Il dataset originale
La correlazione
La retta di regressione

704SM Biostatistica

Massimo Borelli

ottobre 2016

UNIVERSITÀ DEGLI STUDI DI TRIESTE
Dipartimento di Scienze della Vita

SOCIETA' DEI MATEMATICI
E NATURALISTI DI MODENA
www.socnatmatmo.unimore.it

Il dataset originale
La correlazione
La retta di regressione

Francis Galton

Sir Francis Galton, FRS (/frænsɪs ˈɡɔːltən/; 16 February 1822 – 17 January 1911) was an English Victorian progressive, polymath, psychologist,[1][2] anthropologist, eugenicist, tropical explorer, geographer, inventor, meteorologist, proto-geneticist, psychometrician, and statistician. He was knighted in 1909.

Galton produced over 340 papers and books. He also created the statistical concept of correlation and widely promoted regression toward the mean. He was the first to apply statistical methods to the study of human differences and inheritance of intelligence, and introduced the use of questionnaires and surveys for collecting data on human communities, which he needed for genealogical and biographical works and for his anthropometric studies.
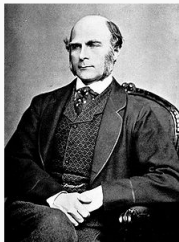
He was a pioneer in eugenics, coining the term itself[3] and the phrase "nature versus nurture".[4] His book *Hereditary Genius* (1869) was the first social scientific attempt to study genius and greatness.[5]

As an investigator of the human mind, he founded psychometrics (the science of measuring mental faculties) and differential psychology and the lexical hypothesis of personality. He devised a method for classifying fingerprints that proved useful in forensic science. He also conducted research on the power of prayer, concluding it had none by its null effects on the longevity of those prayed for.[6]

As the initiator of scientific meteorology, he devised the first weather map, proposed a theory of anticyclones, and was the first to establish a complete record of short-term climatic phenomena on a European scale.[7] He also invented the Galton Whistle for testing differential hearing ability.[8]

He was cousin of Douglas Strutt Galton and half-cousin of Charles Darwin.

Sir Francis Galton

Born	16 February 1822 Birmingham, England
Died	17 January 1911 (aged 88) Haslemere, Surrey, England
Residence	England
Nationality	English
Fields	Anthropology and polymathy
Institutions	Meteorological Council Royal Geographical Society
Alma mater	King's College London

Credits: http://en.wikipedia.org/wiki/Francis_Galton

Il dataset originale
La correlazione
La retta di regressione

La correlazione

ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.
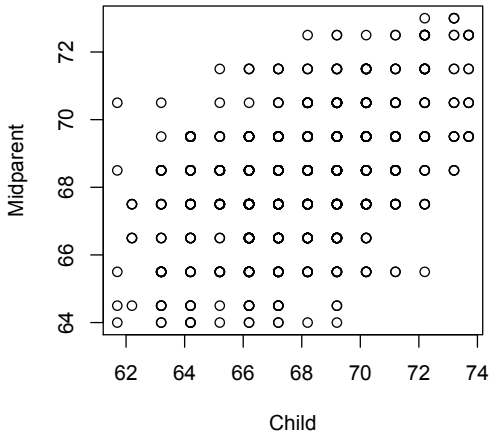
[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some

Il dataset originale
La correlazione
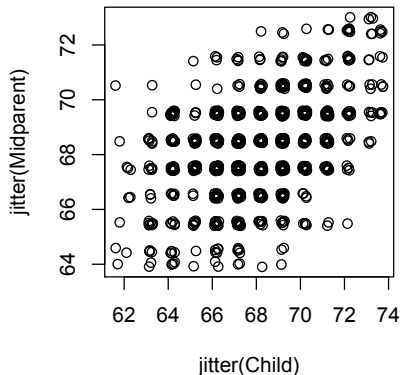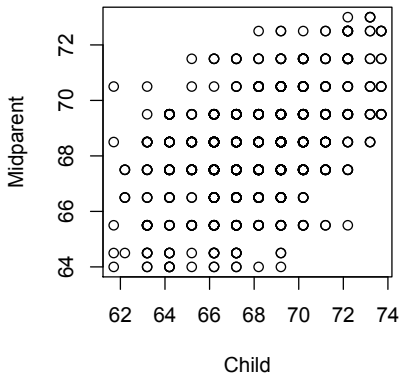La retta di regressione

