The Correlation Principle

Estimation with (Nonparametric) Correlation Coefficients

The Correlation Principle for the estimation of a parameter theta:

A Statistical inference of procedure should be consistent with the assumption that any explanation of a set of data should be accompanied by theta-hat, a value of theta, that makes some correlation coefficient zero.

Historical: the combination of observations

Although extensive methods have been developed for linear models, generalized linear models, non-linear models, and time series models, as well as estimation of parameters for a particular distribution, we only have time to give one result. A nonparametric correlation coefficient measures monotonicity rather than linearity. It will be shown how to measure linearity with any nonparametric correlation coefficient. The Greatest Deviation CORRELATION COEFFICIENT (GD) will be used and its robustness demonstrated.

An exact quadratic relationship

Let y = 3+ 0.5 * x**2 with no error

let r(x,y) be Pearson's CORRELATION COEFFICIENT

on a set of bivariate x,y data. Let GD(x,y) be the GD

correlation on data. Let x=0(1/4)5 and y as above.

Then y ranges from 3 to 15 and 1/2. Note that there is a perfect monotonic relationship between x and y.

Correlation values and regression

$r(x,y) = 0.9655$, $\quad GD(x,y) = 1$

In order to get more information from GD, the regression of y on x is performed. LS stands for least squares regression. Regression for the slope is done by solving the following equations:

LS; $r(x,y-bx) = 0$ and GD; $GD(x,y-bx) = 0$.

The intercepts are obtained by taking the mean and median of the uncentered residuals, $y-bx$, respectively.

We are using quantile type plot reasoning in the next slide on the residuals.

The regression results

LS: yhat = 1.0208 + 2.5000 * x 10 MAY2002

GD: yhat = 0.890625 + 2.4375 * x

The GD residuals, y-yhat = res are such that GD(x,res) = 0

This is the starting point to compute a CORRELATION COEFFICIENT from a NPCC that measures linearity.

Let y-ord be the y data ordered and similarly, res-ord. Now perform the regression of res-ord on y-ord, by solving

