MULTIPLE LINEAR REGRESSION

and its applications in Bioinformatics

Presenter: Duy Nguyen
Supervisor: PD Dr. Michael Hutter

CONTENT

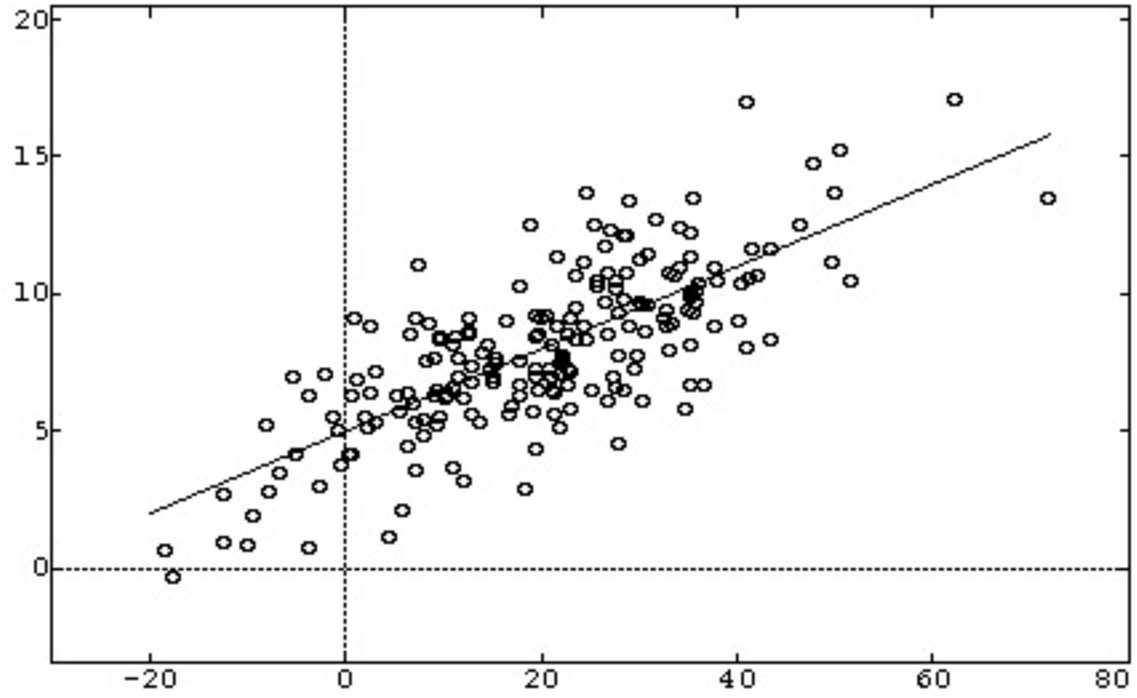
- Introduction
- How to derive a MLR formula
- Validation of MLR model
- Application in Bioinformatics
- Summary

INTRODUCTION

- Good at math => good at statistics as well?
- Expensive + good-looking + popular brand => good quality?
- Do age & smoking habit contribute to heart disease?

INTRODUCTION

Regression analysis (RA): observe the relationship between 1 dependent variable & several independent variables



http://en.wikipedia.org/wiki/Regression_analysis

$$y_i = b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + b_0 + \varepsilon_i$$

$b_1, b_2 .. b_k$: regression coefficients

b_0: constant (intercept)

ε_i : error

DERIVE REGRESSION EQUATION

$$y_i = b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + b_0 + \varepsilon_i$$

$$y'_i = b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + b_0$$

We obtain best fitting line when:

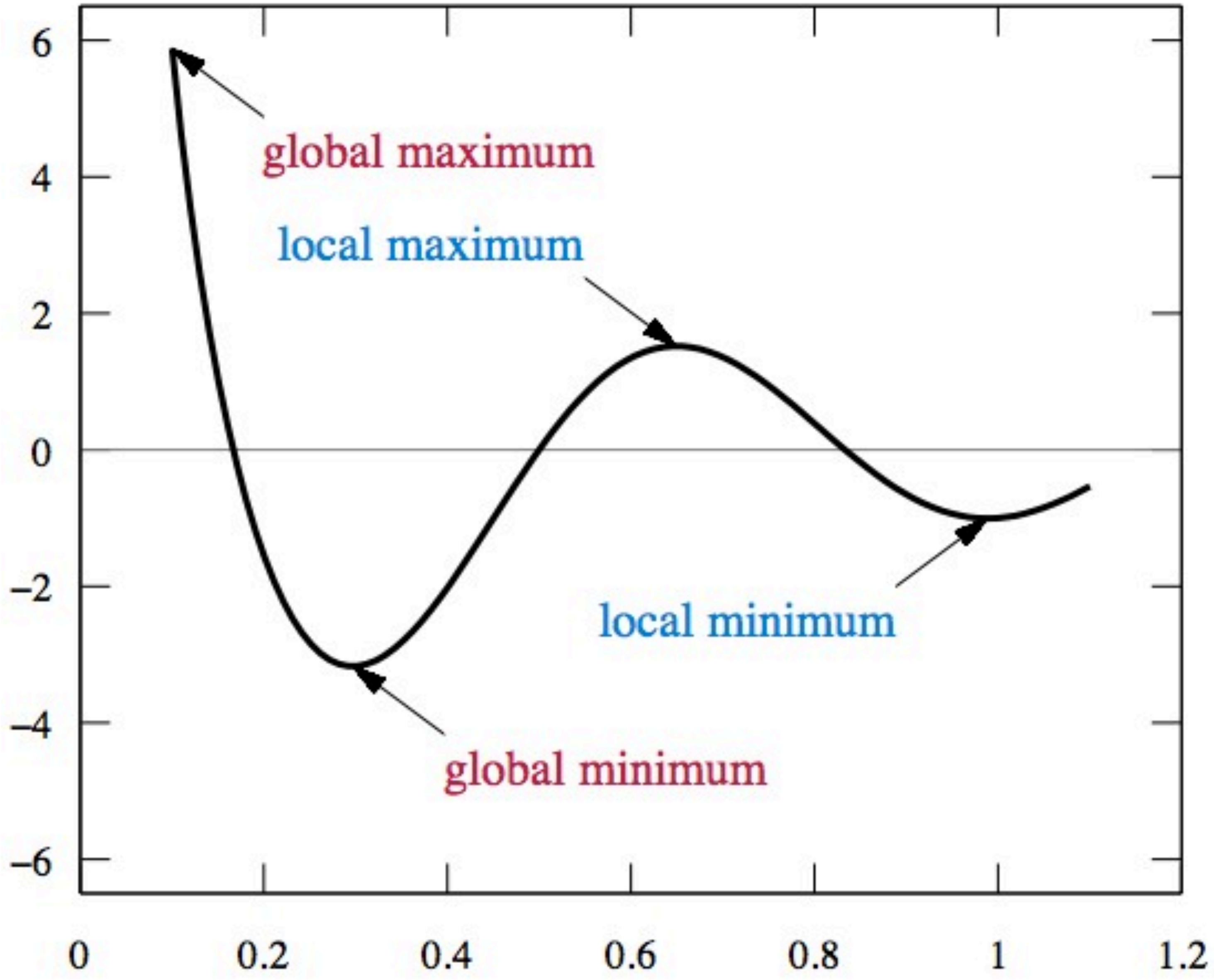
$$\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} \left(y_i - y'_i \right)^2 \text{ is a minimum}$$

DERIVE REGRESSION EQUATION

How to find *minima* of a function?

$$f'(x) = 0 \Rightarrow x_0$$

$$f''(x_0) > 0$$



http://en.wikipedia.org/wiki/Maxima_and_minima

DERIVE REGRESSION EQUATION

$$G = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} \left(y_i - y'_i\right)^2$$

