Multivariable analysis

- Prof. Giuseppe Verlato
- Unit of Epidemiology & Medical Statistics, Department of Diagnostics & Public Health, University of Verona

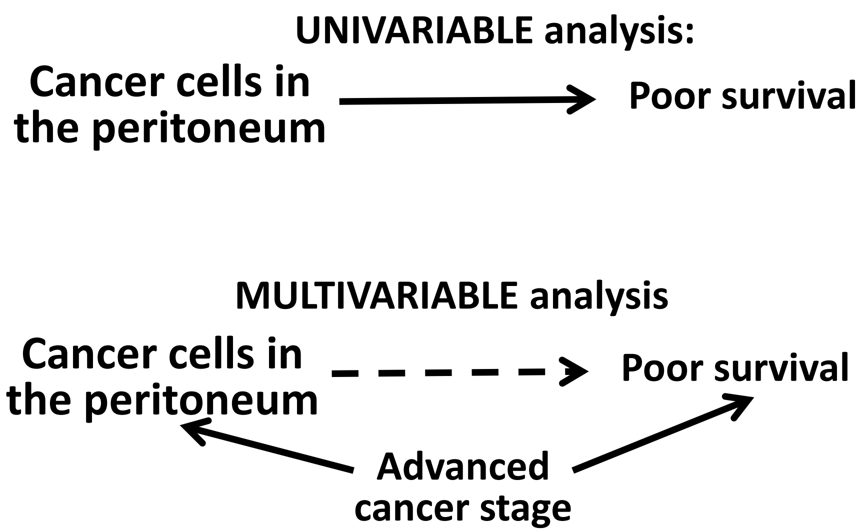
UNIVARIABLE analysis:
enough for a thesis

One X → One Y

MULTIVARIABLE analysis:
necessary for a scientific article

X1 →
X2 → One Y
X3 →

In gastric cancer patients



de Manzoni G, Verlato G, et al. Peritoneal cytology does not increase the prognostic information provided by TNM in gastric cancer. World J Surg 2006; 30: 579-584

Confounding bias:

In the *Verona Diabetes Study* diabetic women experienced about the same mortality as men (Relative Risk = 0.97, CI 95% 0.88-1.07).

Diabetic women were older than diabetic men. Women were 68.3±12.2 years old (mean ± standard deviation) while men were 62.2±13.0 years old.

Hence, when age is taken into account in a multivariable model, diabetic women had a lower mortality than men (Relative Risk = 0.64, CI 95% 0.58-0.71).

STAGE MIGRATION is a type of information bias: if the surgical procedure is extended, the tumor is better staged and often becomes apparently more advanced

BIVARIATE ANALYSIS

In patients undergoing gastrectomy for gastric cancer

Excised lymph nodes ⟷ metastatic lymph nodes
T di Kendall = 0.192
P<0.001

N.B. A non-parametric statistics (Kendall's tau, Kendall rank correlation coefficient) is used instead of a parametric statistics (r), as the distributions of excised and metastatic lymph nodes are highly skewed.

Excised lymph nodes ⟷ metastatic lymph nodes: **T = 0.192, P<0.001**
Excised lymph nodes ⟷ T tier: **T = 0.151, P<0.001**
metastatic lymph nodes ⟷ T tier: **T = 0.525, P<0.001**

T tier (depth of tumor invasion) strongly correlates with the number of metastatic (positive) lymph nodes, but also slightly with the number of excised nodes. Could it be a confounder ?

TRIVARIATE ANALYSIS

Excised lymph nodes ⟷ metastatic lymphnodes: **T = 0.134, P<0.001**
Excised lymph nodes ⟷ T tier: **T = 0.060, P=0.006**
metastatic lymphnodes ⟷ T tier: **T = 0.511, P<0.001**

When controlling for T stage by the Kendall's PARTIAL rank correlation coefficient, the correlation between excised and metastatic lymph nodes becomes weaker but does not disappear.

In conclusion, the number of metastatic lymph nodes increases with increasing number of excised lymph nodes, even if tumor stage is kept constant through a statistical technique. Stage migration is further supported.

G de Manzoni, G Verlato, et al (2003) The new TNM classification of lymph node metastasis minimises stage migration problems in gastric cancer patients. Br J Cancer 87: 171-174

QUANTITATIVE VARIABLES

Univariable analysis: 1 Y, 1 X

1) Simple linear regression

$y = \beta_0 + \beta_1 x + \varepsilon$, *where X both Y are both quantitative variables*

Multivariable analysis: 1 Y, several Xs

2) Multiple linear regression

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, *where Xs and Y are both quantitative variables*

3) Analysis of Variance (ANOVA)