GD(y-ord, res-ord - s * y-ord) = 0

this gives s, the slope as 0.2625. Note 1-(0.2625)**2 =0.9311

0.9311 estimates the square of a CORRELATION COEFFICIENT

Comparison to LS

Recall from LS that SS(res)/SSY + SS(reg)/SSY = 1

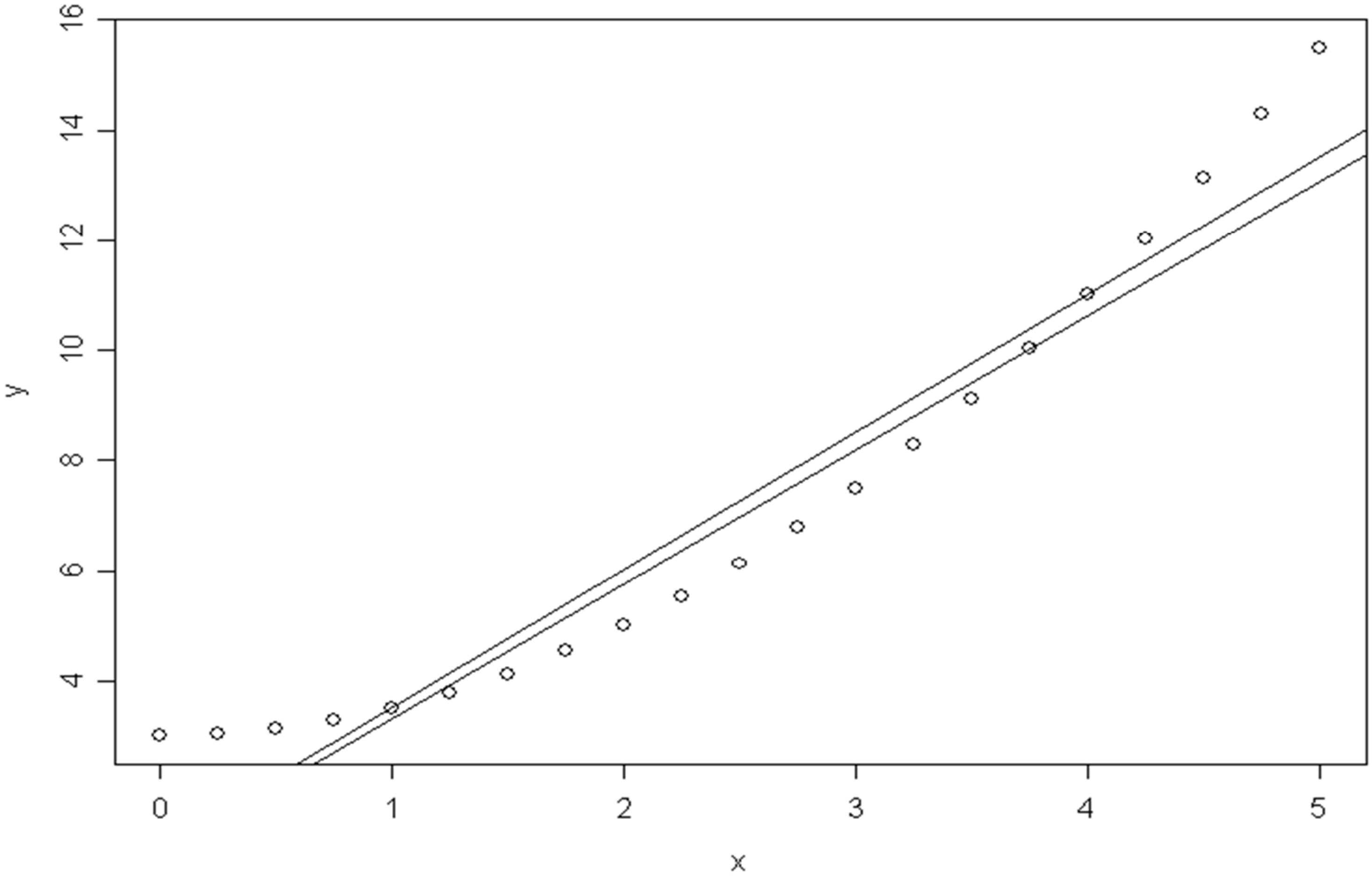
From the LS regression above, SD(res) = 1.0737