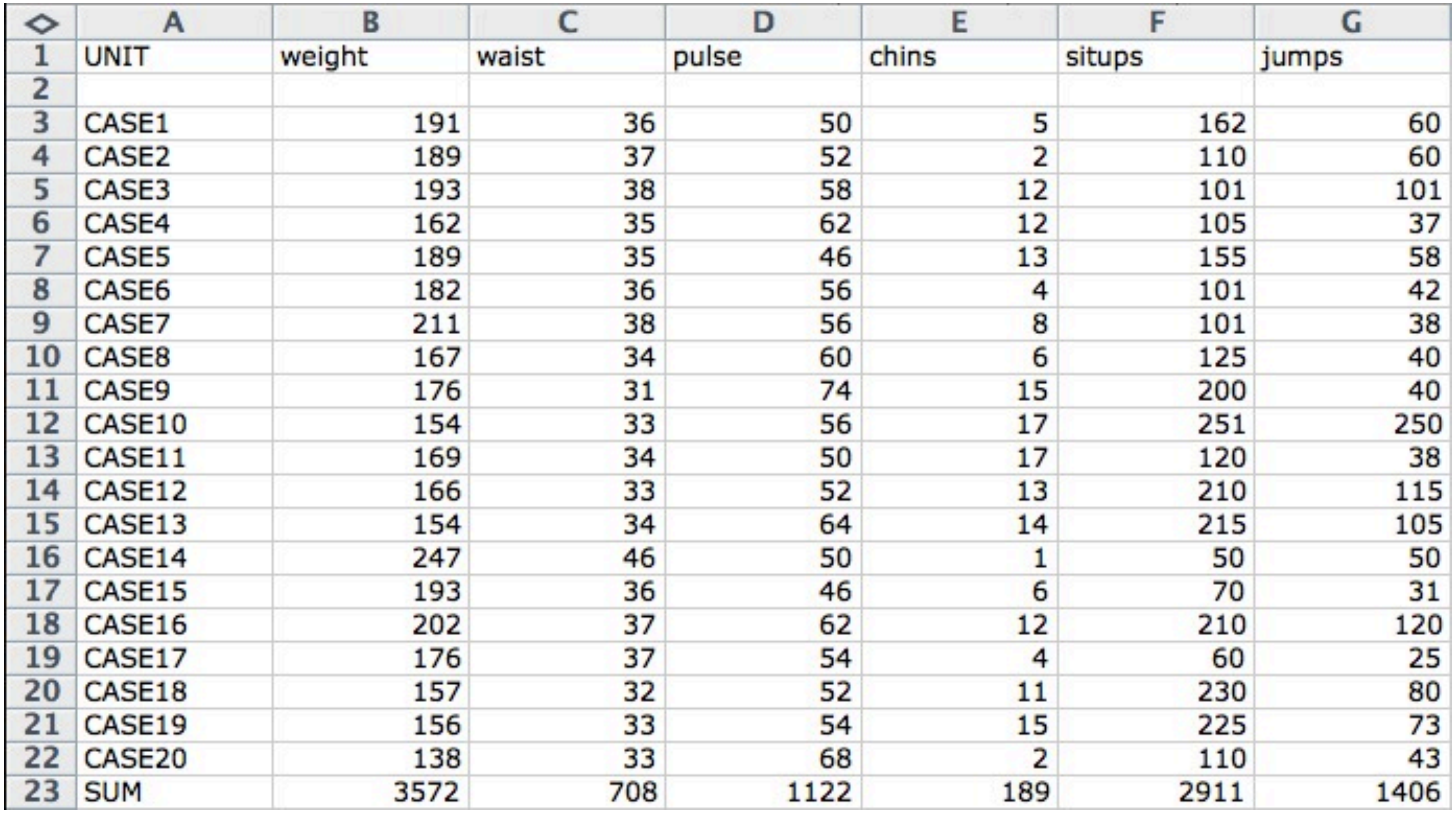
- Apply partial derivative for every regression coefficient in G

$$\frac{\delta G}{\delta b_i} := 0 \Rightarrow \text{found } b_i$$

$$\text{second derivative } > 0 \text{ if } \sum x_i \neq 0$$

DERIVE REGRESSION EQUATION

- Example (Bill Miller’s DataFiles)



◇	A	B	C	D	E	F	G
1	UNIT	weight	waist	pulse	chins	situps	jumps
2							
3	CASE1	191	36	50	5	162	60
4	CASE2	189	37	52	2	110	60
5	CASE3	193	38	58	12	101	101
6	CASE4	162	35	62	12	105	37
7	CASE5	189	35	46	13	155	58
8	CASE6	182	36	56	4	101	42
9	CASE7	211	38	56	8	101	38
10	CASE8	167	34	60	6	125	40
11	CASE9	176	31	74	15	200	40
12	CASE10	154	33	56	17	251	250
13	CASE11	169	34	50	17	120	38
14	CASE12	166	33	52	13	210	115
15	CASE13	154	34	64	14	215	105
16	CASE14	247	46	50	1	50	50
17	CASE15	193	36	46	6	70	31
18	CASE16	202	37	62	12	210	120
19	CASE17	176	37	54	4	60	25
20	CASE18	157	32	52	11	230	80
21	CASE19	156	33	54	15	225	73
22	CASE20	138	33	68	2	110	43
23	SUM	3572	708	1122	189	2911	1406

DERIVE REGRESSION EQUATION

- For simplification, only 2 independent variables are chosen: weight & waist

$$jumps = a * weight + b * waist + c$$

$$y' = ax_1 + bx_2 + c$$

DERIVE REGRESSION EQUATION

$$G = \sum_{i=1}^{20} \left(y_i - y'_i \right)^2 = \sum_{i=1}^{20} \left(y_i - \left(a x_{i1} + b x_{i2} + c \right) \right)^2 \text{ is minimum}$$

$$\frac{\delta \mathrm{G}}{\delta \mathrm{a}} = 2 \sum_{i=1}^{20} \left(y_i - \left(a x_{i1} + b x_{i2} + c \right) \right) \left(-x_{i1} \right) = 0$$

$$\frac{\delta \mathrm{G}}{\delta b} = 2 \sum_{i=1}^{20} \left(y_i - \left(a x_{i1} + b x_{i2} + c \right) \right) \left(-x_{i2} \right) = 0$$

$$\frac{\delta \mathrm{G}}{\delta \mathrm{c}} = 2 \sum_{i=1}^{20} \left(y_i - \left(a x_{i1} + b x_{i2} + c \right) \right) \left(-1 \right) \quad = 0$$

