Data analysis and Geostatistics

Short Course on the use of statistical techniques in the geosciences

Vincent van Hinsberg • McGill University

Geotop Short Course in Data Analysis and Geostatistics

Time series analysis

Time series analysis

Time is a critical variable in geology and a whole subfield of geostatistics is devoted to it: time series analysis

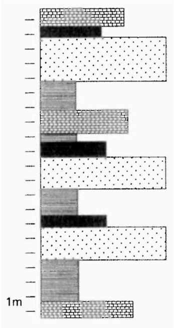
aims: detect trends and systematics with time for process identification and to predict the future

time is only rarely absolute, in most cases we have only qualitative information on time (strat sequence, growth zoning, younger-older)

Time series analysis - Markov chain

Systematics in the lithology changes for a log (time is qualitative)

strat log



transition matrix

to

from

	sh	s	c	l	total
sh	0	3	0	1	4
s	0	0	3	0	3
c	2	0	0	1	3
l	2	0	0	0	2
total	4	3	3	2	12

transition prob matrix

to

from

	sh	s	c	l	total
sh	0	3/4	0	1/4	1
s	0	0	1	0	1
c	2/3	0	0	1/3	1
l	1	0	0	0	1
total					

find systematic chains:

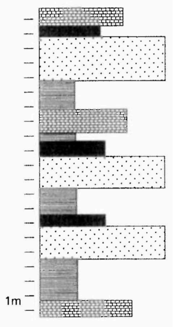
sh —> s —> c —> sh
l —> sh —> s —> c —> l

are these sequences significant or pure chance ?

Time series analysis - Markov chain

Systematics in the lithology changes for a log (time is qualitative)

strat log



transition matrix

to

from

	sh	s	c	l	total
sh	4	3	0	1	8
s	0	4	3	0	7
c	2	0	0	1	3
l	2	0	0	0	2
total	8	7	3	2	20

transition prob matrix

to

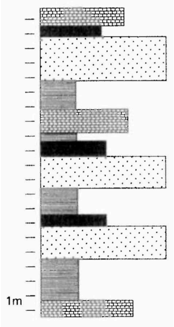
from

	sh	s	c	l	total
sh	1/2	3/8	0	1/8	1
s	0	4/7	3/7	0	1
c	2/3	0	0	1/3	1
l	1	0	0	0	1
total					

Time series analysis - Markov chain

Systematics in the lithology changes for a log (time is qualitative)

strat log



transition matrix

to

from

	sh	s	c	l	total
sh	4	3	0	1	8
s	0	4	3	0	7
c	2	0	0	1	3
l	2	0	0	0	2
total	8	7	3	2	20

transition prob matrix

to

from

	sh	s	c	l	total
sh	1/2	3/8	0	1/8	1
s	0	4/7	3/7	0	1
c	2/3	0	0	1/3	1
l	1	0	0	0	1
total					

to

from

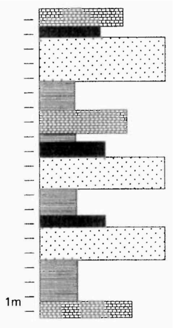
	sh	s	c	l	total
sh	8/20	7/20	3/20	2/20	1
s	8/20	7/20	3/20	2/20	1
c	8/20	7/20	3/20	2/20	1
l	8/20	7/20	3/20	2/20	1
total					

random
transition
prob matrix

Time series analysis - Markov chain

Systematics in the lithology changes for a log (time is qualitative)

strat log



transition matrix

to

from

	sh	s	c	l	total
sh	4	3	0	1	8
s	0	4	3	0	7
c	2	0	0	1	3
l	2	0	0	0	2
total	8	7	3	2	20

random transition matrix

to

from

	sh	s	c	l	total
sh	3.2	2.8	1.2	0.8	8
s	2.8	2.45	1.05	0.7	7
c	1.2	1.05	0.45	0.3	3
l	0.8	0.7	0.3	0.2	2
total	8	7	3	2	20

random
transition
prob matrix

to

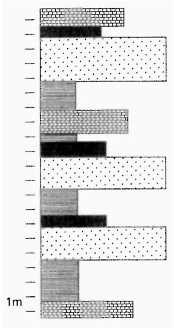
from

	sh	s	c	l	total
sh	8/20	7/20	3/20	2/20	1
s	8/20	7/20	3/20	2/20	1
c	8/20	7/20	3/20	2/20	1
l	8/20	7/20	3/20	2/20	1
total					

Time series analysis - Markov chain

Systematics in the lithology changes for a log (time is qualitative)

strat log



transition matrix

to

from

	sh	s	c	l	total
sh	4	3	0	1	8
s	0	4	3	0	7
c	2	0	0	1	3
l	2	0	0	0	2
total	8	7	3	2	20

observed

random transition matrix

to

from

	sh	s	c	l	total
sh	3.2	2.8	1.2	0.8	8
s	2.8	2.45	1.05	0.7	7
c	1.2	1.05	0.45	0.3	3
l	0.8	0.7	0.3	0.2	2
total	8	7	3	2	20

expected

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

X^2_{calc} = 19.23

df = (c -1)(r-1) = 9

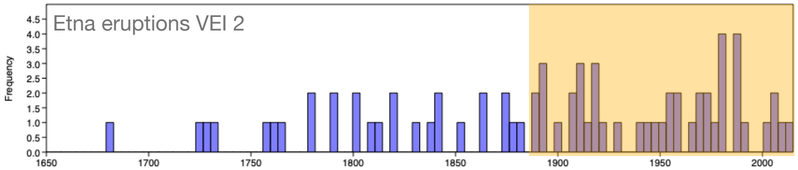
$X^2_{0.05:9}$ = 16.92 : calc exceed critical : not random

Time series analysis - randomness of

same number of eruptions through time

however, non uniform spacing in time?

The past is the key to the future: but only if the past was non-random !



periode 1890 - 2015: 44 eruptions

split this up in 25 year time periods:

	obs	exp
1890 - 1915:	11	8.8
1915 - 1940:	5	8.8
1940 - 1965:	8	8.8
1965 - 1990:	14	8.8
1990 - 2015:	6	8.8

$X^2_{calc} = 6.23$

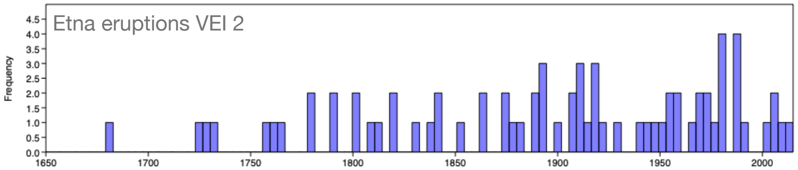
$X^2_{0.05:4} = 9.488$

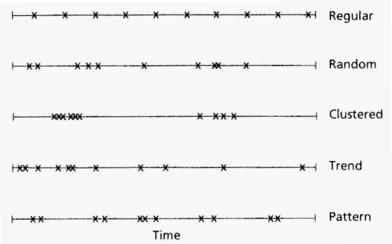
d.f. = class - 1 = 5 - 1 = 4

✗

Time series analysis - randomness of events

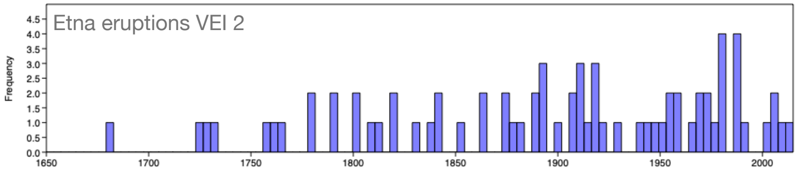
The past is the key to the future: but only if the past was non-random !





