Statistics 215a – 11/3/03 – D. R. Brillinger

r^2 and R^2. Squared coefficients of correlation and multiple correlation

Correlation and association are vague concepts

data, (x_i^T, y_i), $i=1,\ldots,n$

first entry of vector x is 1

1. describe y by m

   OLS $SS_1 = \sum (y_i - \bar{y})^2$

2. describe y by $x^T\beta$

   OLS $SS_2 = \sum (y_i - x_i^T b)^2$

3.

   $R^2 = 1 - SS_2/SS_1$

Properties.

1. $0 \le R^2 \le 1$

2. historical

measures of linearity

how well can y be approximated by **linear** function of x?

3. r =

$$\Sigma(x_i - \bar{x})(y_i - \bar{y})/\sqrt{\{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2\}}$$

4. Tukey and Winsor's *Society for the Suppression of Correlation Coefficients*

5. Remember Cleveland, Diaconnis, McGill example

6. Identities - ANOVA

$$\Sigma\, (y_i - x_i^T b)^2 = (1 - R^2)\, \Sigma(y_i - \bar{y})^2$$

Robust variants.

a) $1 - \sum |y_i - x_i^T b_*| / \sum |y_i - m_*|$

m_*, b_* being L_1 variants

b) Splus has cor(,trim=)

Andrews data: (.02,.04), (.99,1.03), (2.01,1.97), (2.98,2.96), (4.03,3.97), (5.01,4.98), (6.05,6.07), (6.98,7.03), (8.07,8.00), (9.03,8.96), (25.00,-25.00)

With trim = .1

$r = -.7325$, $r_{robust} = .9999$

c) biweight midcorrelation (NIST)

d) rank correlation

Statistics 215a – 11/3/03 – D. R. Brillinger

Exploratory time series analysis.

Time series – a succession of measurements of a quantity through time (or space, or …)

y_t, t in a set **T**, e.g. **T** = [0,T) or **T** = {0,1,..,T-1} or **T** = {τ_1, τ_2, …}

A function, a wiggly line, …

y is the response and t the explanatory or independent or exogenous variable.

Sometimes t is referred to as the parameter.

There are no repeated t’s

Interests/goals:

to express the dependence of y on t

prediction

model

surprises

…

The paradigm

response = fit + residual

$y_t = m_t + r_t$

remains appropriate

Visualization.

Tufte (1983) “A time-series plot is the most frequently used form of graphic design.”

There are various ways to display:

1. Connected symbols (e.g. points & lines)

2. Symbols (e.g. points)

   good for long term behavior

   cannot appreciate middle and high frequency behavior

   cannot perceive the order of the series over short time periods

3. Connected graph

good for smooth series

individual data points not unambiguously portrayed

irregular sampling can be unclear

4. Vertical graph

good when need to see individual values

good when series long (can pack tightly)

not good when strong trend

good about central value

Which to use depends on the situation

The plots display characteristics

e.g. trend, cycles, seasonal, steps, …

m_t may be polynomial, trigonometric, $\{y_{t-1}, y_{t-2}, \ldots, y_{t-p}\}$

One may seek a decomposition

Difficulties

T very large (speech)

strong trend

outliers

very rapid oscillations

missing values

A standardization

$(y_t - m_t)/s_t$

Methods.

stacking (Buys-Ballot)

useful when there exists a special period (two-way table)

parallel boxplots

fitting description robustly

Vector case.

use several line types, colors

forces comparisons