and SD(Y) = 4.0167, so that SS(res)/SSY = the square of

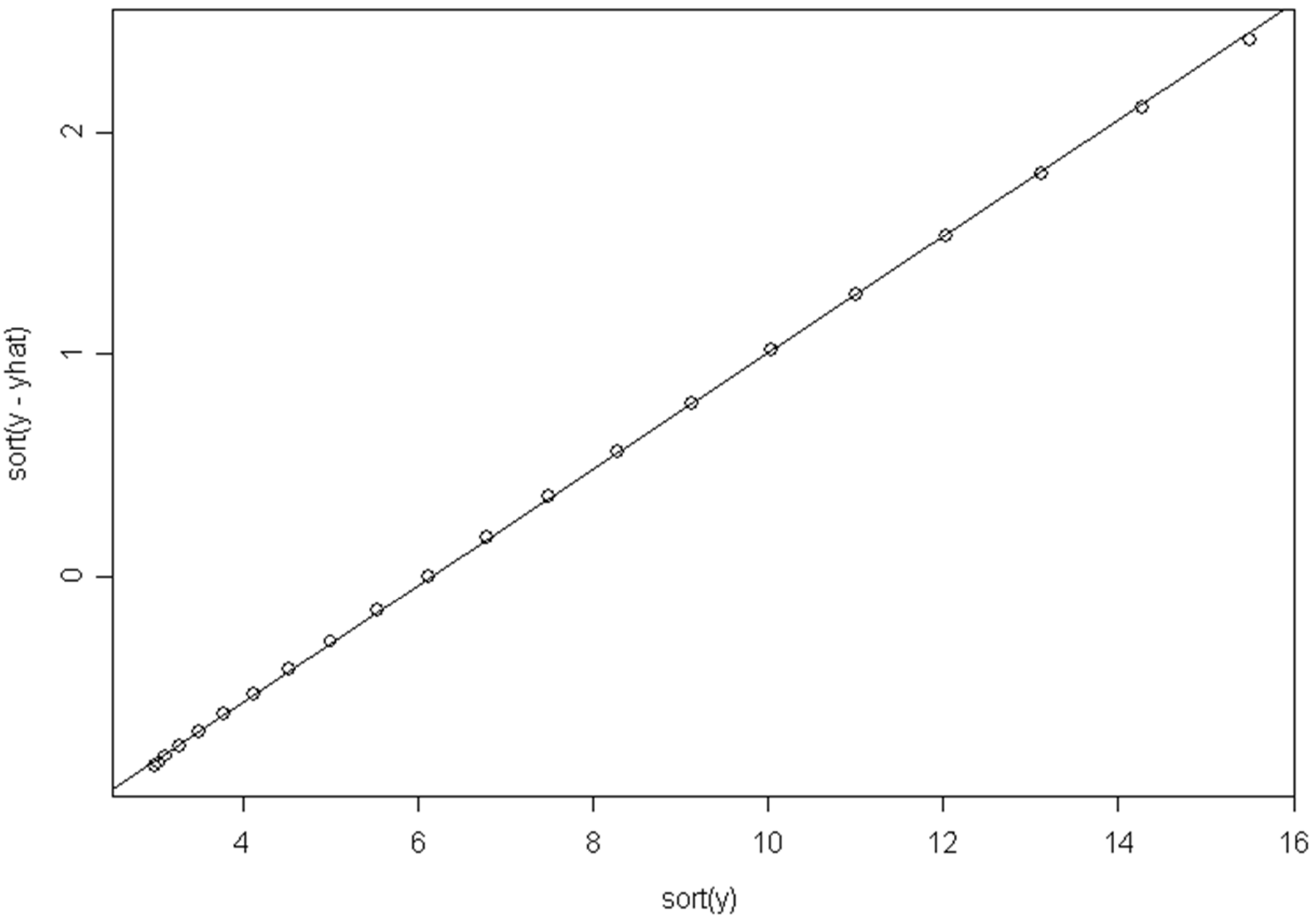
1.0737/4.0167 which equals 0.2673. Note that this is

very close to the slope coming from the second GD regression, that is, 0.2625. We are doing a regression with GD to obtain the proportion of Y explained by the explanatory variable. (in a linear relationship!!!)