Time series analysis - randomness of events

The past is the key to the future: but only if the past was non-random !



periode 1890 - 2015: 44 eruptions, 43 intervals

time between eruptions:

	obs
<2:	19
2-4:	14
5-7:	5
8-10:	5

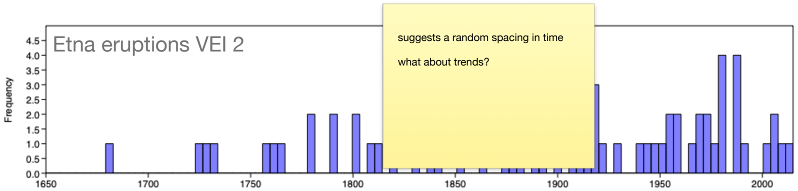
we will calculate the expected random occurrence from the Poisson distribution (2.3.7.3):

$E_j = T \cdot e^{(-n/T)} \cdot (n/T)^j / j!$

where n = total no. events = 44, T = no. intervals = 43

Time series analysis - randomness of events

The past is the key to the future: but only if the past was non-random !



suggests a random spacing in time

what about trends?

periode 1890 - 2015: 44 eruptions, 43 intervals

time between eruptions:

	obs	exp	exp
<2:	19	15.46	15.46
2-4:	14	15.81	15.81
5-7:	5	8.09	8.09
8-10:	5	2.76	3.61
11-13:		0.71	
14-16:		0.14	

$X^2_{calc} = 2.74$

d.f. = class - 1 = 3

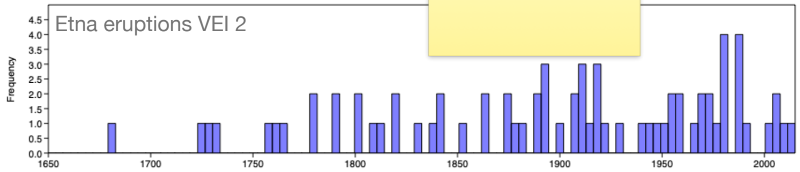
$X^2_{0.05:3} = 7.815$

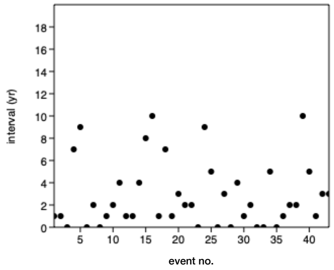
✗

Time series analysis - randomness of events

The past is the key to the future: but c non-random !

the occurrence of an eruption has no info on occurrence of another





event number does not have a normal distribution: robust test

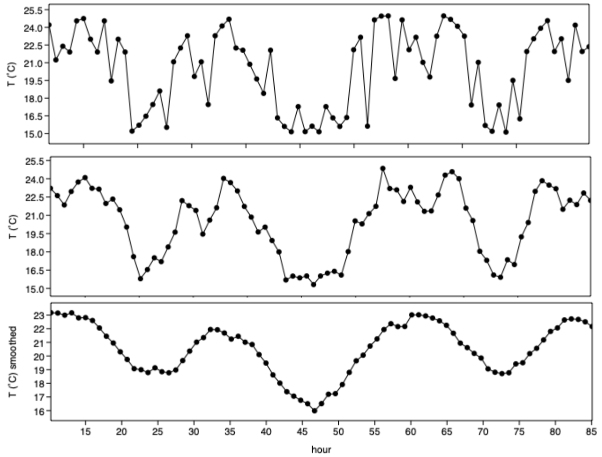
Spearman r analysis: H_o = no trend

H_A = trend

r′ = 0.07, p_{uncorr} = 0.65

Time series analysis - systematics with time

Time series data consists of noise + signal: smoothing allows for noise to be reduced by assuming its frequency to be different

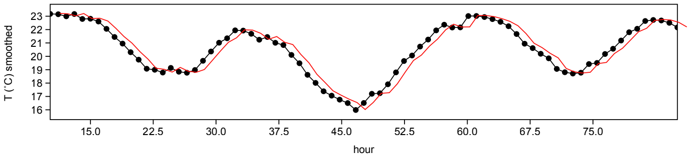


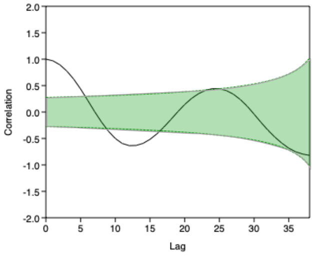
smoothing:
3 point average

smoothing:
weighted
polynomial

Time series analysis - systematics with time

Assessing periodicity in your time series: auto-correlation

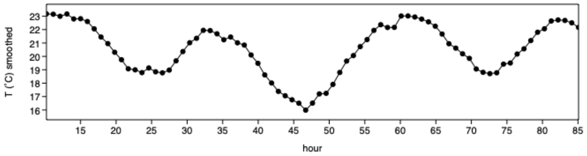




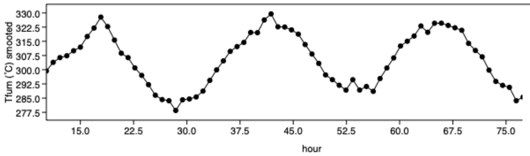
95% confidence window

Time series analysis - systematics with time

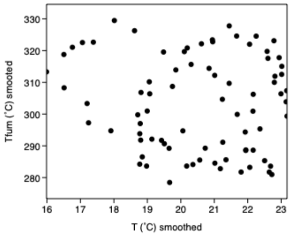
Analysing multiple variables against time: cross-correlation



