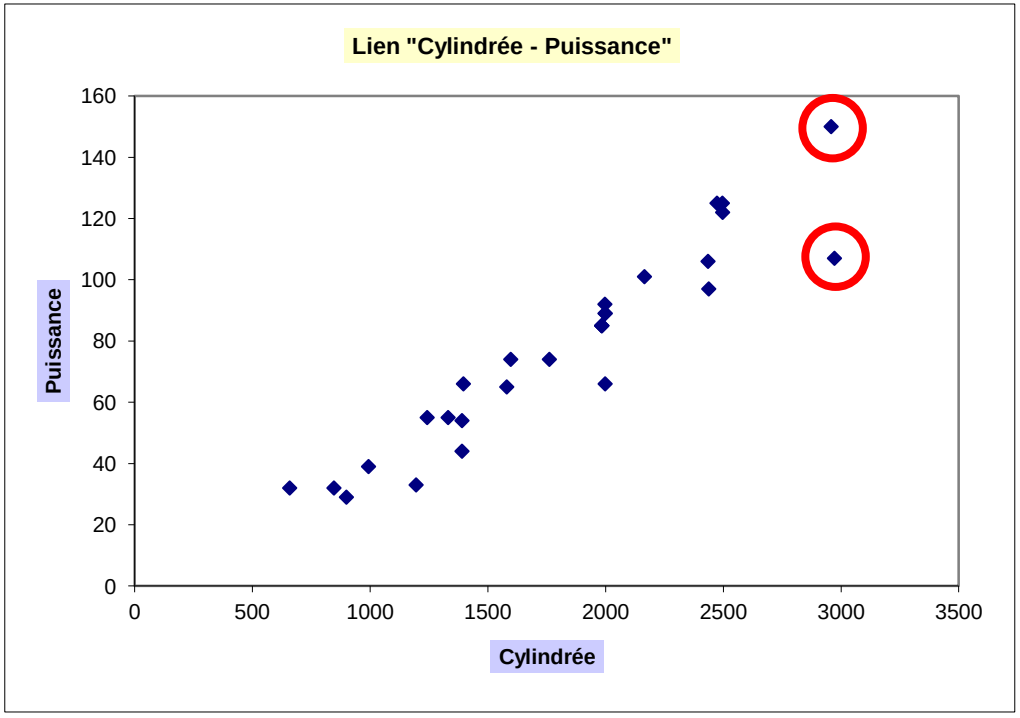
$$= \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2} \times \sqrt{\sum_i y_i^2 - n\bar{y}^2}}$$

Numero	Modele	Cylindree	Puissance	XY	X²	Y²
1	Daihatsu Cuore	846	32	27072	715716	1024
2	Suzuki Swift 1.0 GLS	993	39	38727	986049	1521
3	Fiat Panda Mambo L	899	29	26071	808201	841
4	VW Polo 1.4 60	1390	44	61160	1932100	1936
5	Opel Corsa 1.2i Eco	1195	33	39435	1428025	1089
6	Subaru Vivio 4WD	658	32	21056	432964	1024
7	Toyota Corolla	1331	55	73205	1771561	3025
8	Opel Astra 1.6i 16V	1597	74	118178	2550409	5476
9	Peugeot 306 XS 108	1761	74	130314	3101121	5476
10	Renault Safrane 2.2. V	2165	101	218665	4687225	10201
11	Seat Ibiza 2.0 GTI	1983	85	168555	3932289	7225
12	VW Golt 2.0 GTI	1984	85	168640	3936256	7225
13	Citroen ZX Volcane	1998	89	177822	3992004	7921
14	Fiat Tempra 1.6 Liberty	1580	65	102700	2496400	4225
15	Fort Escort 1.4i PT	1390	54	75060	1932100	2916
16	Honda Civic Joker 1.4	1396	66	92136	1948816	4356
17	Volvo 850 2.5	2435	106	258110	5929225	11236
18	Ford Fiesta 1.2 Zetec	1242	55	68310	1542564	3025
19	Hyundai Sonata 3000	2972	107	318004	8832784	11449
20	Lancia K 3.0 LS	2958	150	443700	8749764	22500
21	Mazda Hachtback V	2497	122	304634	6235009	14884
22	Mitsubishi Galant	1998	66	131868	3992004	4356
23	Opel Omega 2.5i V6	2496	125	312000	6230016	15625
24	Peugeot 806 2.0	1998	89	177822	3992004	7921
25	Nissan Primera 2.0	1997	92	183724	3988009	8464
26	Seat Alhambra 2.0	1984	85	168640	3936256	7225
27	Toyota Previa salon	2438	97	236486	5943844	9409
28	Volvo 960 Kombi aut	2473	125	309125	6115729	15625
n		**Moyenne**			**Somme**	
28		1809.07	77.71	4451219	102138444	197200