Regressions: LS and GD on Quadratic Data



Regression for error estimate



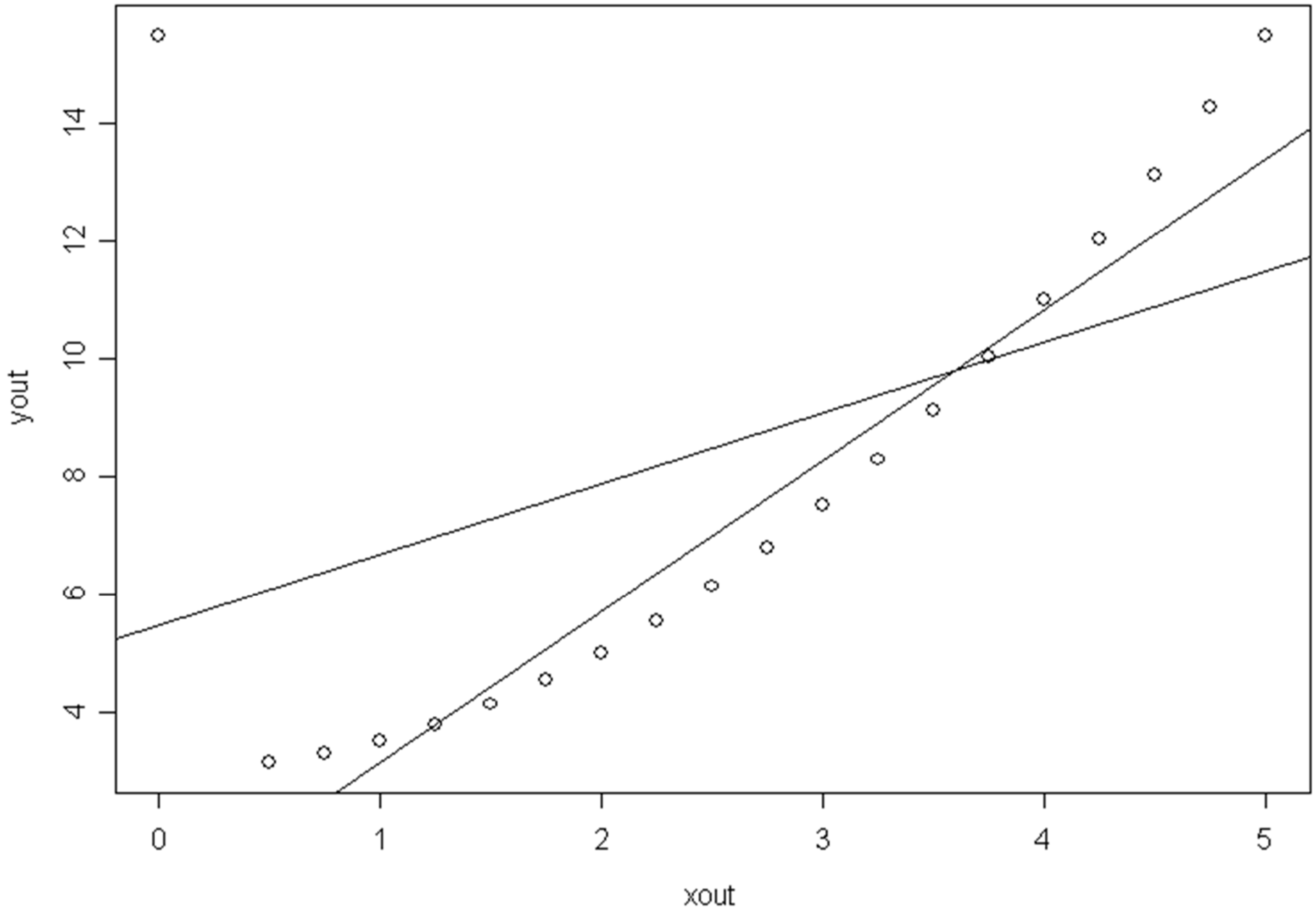
The Outlier Data

The data is the same except for the smallest two x values, in which there are two tied values at x=0, with y =15.5. Again label the data x and y.