air temperature



fumarole temperature

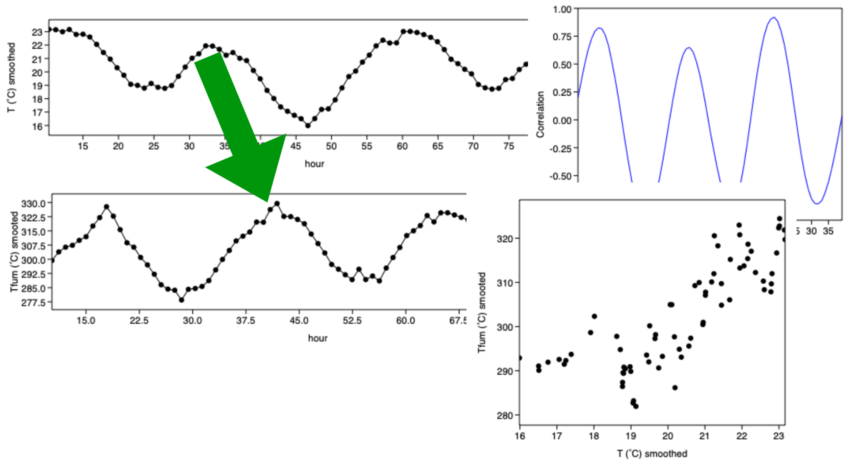


r = -0.09, p_{uncorr} = 0.41

r′ = -0.07, p_{uncorr} = 0.57

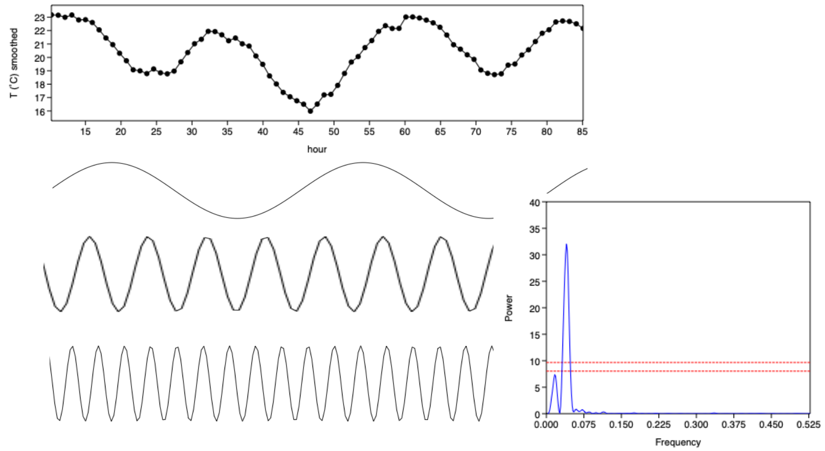
Time series analysis - systematics with time

Analysing multiple variables against time: cross-correlogram



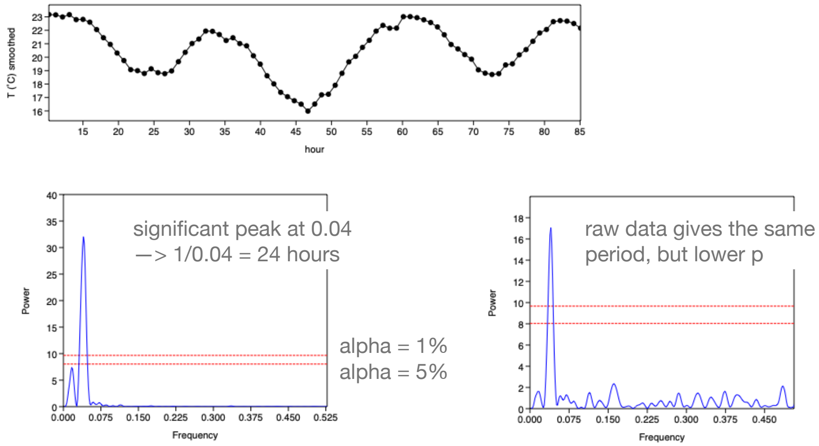
Time series analysis - periodicity

Analysing variables against time: periodogram



Time series analysis - periodicity

Analysing variables against time: periodogram

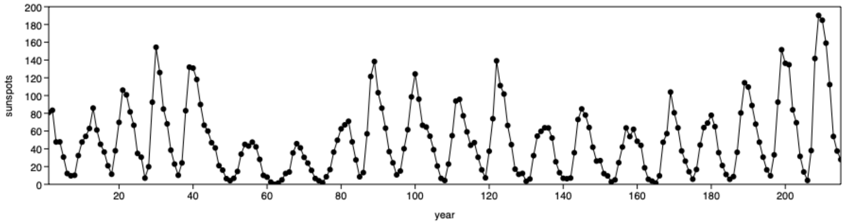


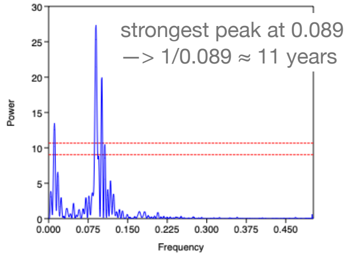
significant peak at 0.04
—> 1/0.04 = 24 hours

alpha = 1%
alpha = 5%

raw data gives the same
period, but lower p

Time series analysis - periodicity





strongest peak at 0.089
—> 1/0.089 ≈ 11 years

This plot is essentially a histogram of
the sum of squares for each frequency:

the amount of variance in the time
series explained by each frequency

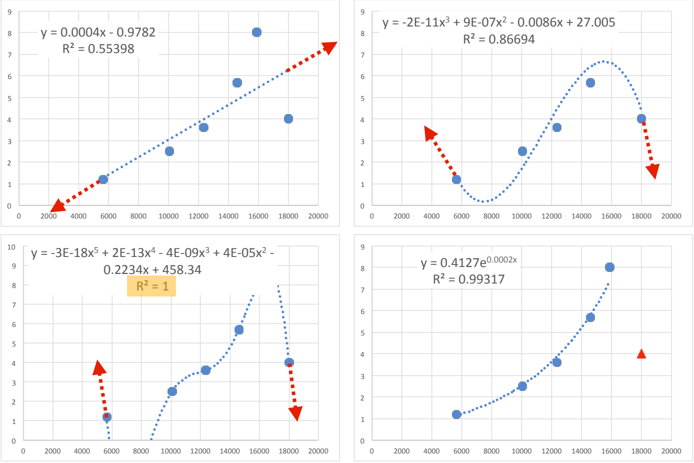
Geotop Short Course in Data Analysis and Geostatistics

Regression analysis and curve fitting

geotop

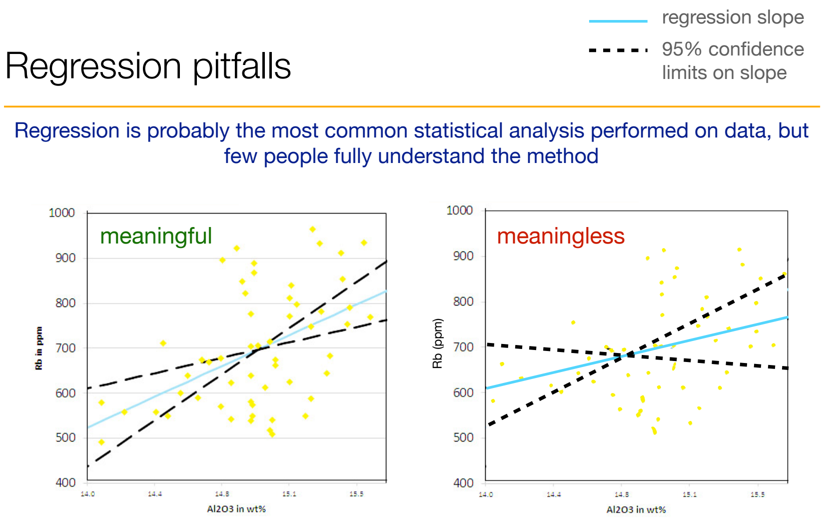
Regression pitfalls

Regression is probably the most common statistical analysis performed on data, but few people fully understand the method