Numérateur	514679.571
Dénominateur	543169.291
Corrélation	**0.9475**

Excel function → Coef.Corr.Excel 0.9475

Correlation: visual inspection



A statistical indicator gives only one point of view, a graphical analysis is also essential (e.g. to identify unusual situations, outliers, etc.)

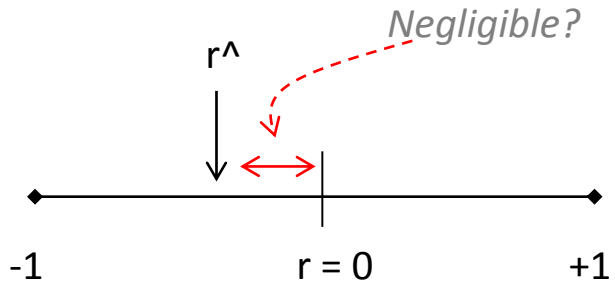
3. SIGNIFICANCE TESTING

Significance testing (1)

Testing the existence of a linear relationship between X and Y

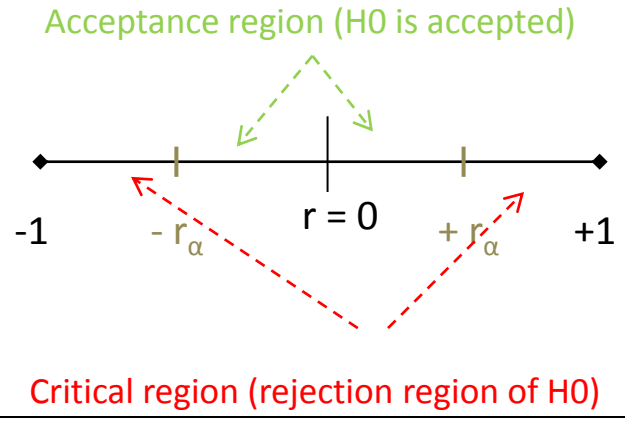
$$\begin{cases} H_0 : r = 0 \\ H_1 : r \neq 0 \end{cases}$$

(X,Y) independents → r = 0; but r = 0 does not mean that (X,Y) are independent, it means that there is not a linear relationship.



How to proceed?

- We want to know if r is significantly different to 0.
- We calculate a sample estimate of r (r^).
- In order to define the thresholds around 0, we specify the alpha level P(reject H0/H0 is true in fact i.e. r = 0), and we obtain r_α
- But for that, we must know the sampling distribution of r^ under H0



Significance testing (2)

Student's t–distribution

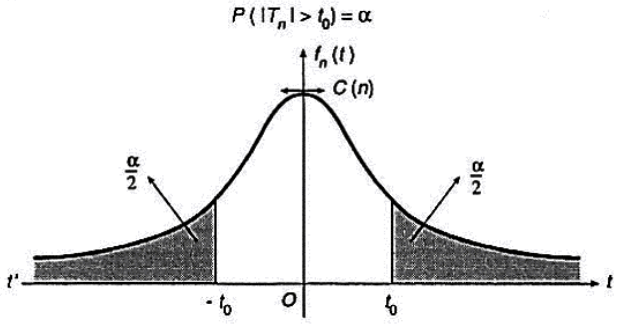
$$t = \frac{\hat{r}}{\sqrt{\frac{1-\hat{r}^2}{n-2}}} \equiv \Im(n-2)$$