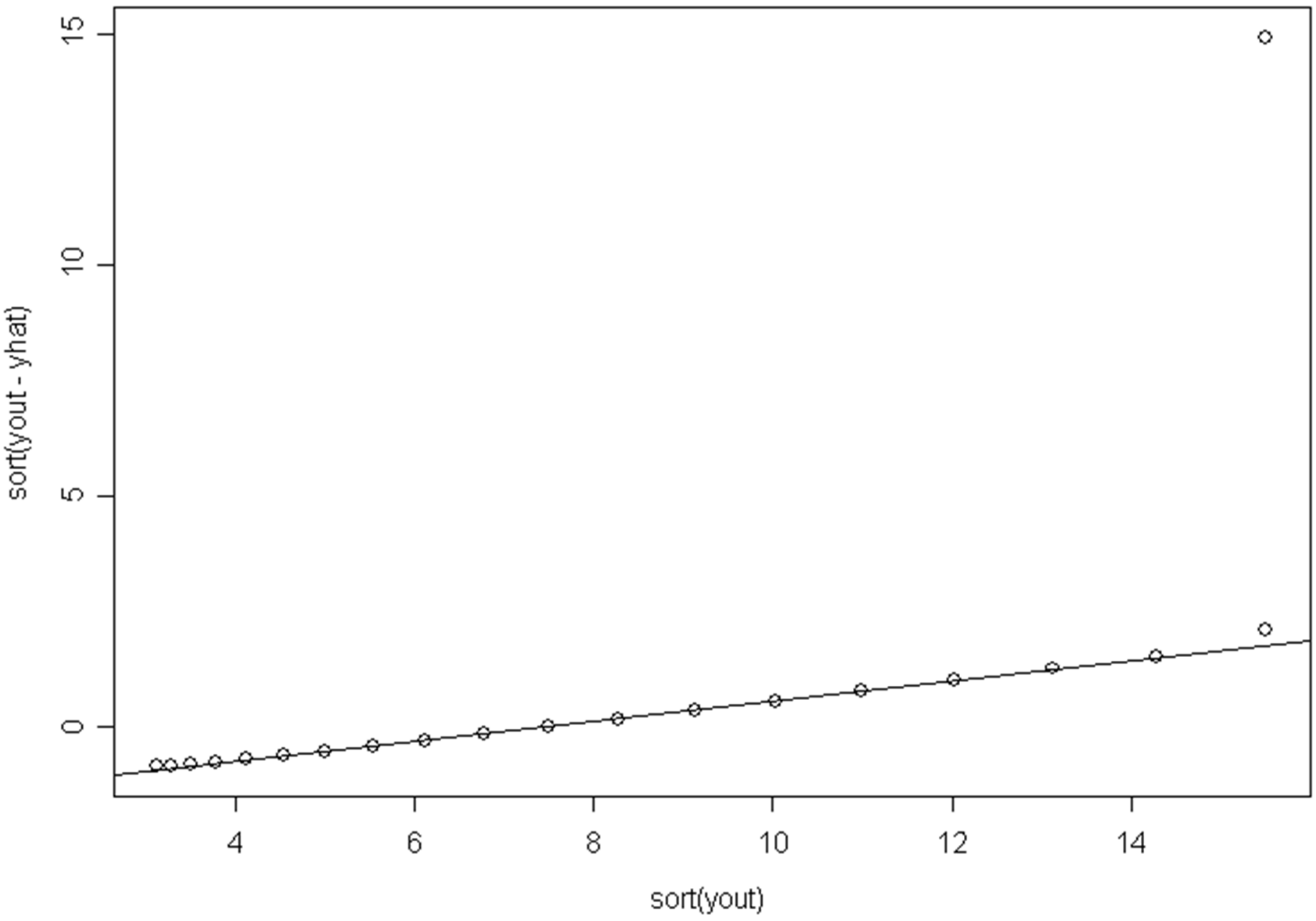
r(x,y) = 0.4260, GD(x,y) = 0.7000.

Regression results: LS y = 5.4702 + 1.20155 * x

GD y=0.578125 + 2.5625 * x


yout
4
6
8
10
12
14
0
1
2
3
4
5
xout

Error estimate with outliers



Outlier Results

r(x1,y1) = 0.42602, GD(x1,y1) = 0.7

Regressions: LS, y1 = 5.470 + 1.201 * x1

GD, y1 = 0.5781 + 2.5625 * x1

The GD linear CORRELATION COEFFICIENT estimate

let res1 be the residuals from the GD regression.

Solve for s in GD(y1-ord, res1-ord -s * y1-ord) = 0

This yields s = 0.21870 and 1-s**2 = 0.9521, an estimate of a square of a CORRELATION COEFFICIENT

Comparison between LS and GD

It is clear that GD is robust but LS or Pearson’s r is not.

Correlations and regression coefficients are dramatically different. Both GD directly and a linear correlation with GD regression leads to a much higher correlation than classical statistics.

Good data is indicated when sin(pi *GD/2), which is always greater than GD, is close to Pearson’s r. This is a normal distribution conversion to population values of what is estimated.

Absolute rather than relative comparison

Someone may say that LS is better no matter what because it gives estimates of residual SS and SS of the centered y-data (the response variable) rather than the relative estimate of SS(res)/SS(y-centered) through GD regression.

However, using regressions as illustrate above, an ordered norm comparable to the classical norm can be defined to overcome this criticism. Let the following be notation

$$\left\| y \right\| \text{and} \left\| y \right\|^{o}$$

Ordered Norm is indicated by Onorm

Without defining it, we will instead use it to illustrate that for this example, we get the same useful robustness and more direct measures. We list the classical and order norm for the y data and for the residuals for the respective regression. We do this both for the good (but quadratic) data and the same data with the two outliers.