Regression pitfalls

Regression is probably the most common statistical analysis performed on data, but few people fully understand the method



the slope is always positive —> there is always a positive relation between the variables

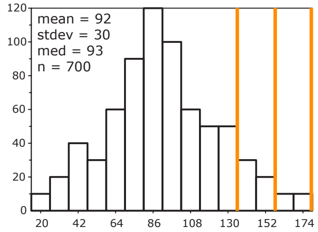
the slope is both positive and negative —> there can be both a positive and negative correlation

---

Regression analysis

The conc. of a heavy metal in soils from all over Europe:

determine the natural background so you can set pollution criteria



nice continuous distribution of the data;
can describe it with a mean/median and stdev/IQR

conclusion;
spread is large in the data, but there are no clear signs of pollution

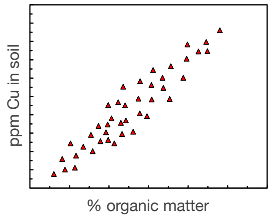
however; some samples were from heavily polluted sites, so why don't they jump out in the total data set?

unlikely to be one background value: will depend on soil type, composition etc

Regression analysis

The content of a heavy metal in soils from all over Europe:

organic matter content completely controls the conc of this heavy metal:



any soil with high organic matter content will have a natural enrichment

pollution will be an enrichment beyond that caused by organic matter

but how can we correct for the organic matter contribution ?

need to quantify the relation between organic matter and heavy metal content

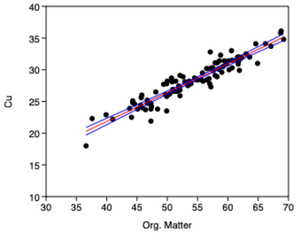
allows organic matter influence to be subtracted from the bulk composition so soils can be directly compared

to quantify this relation: use regression analysis

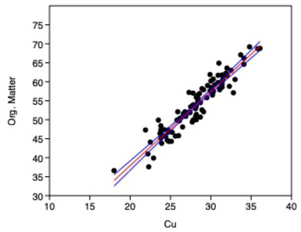
Regression analysis- linear model

conduct a regression analysis on this data set:

Identify the dependent and the independent variable



$$\hat{Y} = b_0 + b_1 X_i$$



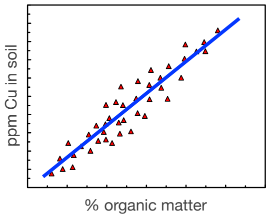
Cu = 0.455*Org. Matter + 3.64

Org. Matter = 1.96*Cu + 1.18

$Y = b_0 + b_1X$ → $X = (Y - b_0) / b_1 = -b_0/b_1 + 1/b_1Y$ → Org. Matter = 2.20*Cu - 8

Regression analysis- linear model

conduct a regression analysis on this data set:



$$\hat{Y} = b_0 + b_1 X_i$$

where $\hat{Y} =$ estimated value of Y at X_i

$b_0 =$ the intercept (Cu when no organics)

$b_1 =$ the slope of the data array

X (% organic matter) is the **independent** variable,

whereas Y (Cu content) is the **dependent** variable as it is a function of X

this regression equation is an estimate of the population equivalent:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where ε_i is an uncertainty term related to the variance in the data

---

Regression analysis - assumptions

assumptions (or requirements) for linear regression analysis:

ε_i - has to be normally distributed with a mean of 0 and variance σ_ε^2

equal distribution of points on either side of the regression curve as well as along the curve (throughout the data range)

i.e. the deviation from a perfect fit and should therefore be centred on your fit

for every value of X, the corresponding values of Y are normally distributed

if this is not the case: have to switch to a robust regression technique e.g. saturation level regression

μ_Y for every X has to lie on a linear trend with σ_ε^2 variance around this trend (when fitting a linear trend)

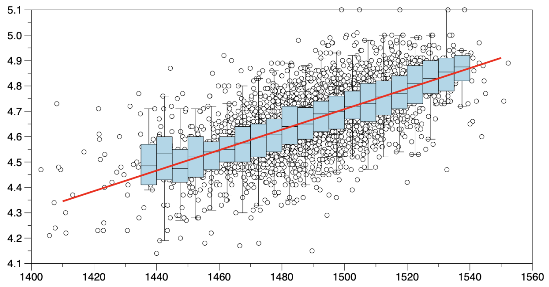
i.e. the μ_Y values should correctly describe the trend that you are modeling

Regression analysis - assumptions

assumptions (or requirements) for linear regression analysis:

μ_Y for every X has to lie on a linear trend with σ_ε^2 variance around this trend (when fitting a linear trend)

i.e. the μ_Y values should correctly describe the trend that you are modeling



Regression analysis - testing of the assumptions

assumptions (or requirements) for linear regression analysis that need to be tested:

1. that the regression coefficients and the intercept are meaningful (if not, the non-significant ones need to be removed from the regression model)
2. that the overall model is significant (using an ANOVA analysis, R^2 is not sufficient)
3. that the assumptions are met (residual distribution)
4. that the model is not overly dependent on a single datapoint or variable; i.e. an outlier (Variance Inflation Factor)

Regression analysis - ANOVA

Let's have a look at the data uncertainties in regression analysis

original data have associated uncertainty:

σ_x^2 and σ_y^2, however σ_y^2 is not independent:

$\sigma_y^2 = \beta_1^2\, \sigma_x^2 + \sigma_\varepsilon^2$

where the first part describes the uncertainty explained by the regression and the second part the uncertainty that is not

The total deviation from the mean (i.e. the sum of squares) is of course preserved, so;

$$SS_{TOT} = SS_x + SS_y = SS_{\beta 1x+\beta o} + SS_\varepsilon = SS_{\hat{y}} + SS_\varepsilon$$

where the latter two represent the deviation along the regression curve and the deviation around the regression fit respectively

Regression analysis - ANOVA

We can use the sums of squares to determine goodness-of-fit;

When $SS_{\hat{y}} >> SS_\varepsilon$ you have a good regression fit as most of the variance resides in the regression and there is only minimal variance remaining around this curve

When $SS_{\hat{y}} \leq SS_\varepsilon$ you have a poor regression fit as the deviation from your fit is equal or even larger than that along your fit

$$SS_{\hat{y}} = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$$

the deviation between the predicted and the mean of Y = $SS_{Regression}$

$$SS_\varepsilon = \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2$$

the deviation between the predicted and real value of Y = $SS_{Deviation}$

Regression analysis - ANOVA

The ratio between SS_R and SS_{TOT} is an indicator for the goodness-of-fit; the coefficient of determination R^2

$$R^2 = \frac{SS_R}{SS_{TOT}}$$

$R^2 = 1$: perfect regression fit as regression describes the full variance in the data ($SS_R = SS_{TOT}$)

$R^2 \approx 0$: no fit as the regression part of the variance is negligible ($SS_R << SS_{TOT}$)

Note: $R \neq r$

both relate the variance along a trend to the total variance in your data, but they are based on different assumption and have different requirements on the input data !