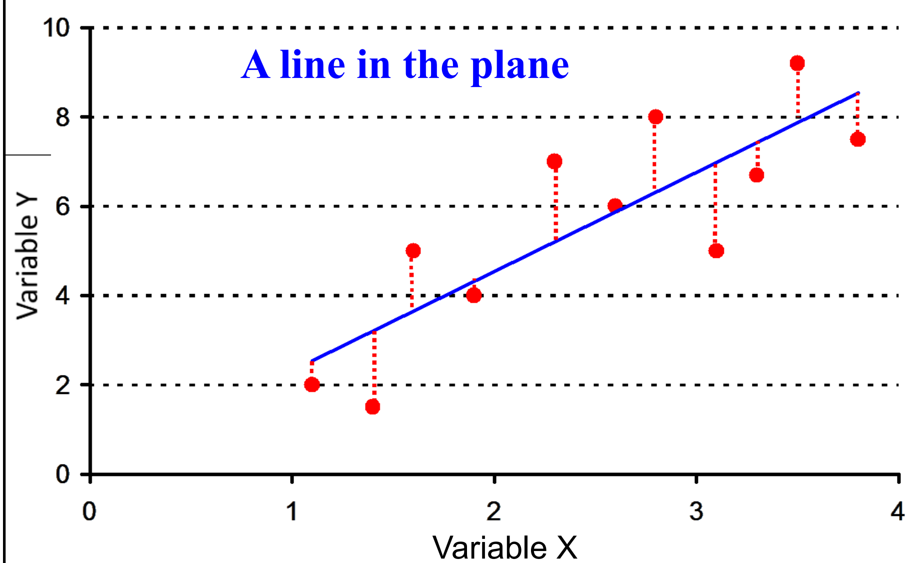
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, *where Y quantitative, Xs qualitative*

4) Analysis of covariance (ANCOVA)

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, *where Y quantitative, Xs both qualitative and quantitative*

Model of simple linear regression

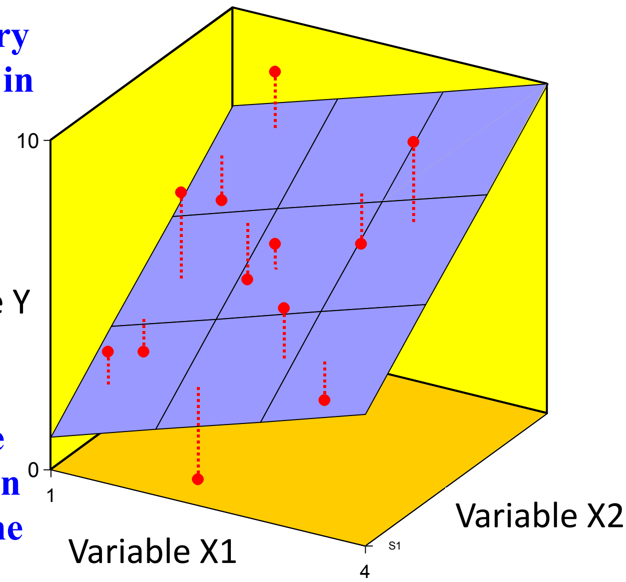
$$y = \beta_0 + \beta x + \varepsilon$$



Multiple linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

With 2 explanatory variables, a plane in the space.



With several variables, multiple linear regression can be represented in the hyperspace

And with NOMINAL VARIABLES ?

2 VARIABLES (both qualitative):
Chi-squared test, Fisher's exact test

3 qualitative VARIABLES (2 variables + 1 stratification factor): Mantel-Haenszel test

SEVERAL VARIABLES:

y dichotomous (healthy/sick) → LOGISTIC model

y polytomous (never smoker, ex-smoker, current smoker) → MULTINOMIAL model

LOGISTIC REGRESSION MODEL

19 / (19+132)
0 / (0+9)
11 / (11+52)
6 / (6+97)

prevalence

LOG-LINEAR MODEL

19	132
0	9
11	52
6	97

counts

POISSON'S MODEL

19 / 1510 persone-anno
0 / 90 persone-anno
11 / 630 persone-anno
6 / 103 persone-anno