By SSY is meant the centered y norm data, but,of course GD would use an ordered-mean (median) for the centering. Since residuals are already centered, no additional centering is done.

Norm and Order Norm

	Good Data		Outlier Data	
	Norm	Order Norm	Norm	Order Norm
data	37.853	36.198	43.533	43.040
location	7.2708	6.9843	8.4598	8.2452
Centered Norm	17.963	16.909	19.804	20.611

Residual Norms

For the “Good Data”

Norm(res) =4.6805 for LS, so Pearson’s r is

1-(4.6805/17.963)**2 = 0.9320 or r = 0.9654.

For GD, Onorm(res) = 4.5519 so Linear Correlation

via GD is 1 - (4.5519/16.909)**2 = 0.927531.

This very close to the earlier relative approach

1 - s**2 =1- (0.2625)**2 = 0.9311

Onorm for “Outlier Data”

Onorm(res) = 4.6397 and so the correlation squared is

1 - (4.6397/20.611)**2 = 0.9493.

This compares to 1 - s**2 = 1 - (0.21870)**2 =0.9521.

GD(x1,y1) = 0.7 and sin(pi*0.7/2) = 0.8910

Summary

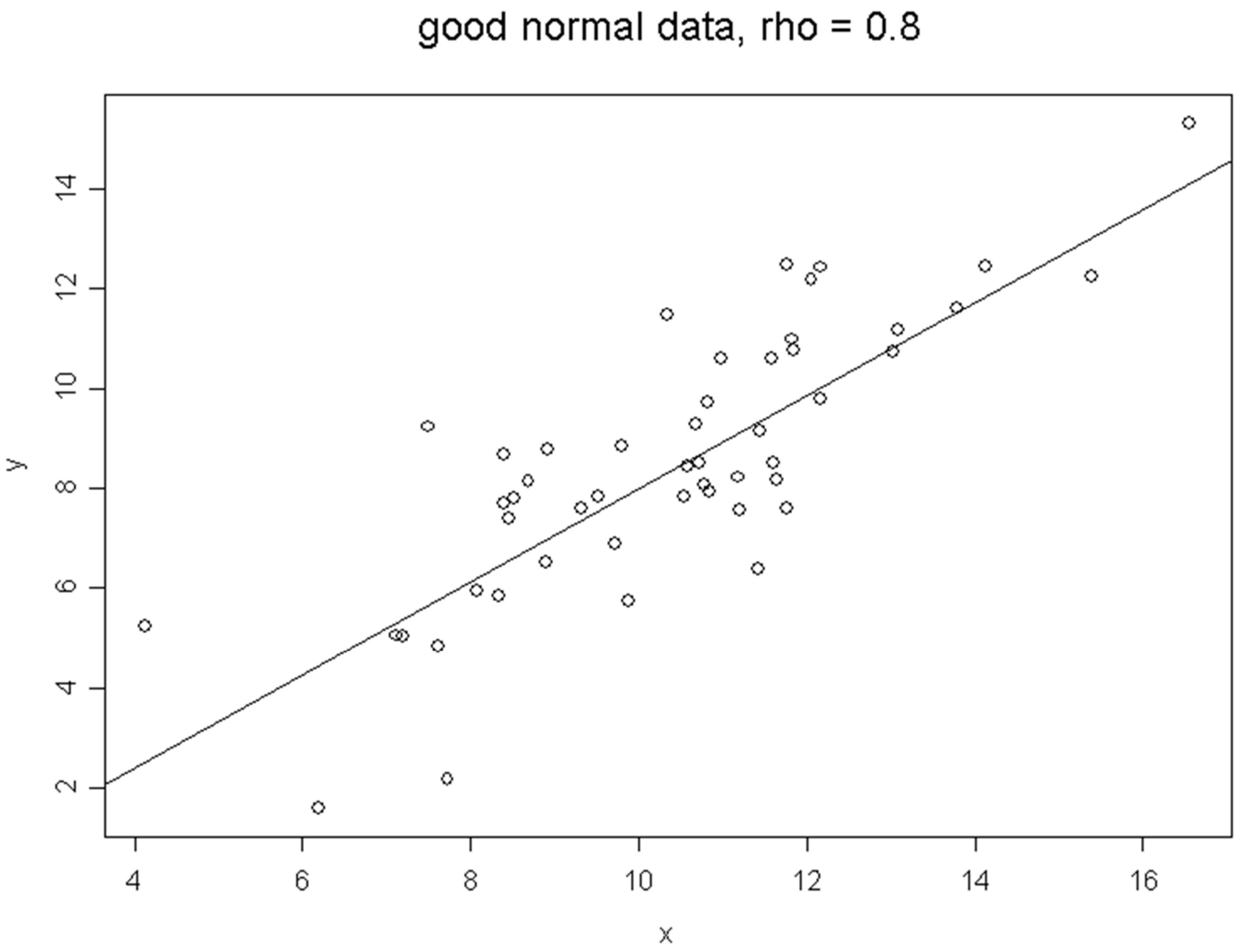
•Tied values on the outlier example were chosen to demonstrate that tied values do not alter the GD or Onorm method. The numerical methods work even if there are all tied values.