Regression analysis - ANOVA

Distribution of variance in regression analysis

var source	sum of squares	d.f.	variance
regression	SS_R	1	MS_R
deviation	SS_D	n - 2	MS_D
total	SS_{TOT}	n - 1	

MS = mean square

what are the d.f. for each contribution?

deviation: need β_1 and β_0 coefficients to determine the predicted value of Y, which you need for SS_D, so the d.f. = n - 2

regression: only 1 degree of freedom as the slope fixes the relation between the variables; can only shift curve up or down

total d.f.: essentially the deviation in Y_i from the mean of Y, so n - 1

Regression analysis - ANOVA

var source	sum of squares	d.f.	variance
regression	SS_R	1	MS_R
deviation	SS_D	n - 2	MS_D
total	SS_{TOT}	n - 1	

MS = mean square

variance = sum of squares divided by the degrees of freedom:

$s^2_D = MS_D = SS_D / n\text{-}2$ and $s^2_R = MS_R = SS_R / 1$

This can be used to determine whether the regression fit is significant following our earlier ANOVA approach:

MS_R has to be significantly larger than MS_D at alpha:

F-test on the ratio of MS_R and MS_D (H_0; $MS_R = MS_D$)

Regression analysis

What if the fit is not significant ?

1. there is no correlation between the variables
   plot the data in a scatter diagram and check
2. the correlation is weak and not significant due to lack of data
   obtain more data or accept a larger value of alpha
3. the data are correlated, but the correlation is not linear
   repeat the same exercise using a more appropriate curve:

quadratic: $Y = b_1X + b_2X^2 + b_0$

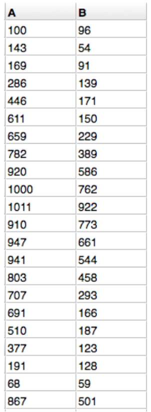
exponential: $Y = b_0 \, EXP(b_1X)$

reciprocal: $Y = 1 / (b_1X + b_0)$

multiple linear: $Y = b_1X_1 + b_2X_2 + b_3X_3 + b_0$

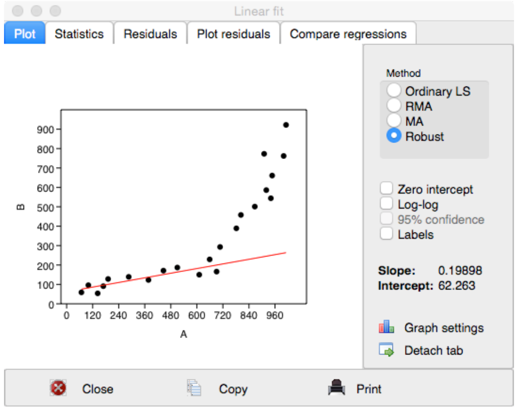
Linear regression with the statistics package PAST

data



A	B
100	96
143	54
169	91
286	139
446	171
611	150
659	229
782	389
920	586
1000	762
1011	922
910	773
947	661
941	544
803	458
707	293
691	166
510	187
377	123
191	128
68	59
867	501

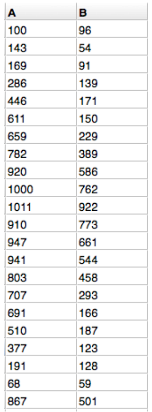
linear fit



Slope: 0.19898
Intercept: 62.263

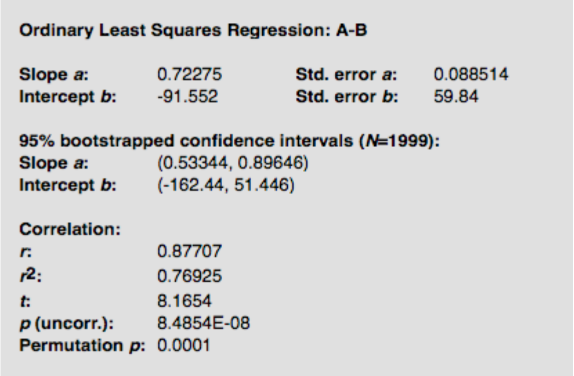
Linear regression with the statistics package PAST

data



A	B
100	96
143	54
169	91
286	139
446	171
611	150
659	229
782	389
920	586
1000	762
1011	922
910	773
947	661
941	544
803	458
707	293
691	166
510	187
377	123
191	128
68	59
867	501

linear fit - statistics



Ordinary Least Squares Regression: A-B

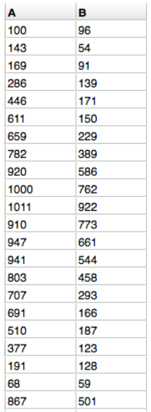
Slope *a*: 0.72275 **Std. error *a*:** 0.088514
Intercept *b*: -91.552 **Std. error *b*:** 59.84

95% bootstrapped confidence intervals (*N*=1999):
Slope *a*: (0.53344, 0.89646)
Intercept *b*: (-162.44, 51.446)

Correlation:
***r*:** 0.87707
***r*2:** 0.76925
***t*:** 8.1654
***p* (uncorr.):** 8.4854E-08
Permutation *p*: 0.0001

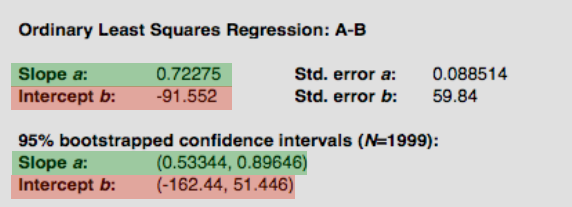
Linear regression with the statistics package PAST

data



A	B
100	96
143	54
169	91
286	139
446	171
611	150
659	229
782	389
920	586
1000	762
1011	922
910	773
947	661
941	544
803	458
707	293
691	166
510	187
377	123
191	128
68	59
867	501

linear fit: are the coefficients significant ?



Ordinary Least Squares Regression: A-B

Slope *a*: 0.72275 **Std. error *a*:** 0.088514
Intercept *b*: -91.552 **Std. error *b*:** 59.84

95% bootstrapped confidence intervals (*N*=1999):
Slope *a*: (0.53344, 0.89646)
Intercept *b*: (-162.44, 51.446)

H_0; $a = 0, b = 0$ $t_{\alpha,df}$ = (a - 0) / stdev

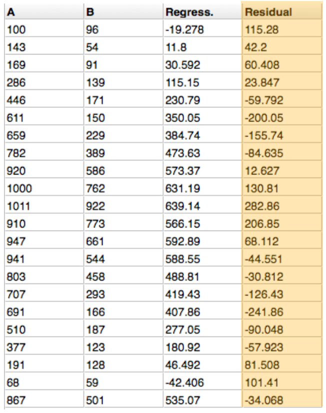
H_A; $a \neq 0, b \neq 0$ $t_{\alpha,df}$ = (b - 0) / stdev

t (slope) $_{calc}$ = 8.16 > t $_{\alpha,df}$ = 2.08 -> reject H_0

t (intercept) $_{calc}$ = -1.59 < t $_{\alpha,df}$ = -2.08 -> accept H_0

Linear regression with the statistics package PAST

data



A	B	Regress.	Residual
100	96	-19.278	115.28
143	54	11.8	42.2
169	91	30.592	60.408
286	139	115.15	23.847
446	171	230.79	-59.792
611	150	350.05	-200.05
659	229	384.74	-155.74
782	389	473.63	-84.635
920	586	573.37	12.627
1000	762	631.19	130.81
1011	922	639.14	282.86
910	773	566.15	206.85
947	661	592.89	68.112
941	544	588.55	-44.551
803	458	488.81	-30.812
707	293	419.43	-126.43
691	166	407.86	-241.86
510	187	277.05	-90.048
377	123	180.92	-57.923
191	128	46.492	81.508
68	59	-42.406	101.41
867	501	535.07	-34.068

linear fit: is it significant ?