Incidence rates

SURVIVAL ANALYSIS

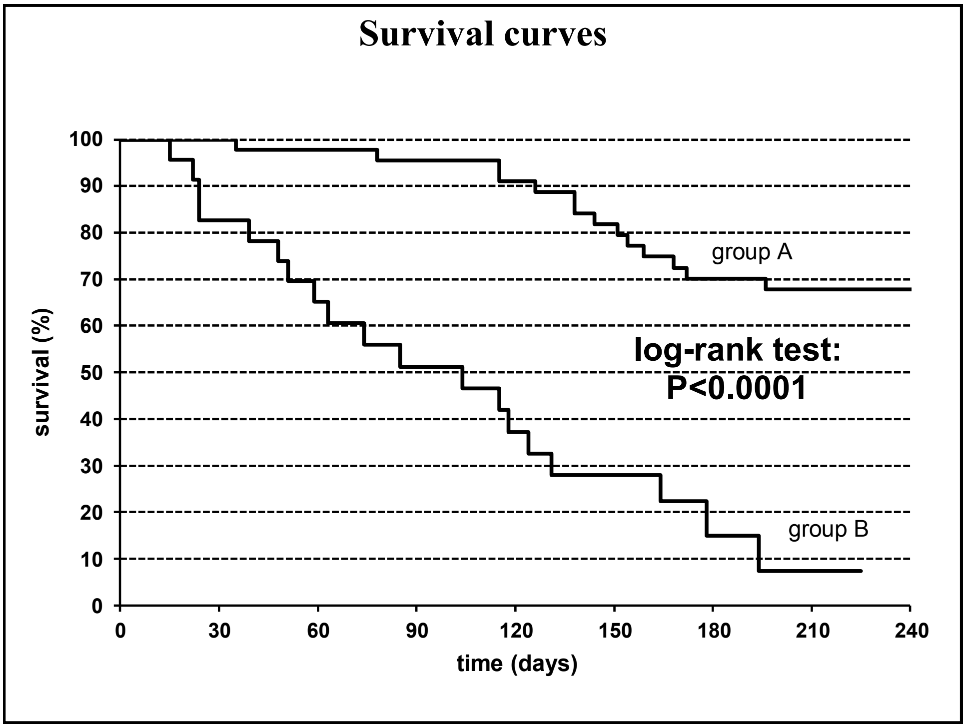
1) UNIVARIABLE ANALYSIS

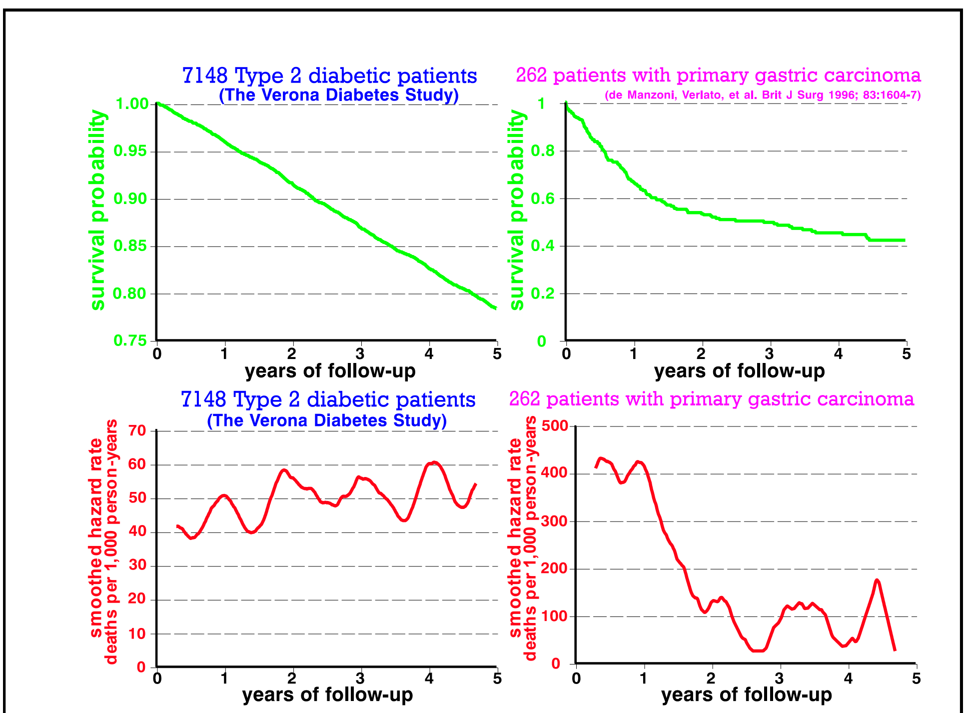
Estimating survival curves: method by Kaplan-Meier

Evaluating significance of differences among survival curves: log-rank test

2) MULTIVARIABLE ANALYSIS

Evaluating the combined prognostic significance of several factors: Cox model


Survival curves
survival (%)
100
90
80
70
60
50
40
30
20
10
0
0
30
60
90
120
150
180
210
240
time (days)
group A
group B
log-rank test:
P<0.0001


7148 Type 2 diabetic patients
(The Verona Diabetes Study)
survival probability
1.00
0.95
0.90
0.85
0.80
0.75
years of follow-up
262 patients with primary gastric carcinoma
(de Manzoni, Verlato, et al. Brit J Surg 1996; 83:1604-7)
survival probability
1
0.8
0.6
0.4
0.2
0
years of follow-up
7148 Type 2 diabetic patients
(The Verona Diabetes Study)
smoothed hazard rate
deaths per 1,000 person-years
70
60
50
40
30
20
10
0
years of follow-up
262 patients with primary gastric carcinoma
smoothed hazard rate
deaths per 1,000 person-years
500
400
300
200
100
0
years of follow-up