• No visualization is really needed to ascertain whether or not the data is “good”. Just compare GD methods and LS methods

•There probably is not time to show that for “good data” the GD and LS method give very comparable results. Example follows.

•GD is robust up to about 30% of the data being “Bad”

•There is a chance that this method may be lost due to lack of interest


good normal data, rho = 0.8
y
2
4
6
8
10
12
14
x
4
6
8
10
12
14
16

regression results

N = 50 for good normal data, r(x,y) = 0.8079, GD(x,y) = 0.48,

sin(pi*GD/2) = 0.6845.

LS: yhat = -1.1275 + 0.9326 *x, sigma = 1.5964

GD: yhat = -1.3111 + 0.9314*x

True Model: y = .8*x + normal(0,3) error, sigma =3

GD and LS or Pearson’s nearly the same, not bad data!!!

GD linear estimates of Corr

First, regress ordered GD residuals on ordered y-data to obtain a direct but relative estimate of SS(res)/SSY, s = 0.64077,

so the linear NP correlation estimate is 1 - s^2 = 0.58941, and the square root of this is 0.7677, close to 0.8

Onorm method

Review by LS: the classical norms give SS(res)/SSY = square of

11.0602/18.7700 = 0.58934, 1- (0.58934)^2 = 0.6526, and the square root of this is Pearson's r, sqrt(0.6526) = 0.8079.

The same method but with GD regession and Onorms.

SS(res)/SSY = 9.3234/15.5392 =.6000,

So 1 - 0.6^2 = 0.64 and the estimate of rho is sqrt(0.64) =0.8000

A result exact to 4 places. This data was just randomly picked as a good normal model.

Estimates of error sigma

LS; The sqrt of SS(res)/48 = 11.0602/sqrt(48) = 1.5964

GD: The sqrt of SS(res)/48 = 9.3234/sqrt(48) = 1.3457

mean(y) = 8.5396, GD Mean is 8.60892

Suggestions

This is a computer age. Every statistical analysis needs at least two distinct methods before drawing conclusions. For regression Pearson's, a robust NP correlation, slopes of the regressions, and a linear NP correlation.

IN NP methods there is not as much selection bias as in the following current methods. (1) trim 5 or 10 or even 20 % of the data, (2) omitting data points if there are deemed too influential, (3) weighting and reweighting the data until stability occurs.

In each of these latter cases the statistician can control the results rather than letting the data speak for itself

Why aren’t there more details?

This work has been in progress for twenty years and there are various papers explaining details. A Web site can be used to find many of them. This Web site is being built to catalog this CORRELATION COEFFICIENT system of estimation

www.math.umt.edu/~gideon

GD is easily computed by hand for untied data, but there is a C program that interfaces with Splus to run any data. These programs and papers explain estimation for location,scale, and regression parameters. This presentation today shows one interesting aspect of the correlation method.