$$SS_D = \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i \right)^2 = SS_{Residual}$$

$$SS_R = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y} \right)^2$$ mean Y = 340

$SS_{TOT} = SS_R + SS_D$ $R^2 = \frac{SS_R}{SS_{TOT}}$

SS_D 345975
SS_R 1153347 $R^2 = 0.77$
SS_{TOT} 1499321

Linear regression with the statistics package PAST

var source	sum of squares	d.f.	variance
regression	$SS_R = 1153347$	1	$s^2_R = 1153347$
deviation	$SS_D = 345975$	n - 2 = 20	$s^2_D = 17299$
total	$SS_{TOT} = 1499321$	n - 1 = 21	

$s^2_D = SS_D / n\text{-}2$ and $s^2_R = SS_R / 1$

For the regression model to be meaningful, s^2_R has to be significantly larger than s^2_D at your chosen confidence level:

F-test on the ratio of s^2_R and s^2_D (H_0; $s^2_R = s^2_D$)

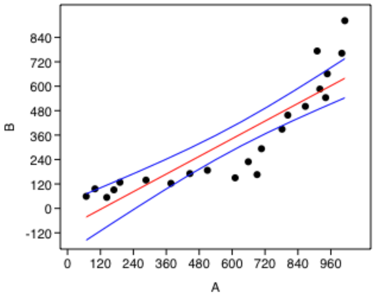
$F_{calc} = 66.67 > F_{0.05,1,20} = 4.35$ The model is meaningful

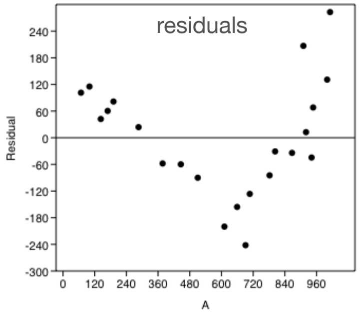
---

Linear regression with PAST

F-ratio is sufficiently high that we can reject the H_0 hypothesis:

the regression fit explains a significant part of the total variance and is therefore meaningful

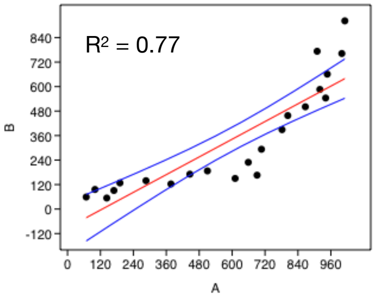




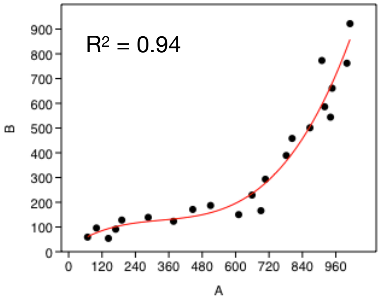
Appropriate fit for this dataset

Even though the linear regression fit is significant, it is not necessarily the most appropriate fit for the data

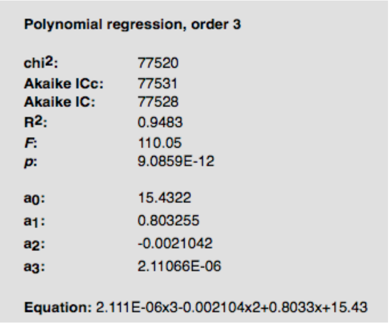
linear fit



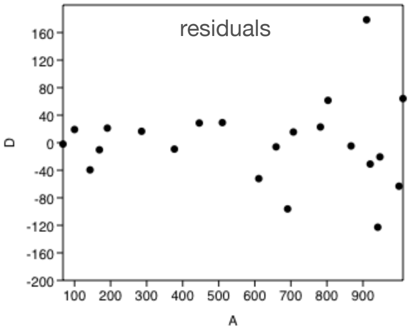
3rd polynomial fit



Cubic regression with PAST



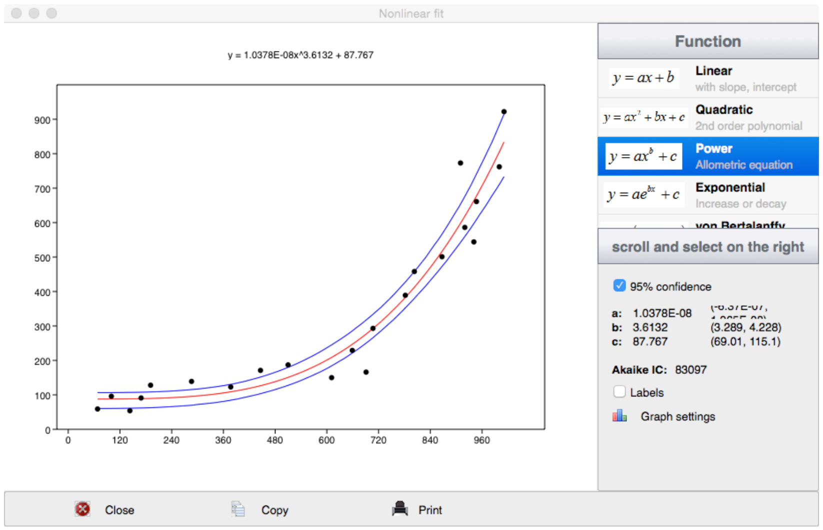
```
Polynomial regression, order 3

chi2:          77520
Akaike ICc:    77531
Akaike IC:     77528
R2:            0.9483
F:             110.05
p:             9.0859E-12

a0:            15.4322
a1:            0.803255
a2:            -0.0021042
a3:            2.11066E-06

Equation: 2.111E-06x3-0.002104x2+0.8033x+15.43
```



F-ratio is higher than before: a more significant model for the data.