Idea: Under H0, the sampling distribution of r^ is unknown, but we can know that of a transformed value of r^

The decision rule becomes:

$$\text{Accept } H_0 (r = 0) \text{ if } |t| < t_{1-\alpha/2}$$

$$\text{Reject } H_0 (r \neq 0) \text{ if } |t| \geq t_{1-\alpha/2}$$



Note:

• Some tools provide often the p-value (probability value, it reflects the strength of the evidence against the null hypothesis)

• The Student's t-distribution is only true in the neighborhood of H0 (r = 0), we cannot use it for other tests (e.g. H0: r = a, where a ≠ 0) or for the calculation of confidence intervals.

Significance testing – An example

Numero	Modele	Cylindree	Puissance
1	Daihatsu Cuore	846	32
2	Suzuki Swift 1.0 GLS	993	39
3	Fiat Panda Mambo L	899	29
4	VW Polo 1.4 60	1390	44
5	Opel Corsa 1.2i Eco	1195	33
6	Subaru Vivio 4WD	658	32
7	Toyota Corolla	1331	55
8	Opel Astra 1.6i 16V	1597	74
9	Peugeot 306 XS 108	1761	74
10	Renault Safrane 2.2. V	2165	101
11	Seat Ibiza 2.0 GTI	1983	85
12	VW Golt 2.0 GTI	1984	85
13	Citroen ZX Volcane	1998	89
14	Fiat Tempra 1.6 Liberty	1580	65
15	Fort Escort 1.4i PT	1390	54
16	Honda Civic Joker 1.4	1396	66
17	Volvo 850 2.5	2435	106
18	Ford Fiesta 1.2 Zetec	1242	55
19	Hyundai Sonata 3000	2972	107
20	Lancia K 3.0 LS	2958	150
21	Mazda Hachtback V	2497	122
22	Mitsubishi Galant	1998	66
23	Opel Omega 2.5i V6	2496	125
24	Peugeot 806 2.0	1998	89
25	Nissan Primera 2.0	1997	92
26	Seat Alhambra 2.0	1984	85
27	Toyota Previa salon	2438	97
28	Volvo 960 Kombi aut	2473	125

r^	0.9475
n	28
ddl (n-2)	26

Test de significativité

t	15.1171
t-théorique (5%)	2.0555
p-value	2.14816E-14

$$t = \frac{\hat{r}}{\sqrt{\frac{1-\hat{r}^2}{n-2}}} = \frac{0.9475}{\sqrt{\frac{1-0.9475^2}{28-2}}} = 15.1171$$

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0.975}(26) = 2.0555$$

Conclusion: we reject the null hypothesis i.e. the correlation is significant at the α = 5% level.

4. CONFIDENCE INTERVAL

Confidence interval

Issue: r^ is a point estimate i.e. it depends on the sample used. With another sample, we obtain another value, slightly (or quite) different.

Solution: A confidence interval is an interval estimate. If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter.

→ To calculate the confidence interval, we must know the sampling distribution of r^, whatever the true value of r.

→ The Student's t distribution is no longer appropriate, it is valid only if "r = 0"

We use another transformation i.e. the "Fisher transformation"

$$\hat{z} = \frac{1}{2} \ln \frac{1+\hat{r}}{1-\hat{r}}$$

z^ follows a normal distribution whatever the true value of r. It can be used for any hypothesis testing, it can be use also for the confidence interval calculation.

With:

$$E[\hat{z}] \approx \frac{1}{2} \ln \frac{1+r}{1-r}$$

$$V[\hat{z}] \approx \frac{1}{n-3}$$

Confidence interval: an example

Numero	Modele	Cylindree	Puissance
1	Daihatsu Cuore	846	32
2	Suzuki Swift 1.0 GLS	993	39
3	Fiat Panda Mambo L	899	29
4	VW Polo 1.4 60	1390	44
5	Opel Corsa 1.2i Eco	1195	33
6	Subaru Vivio 4WD	658	32
7	Toyota Corolla	1331	55
8	Opel Astra 1.6i 16V	1597	74
9	Peugeot 306 XS 108	1761	74
10	Renault Safrane 2.2. V	2165	101
11	Seat Ibiza 2.0 GTI	1983	85
12	VW Golt 2.0 GTI	1984	85
13	Citroen ZX Volcane	1998	89
14	Fiat Tempra 1.6 Liberty	1580	65
15	Fort Escort 1.4i PT	1390	54
16	Honda Civic Joker 1.4	1396	66
17	Volvo 850 2.5	2435	106
18	Ford Fiesta 1.2 Zetec	1242	55
19	Hyundai Sonata 3000	2972	107
20	Lancia K 3.0 LS	2958	150
21	Mazda Hachtback V	2497	122
22	Mitsubishi Galant	1998	66
23	Opel Omega 2.5i V6	2496	125
24	Peugeot 806 2.0	1998	89
25	Nissan Primera 2.0	1997	92
26	Seat Alhambra 2.0	1984	85
27	Toyota Previa salon	2438	97
28	Volvo 960 Kombi aut	2473	125

r^	0.9475
n	28

Calcul de z

z	1.8072
Variance(z)	0.0400
Ecart type(z)	0.2000

Quantile 0.975 - Loi normale

u(0.975)	1.9600

Intervalle de conf. pour z^

bb(z)	1.4152
bh(z)	2.1992

Intervalle de conf. pour r^

bb(r)	0.8886
bh(r)	0.9757

Steps:

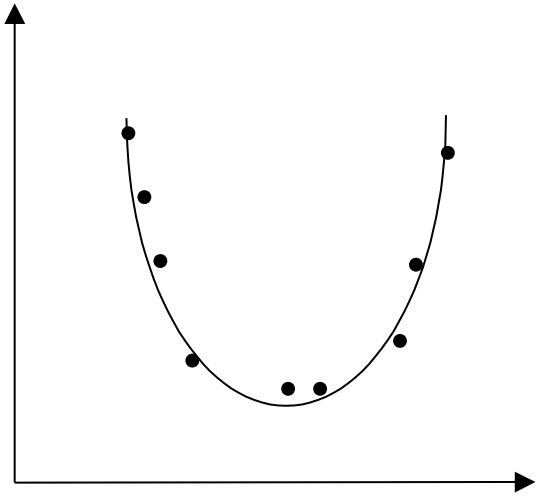
1. We calculate r^
2. We use the Fisher transformation (z^)
3. We calculate the confidence interval of z for (1-α) confidence level.
4. We transform the confidence limits of z into confidence limits for r (using the inverse of the transformation).

→ There is a 95% chance that this interval (0.8886 ; 0.9757) contains the true value of r.

5. PROBLEMATIC SITUATIONS

Non linear relationship - Variable transformations

Non linear, non monotonic relationship



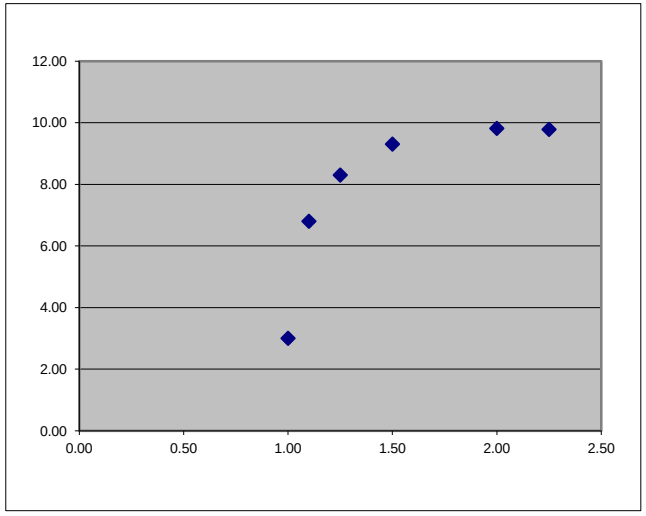
Linearization by variable transformations (e.g. $Z = X^2$)

Y	X	X^2
9.31	-3	9
4.14	-2	4
1.04	-1	1
0.45	0	0
1.47	1	1
4.82	2	4
9.42	3	9
Corrélation (Y,X)	0.04369908	
Corrélation (Y,X^2)	0.99772156	

Determining the appropriate function for the variable transformation is not easy in general.