Il dataset originale

```
indirizzo = http://www.math.uah.edu/stat/data/Galton.txt
galton = read.table(indirizzo, header = TRUE)
attach(galton)
str(galton)
```

```
'data.frame': 928 obs. of 2 variables:
$ Child  : num 61.7 61.7 62.2 62.2 ...
$ Midparent: num 70.5 68.5 67.5 66.5 ...
```
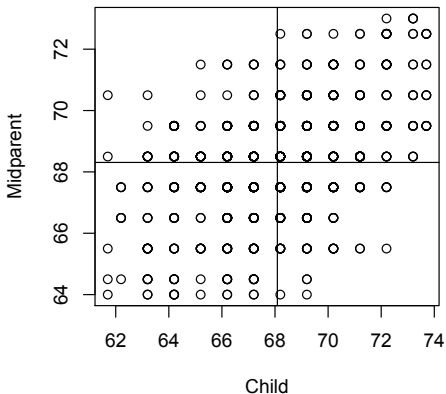
Il dataset originale
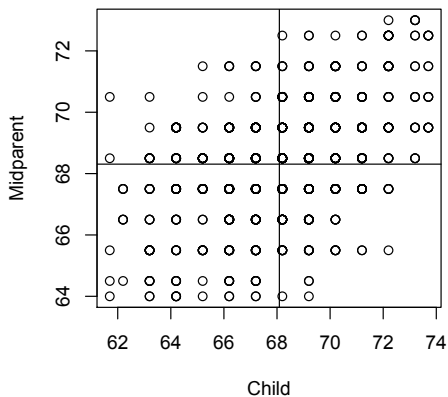La correlazione
La retta di regressione

plot(Child, Midparent)

Il dataset originale
La correlazione
La retta di regressione

par(mfrow = c(1,2))
plot(Child, Midparent)
plot(jitter(Child), jitter(Midparent))

Il dataset originale
La correlazione
La retta di regressione

plot(Child, Midparent)
abline(h = mean(Midparent))
abline(v = mean(Child))

Il dataset originale
La correlazione
La retta di regressione

? Child - mean(Child)
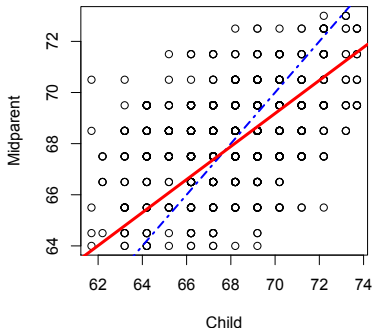
Il dataset originale
La correlazione
La retta di regressione

```
> sum((Child - mean(Child))*((Midparent - mean(Midparent))))
[1] 1913.898
> sum((Child - mean(Child))*((Midparent - mean(Midparent))))/(928-1)
[1] 2.064614
> cov(Child, Midparent)
[1] 2.064614
```

```
> cov(Child, Midparent) / (sd(Child) * sd(Midparent))
[1] 0.4587624
> cor(Child, Midparent)
[1] 0.4587624
```

Il dataset originale
La correlazione
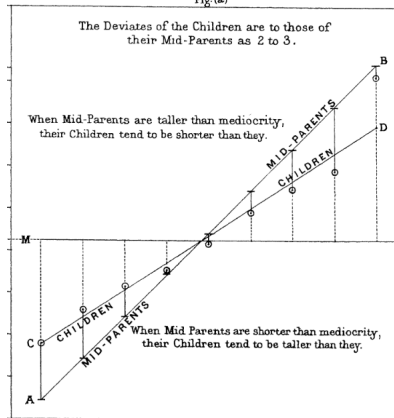**La retta di regressione**

La retta di regressione

```
plot(Child, Midparent)
abline(0,1, col = blue, lty = 4, lwd = 2)
modello = lm(Child ~ Midparent)
abline(modello, col = red, lwd = 3)
```

Il dataset originale
La correlazione
La retta di regressione

La correlazione

Il dataset originale
La correlazione
La retta di regressione

La retta di regressione

questione da approfondire

- adesso che conosciamo il 'razionale' della retta di regressione, come possiamo precisare il fatto che essa sia **la migliore** retta che passa per la nuvola di punti?
  - come sono legati tra loro la correlazione ed i coefficienti della retta di regressione?