Chance of obtaining this result purely by chance: 1 / 100000000000

Regression and curve fitting in PAST



☑ 95% confidence

a: 1.0378E-08
b: 3.6132 (3.289, 4.228)
c: 87.767 (69.01, 115.1)

Akaike IC: 83097

☐ Labels

Multiple linear regression with PAST

the dependent is a linear combination of many independents:

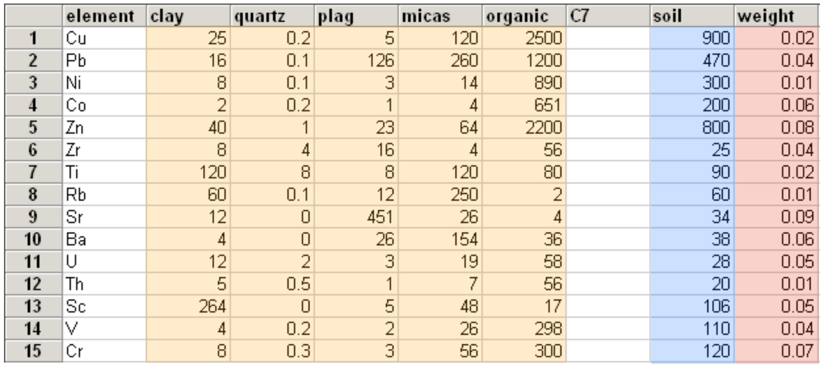
the composition of a soil will be the sum of the compositions of its constituents multiplied by their respective fraction in the soil:

	clay	quartz	plag	micas	organic
Cu	25	0	5	120	2500
Pb	16	0.1	50	260	1200
Ni	8	0	3	14	890
Co	2	0	1	4	651
Zn	40	0.5	23	64	2200
Zr	8	4	16	4	56
Ti	120	8	8	140	80

$$\mathrm{Cu(soil)} = X_{\mathrm{clay}}{*}25 + X_{\mathrm{qtz}}{*}0 + X_{\mathrm{plag}}{*}5 + X_{\mathrm{micas}}{*}120 + X_{\mathrm{organic}}{*}2500$$

Multiple linear regression with PAST

can derive the phase fractions by multiple linear regression



	element	clay	quartz	plag	micas	organic	C7	soil	weight
1	Cu	25	0.2	5	120	2500		900	0.02
2	Pb	16	0.1	126	260	1200		470	0.04
3	Ni	8	0.1	3	14	890		300	0.01
4	Co	2	0.2	1	4	651		200	0.06
5	Zn	40	1	23	64	2200		800	0.08
6	Zr	8	4	16	4	56		25	0.04
7	Ti	120	8	8	120	80		90	0.02
8	Rb	60	0.1	12	250	2		60	0.01
9	Sr	12	0	451	26	4		34	0.09
10	Ba	4	0	26	154	36		38	0.06
11	U	12	2	3	19	58		28	0.05
12	Th	5	0.5	1	7	56		20	0.01
13	Sc	264	0	5	48	17		106	0.05
14	V	4	0.2	2	26	298		110	0.04
15	Cr	8	0.3	3	56	300		120	0.07

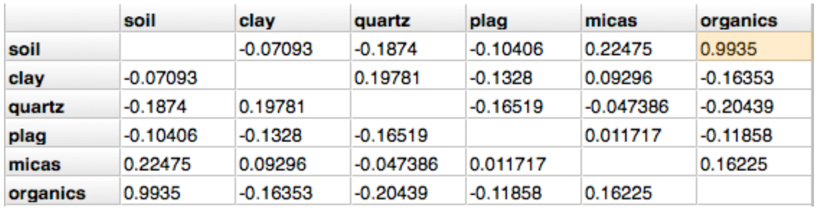
independents

dependent

Multiple linear regression with PAST

can derive the phase fractions by multiple linear regression

Pearson correlation coefficient matrix



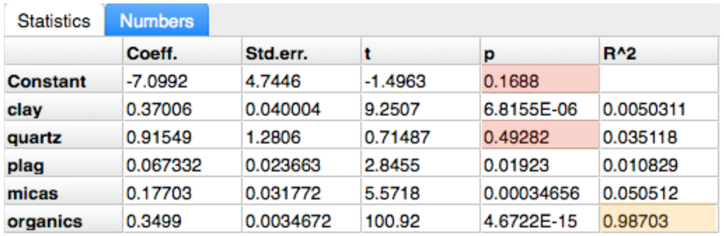
	soil	clay	quartz	plag	micas	organics
soil		-0.07093	-0.1874	-0.10406	0.22475	0.9935
clay	-0.07093		0.19781	-0.1328	0.09296	-0.16353
quartz	-0.1874	0.19781		-0.16519	-0.047386	-0.20439
plag	-0.10406	-0.1328	-0.16519		0.011717	-0.11858
micas	0.22475	0.09296	-0.047386	0.011717		0.16225
organics	0.9935	-0.16353	-0.20439	-0.11858	0.16225	

potential problem: organics strongly dominant control on soil composition

Multiple linear regression with PAST

can derive the phase fractions by multiple linear regression

The regression model:



	Coeff.	Std.err.	t	p	R^2
Constant	-7.0992	4.7446	-1.4963	0.1688	
clay	0.37006	0.040004	9.2507	6.8155E-06	0.0050311
quartz	0.91549	1.2806	0.71487	0.49282	0.035118
plag	0.067332	0.023663	2.8455	0.01923	0.010829
micas	0.17703	0.031772	5.5718	0.00034656	0.050512
organics	0.3499	0.0034672	100.92	4.6722E-15	0.98703

regression coefficients

±

t-test on coeff.

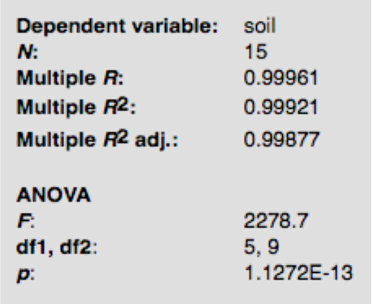
contribution to R²

probability that coefficient is 0

Multiple linear regression with PAST

can derive the phase fractions by multiple linear regression

The regression model:



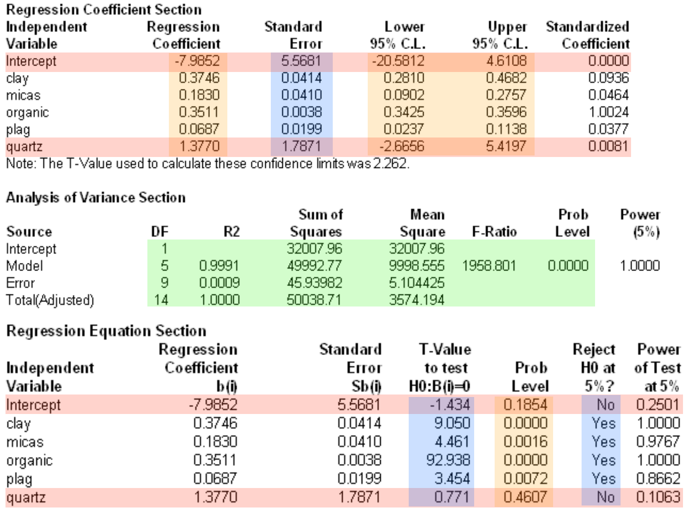
Dependent variable:	soil
***N*:**	15
Multiple *R*:	0.99961
Multiple *R*2:	0.99921
Multiple *R*2 adj.:	0.99877

ANOVA

***F*:**	2278.7
df1, df2:	5, 9
***p*:**	1.1272E-13

very high F_{calc}

Multiple linear regression with NCSS



Regression Coefficient Section

Independent Variable	Regression Coefficient	Standard Error	Lower 95% C.L.	Upper 95% C.L.	Standardized Coefficient
Intercept	-7.9852	5.5681	-20.5812	4.6108	0.0000
clay	0.3746	0.0414	0.2810	0.4682	0.0936
micas	0.1830	0.0410	0.0902	0.2757	0.0464
organic	0.3511	0.0038	0.3425	0.3596	1.0024
plag	0.0687	0.0199	0.0237	0.1138	0.0377
quartz	1.3770	1.7871	-2.6656	5.4197	0.0081

Note: The T-Value used to calculate these confidence limits was 2.262.