Non linear but monotonic relationship - Using the rank transformation

Non linear but monotonic relationship



Transform the data into ranks

X	Y	RX	RY
1.00	3.00	1	1
1.10	6.80	2	2
1.25	8.30	3	3
1.50	9.30	4	4
2.00	9.81	5	6
2.25	9.78	6	5
Corrélation (XY)	0.77588403		0.94285714

→ The Pearson correlation coefficient computed on ranked variables is the "Spearman's rank correlation coefficient"

→ The inferential procedures (hypothesis testing, confidence intervals) remain valid.

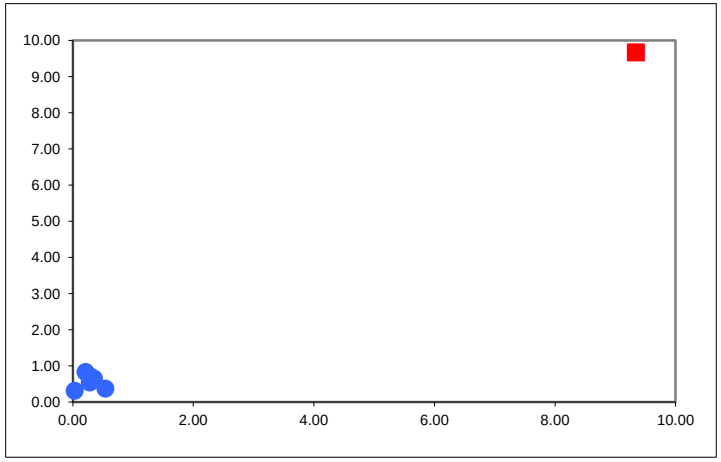
→ This method is not useful for characterizing non monotonic relationship

age	rang moyen	rang aléatoire
15	4	**5**
18	7	7
12	1	1
13	2	2
15	4	**3**
16	6	6
15	4	**4**

In case of ties:

- random ranks (simple but not powerful)
- average the ranks (need more calculations but more powerful)

Outliers problem



The correlation sample estimate is very sensitive to outliers.

	X	Y
1	0.30	0.70
2	0.35	0.65
3	0.54	0.37
4	0.28	0.54
5	0.21	0.83
6	0.03	0.31
7	9.34	9.67

r (6 points)	**0.0185**
r (7 points)	**0.9976**

Transform data into ranks

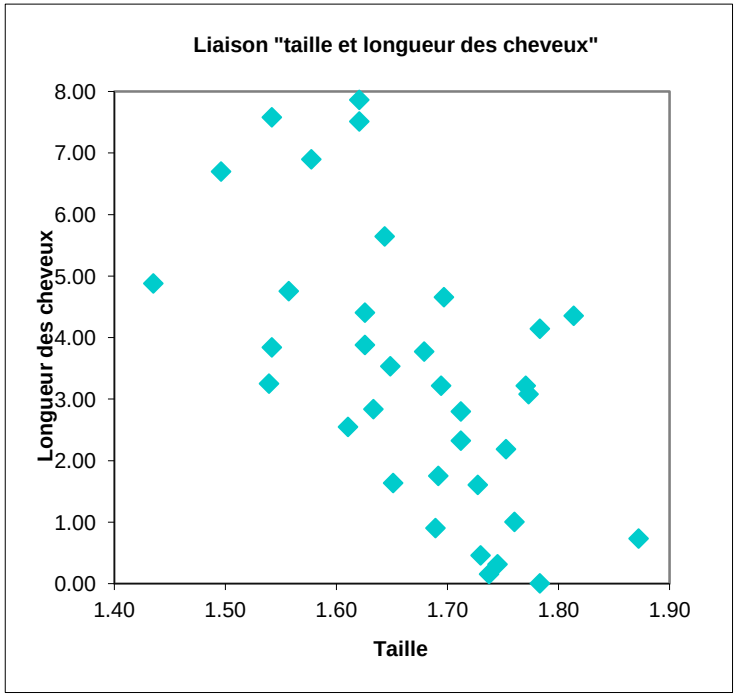
	X	Y	RX	RY
1	0.30	0.70	4	5
2	0.35	0.65	5	4
3	0.54	0.37	6	2
4	0.28	0.54	3	3
5	0.21	0.83	2	6
6	0.03	0.31	1	1
7	9.34	9.67	7	7

Coef. Rang: 0.39285714

The Spearman's rank correlation is not (less) sensitive to outliers.

6. PARTIAL CORRELATION

Some correlations seem mysterious



Length of the hair

Height of the individual

Who can believe that there is a negative relationship between the height of individuals (X) and the length of hair (Y)?

There surely has a third variable (Z) which simultaneously influences X and Y.

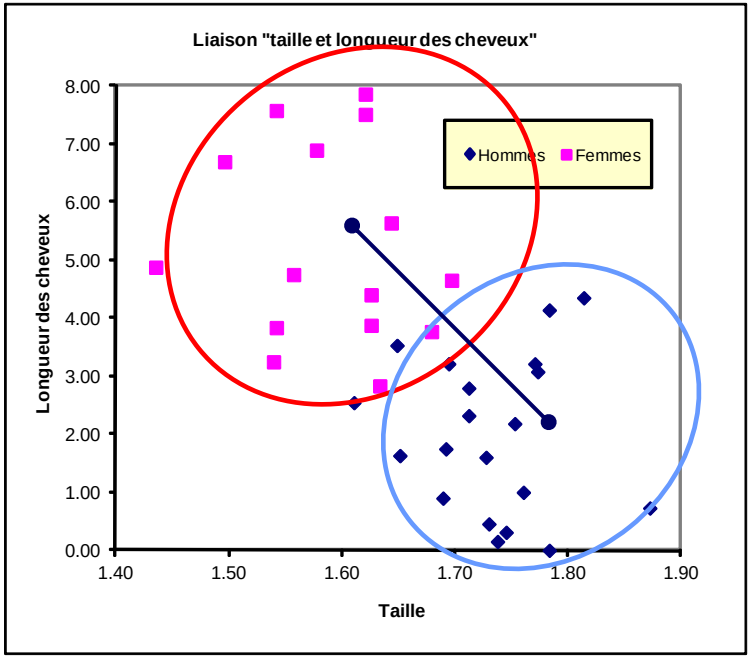
And, in fact, the relationship between Y and X is essentially determined by Z.

Special case: Z is a binary variable (e.g. Sex)

		Cheveux (cm)	Taille (m)
Hommes	1	1.64	1.65
	2	0.32	1.74
	3	1.00	1.76
	4	2.80	1.71
	5	4.35	1.81
	6	2.33	1.71
	7	0.01	1.78
	8	1.75	1.69
	9	3.22	1.77
	10	3.53	1.65
	11	2.55	1.61
	12	3.08	1.77
	13	0.46	1.73
	14	3.22	1.69
	15	2.19	1.75
	16	0.73	1.87
	17	0.16	1.74
	18	0.90	1.69
	19	4.14	1.78
	20	1.61	1.73
Femmes	1	4.66	1.70
	2	3.25	1.54
	3	3.88	1.63
	4	2.84	1.63
	5	4.88	1.44
	6	3.77	1.68
	7	5.64	1.64
	8	4.41	1.63
	9	3.84	1.54
	10	7.58	1.54
	11	7.51	1.62
	12	6.90	1.58
	13	4.76	1.56
	14	6.70	1.50
	15	7.86	1.62

r (hommes)	-0.074
r (femmes)	-0.141

r (global)	-0.602



The computed correlation is essentially influenced by the difference between the conditional centroids.

The within-group correlation is very weak.