Analysis of Variance Section

Source	DF	R2	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1		32007.96	32007.96			
Model	5	0.9991	49992.77	9998.555	1958.801	0.0000	1.0000
Error	9	0.0009	45.93982	5.104425			
Total(Adjusted)	14	1.0000	50038.71	3574.194			

Regression Equation Section

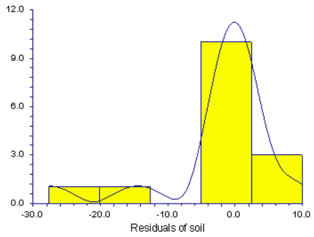
Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0:B(i)=0	Prob Level	Reject H0 at 5%?	Power of Test at 5%
Intercept	-7.9852	5.5681	-1.434	0.1854	No	0.2501
clay	0.3746	0.0414	9.050	0.0000	Yes	1.0000
micas	0.1830	0.0410	4.461	0.0016	Yes	0.9767
organic	0.3511	0.0038	92.938	0.0000	Yes	1.0000
plag	0.0687	0.0199	3.454	0.0072	Yes	0.8662
quartz	1.3770	1.7871	0.771	0.4607	No	0.1063

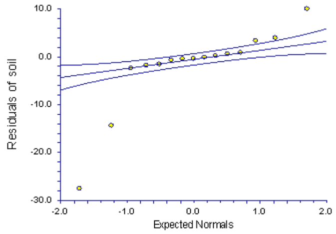
Multiple linear regression with NCSS - checks

Parameter	From PRESS Residuals	From Regular Residuals
Sum of Squared Residuals	1521.904	1112.096
Sum of |Residuals|	97.44209	68.4433
R2	0.9987	0.9994

Multicollinearity Section

Independent Variable	Variance Inflation Factor	Tolerance
clay	0.9857	0.9544
micas	0.8743	0.9445
organic	0.8817	0.8768
plag	0.9257	0.8552
quartz	0.9447	0.9193



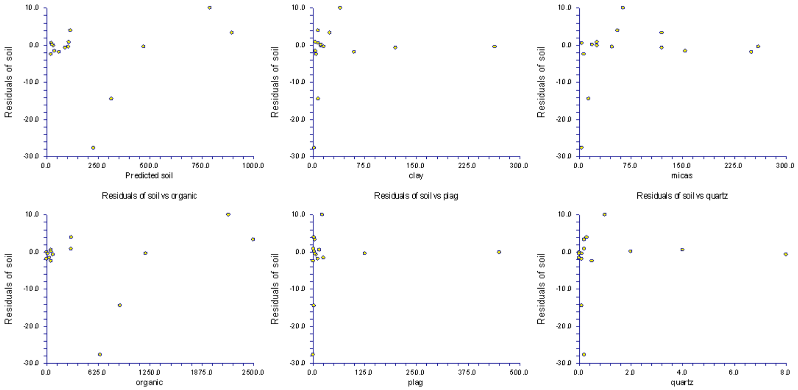


no significant difference between regular and PRESS R²

no significant variance inflation (VIF < 5-10) and tolerance close to 1

residuals are normally distributed

Multiple linear regression with NCSS



no trends between the residuals and the (in)dependent variables

very good regression fit that satisfies all the requirements for regression

---

Regression summary

Regression analysis allows you to define a model for your data that is predictive (both interpolative and extrapolative)

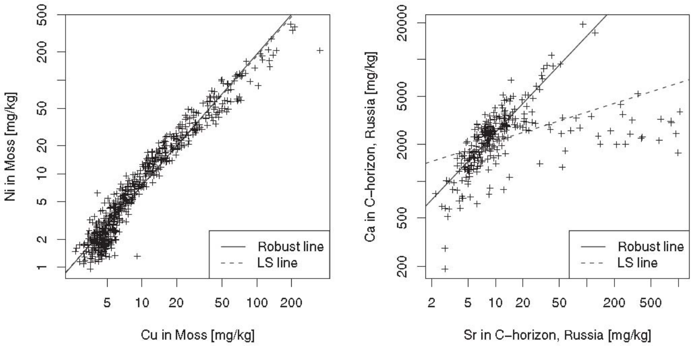
However, have to test that the model is meaningful by testing:

1. that the regression coefficients and the intercept are meaningful (if not, the non-significant ones need to be removed from the regression model)
2. that the overall model is significant (using an ANOVA analysis, R^2 is not sufficient)
3. that the assumptions are met (residual distribution)
4. that the model is not overly dependent on a single datapoint or variable

---

Robust regression

Deviations from normality, such as outliers, can have a major impact on regression coefficients and invalidate results. Unfortunately such datasets cannot always be avoided: use robust regression



Robust regression - Sen slope

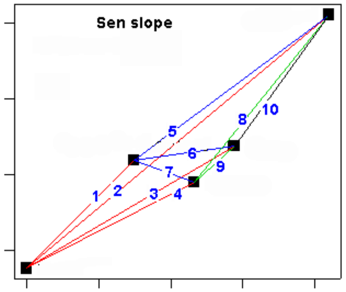
One type of robust regression, which is especially suited to small sets of data is the Sen slope:

The Sen slope involves calculating the slope of each combination of two data points, and then taking the median of these slopes as the robust characteristic slope

slope = Δx / Δy

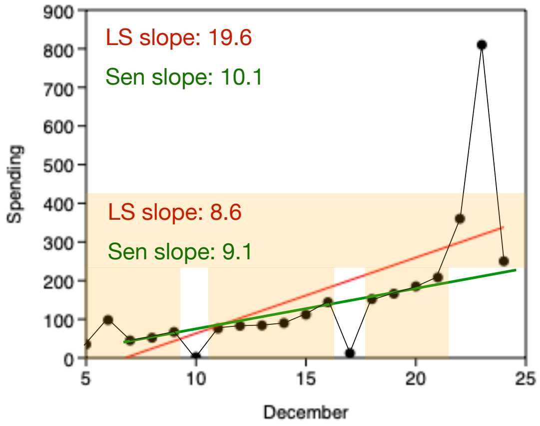
for 5 data points: 10 slopes

Sen slope = median(10 slopes)



Robust regression - Sen slope

date	shopping spending
5-Dec	35
6-Dec	98
7-Dec	45
8-Dec	52
9-Dec	67
10-Dec	2
11-Dec	76
12-Dec	83
13-Dec	84
14-Dec	90
15-Dec	112
16-Dec	144
17-Dec	12
18-Dec	152
19-Dec	166
20-Dec	185
21-Dec	208
22-Dec	360
23-Dec	810
24-Dec	250



Day 4 - topics covered

- Markov chain analysis of transitions (strat. columns)
- (non-)randomness in time series data
- Auto- and cross-correlation among time series data
- Periodicity in time series data
- Dependent vs. independent variables
- Testing linear regression models
- Non-linear regression
- Robust regression