Partial correlation – Z is a quantitative variable

We remove the effect of z on x **and** on y

Correlation between (y, x)

Partial correlation coefficient (correlation between X and Y, by controlling [removing] the effect of Z)

$$r_{xy.z} = \frac{r_{xy} - r_{xz} \times r_{yz}}{\sqrt{(1-r_{xz}^2)}\sqrt{(1-r_{yz}^2)}}$$

Normalization so that $-1 \leq r_{xy.z} \leq +1$

Estimation: we use the sample correlation estimates

$$\hat{r}_{xy.z} = \frac{\hat{r}_{xy} - \hat{r}_{xz} \times \hat{r}_{yz}}{\sqrt{(1-\hat{r}_{xz}^2)}\sqrt{(1-\hat{r}_{yz}^2)}}$$

p^{th}-order partial correlation ($p > 1$): recursive formula

$$r_{xy.zw} = \frac{r_{xy.z} - r_{xw.z} \times r_{yw.z}}{\sqrt{(1-r_{xw.z}^2)}\sqrt{(1-r_{yw.z}^2)}}$$

p = 2 here (Z and W are the controlling variables)

Partial correlation – An example

$$\hat{r}_{xy.w} = \frac{0.8878 - 0.9475 \times 0.8919}{\sqrt{(1-0.9475^2)}\sqrt{(1-0.8919^2)}} = 0.2955$$

		X	Y	W
Numero	Modele	Puissance	Conso	Cylindree
1	Daihatsu Cuore	32	5.7	846
2	Suzuki Swift 1.0 GLS	39	5.8	993
3	Fiat Panda Mambo L	29	6.1	899
4	VW Polo 1.4 60	44	6.5	1390
5	Opel Corsa 1.2i Eco	33	6.8	1195
6	Subaru Vivio 4WD	32	6.8	658
7	Toyota Corolla	55	7.1	1331
8	Opel Astra 1.6i 16V	74	7.4	1597
9	Peugeot 306 XS 108	74	9.0	1761
10	Renault Safrane 2.2. V	101	11.7	2165
11	Seat Ibiza 2.0 GTI	85	9.5	1983
12	VW Golt 2.0 GTI	85	9.5	1984
13	Citroen ZX Volcane	89	8.8	1998
14	Fiat Tempra 1.6 Liberty	65	9.3	1580
15	Fort Escort 1.4i PT	54	8.6	1390
16	Honda Civic Joker 1.4	66	7.7	1396
17	Volvo 850 2.5	106	10.8	2435
18	Ford Fiesta 1.2 Zetec	55	6.6	1242
19	Hyundai Sonata 3000	107	11.7	2972
20	Lancia K 3.0 LS	150	11.9	2958
21	Mazda Hachtback V	122	10.8	2497
22	Mitsubishi Galant	66	7.6	1998
23	Opel Omega 2.5i V6	125	11.3	2496
24	Peugeot 806 2.0	89	10.8	1998
25	Nissan Primera 2.0	92	9.2	1997
26	Seat Alhambra 2.0	85	11.6	1984
27	Toyota Previa salon	97	12.8	2438
28	Volvo 960 Kombi aut	125	12.7	2473

n	28

Corrélations brutes		
Puissance	Conso	0.88781
Puissance	Cylindrée	0.94755
Conso	Cylindrée	0.89187

Corrélation partielle	
r_xy.z	0.29553

Test de significativité	
t	1.54673
t(0.975 ; 25)	2.38461
p-value	0.13450

Intervalle de confiance à 95%	
z	0.30461
e.t.	0.20412
u(0.975)	1.95996
bb(z)	-0.09546
bh(z)	0.70469
bb (r)	-0.09517
bh (r)	0.60734

Significance testing

$$t = \frac{\hat{r}_{xy.z}}{\sqrt{\frac{1-\hat{r}^2_{xy.w}}{n-p-2}}} \equiv \Im(n-p-2)$$

Confidence interval (using the Fisher transformation)

$$\hat{z} = \frac{1}{2}\ln\frac{1+\hat{r}}{1-\hat{r}}$$

normally distributed…

$$E[\hat{z}] \approx \frac{1}{2}\ln\frac{1+r}{1-r}$$

$$V[\hat{z}] \approx \frac{1}{n-p-3}$$

References

- HSC Learning Repository, University of the West of England, 2014.
- L. Simon, STAT 501, "Regression Methods", PennState University.
- M. Plonsky, "Correlation", Psychological Statistics, 2014.