DERIVE REGRESSION EQUATION

$$a\sum_{i=1}^{20} x_{i1}^2 + b\sum_{i=1}^{20} x_{i1}x_{i2} + c\sum_{i=1}^{20} x_{i1} = \sum_{i=1}^{20} y_i x_{i1}$$

$$a\sum_{i=1}^{20} x_{i1}x_{i2} + b\sum_{i=1}^{20} x_{i2}^2 + c\sum_{i=1}^{20} x_{i2} = \sum_{i=1}^{20} y_i x_{i2}$$

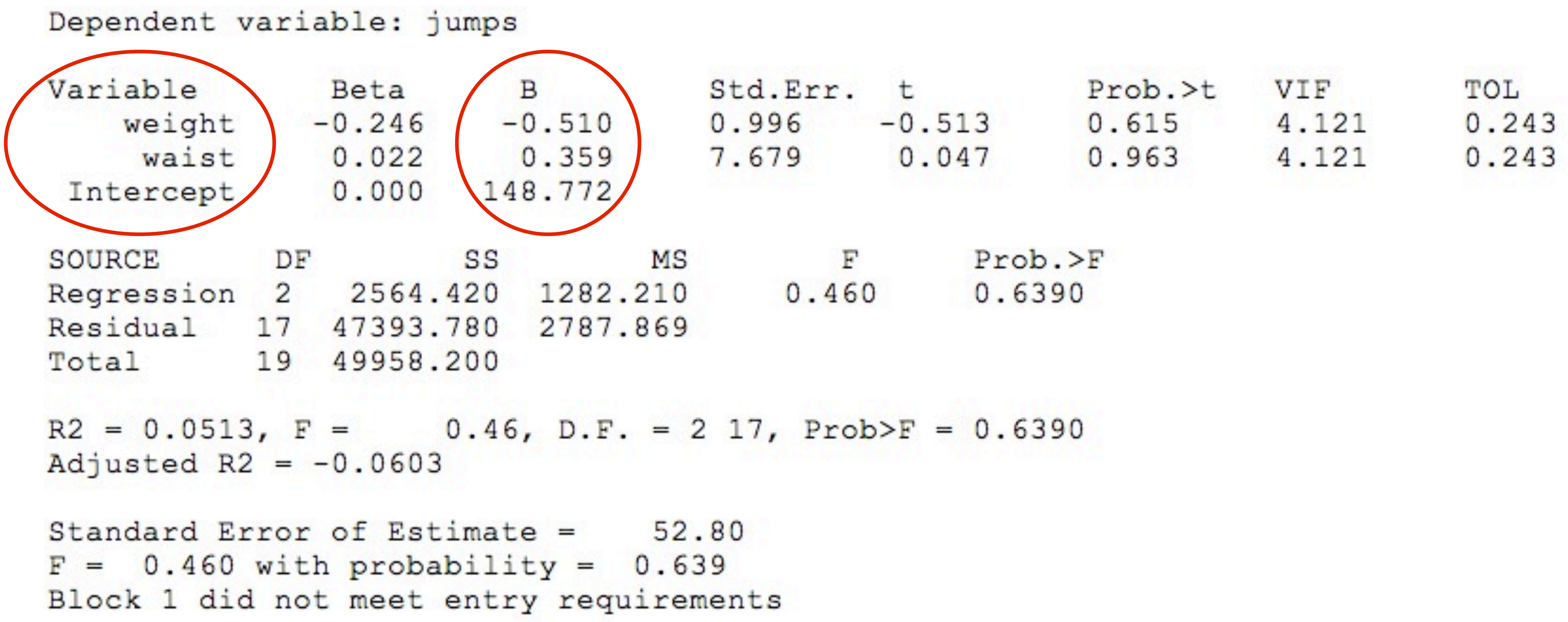
$$a\sum_{i=1}^{20} x_{i1} + b\sum_{i=1}^{20} x_{i2} + \sum_{i=1}^{20} c = \sum_{i=1}^{20} y_i$$

Data →

$$\sum_{i=1}^{20} x_{i1}^2 = 649542 \quad \sum_{i=1}^{20} x_{i2}^2 = 25258 \quad \sum_{i=1}^{20} x_{i1} = 3572 \quad \sum_{i=1}^{20} x_{i2} = 708$$

$$\sum_{i=1}^{20} y_i = 1406 \quad \sum_{i=1}^{20} x_{i1}x_{i2} = 127756 \quad \sum_{i=1}^{20} y_i x_{i1} = 245668 \quad \sum_{i=1}^{20} y_i x_{i2} = 49175$$

DERIVE REGRESSION EQUATION



```
Dependent variable: jumps

Variable        Beta          B        Std.Err.   t          Prob.>t   VIF        TOL
    weight    -0.246     -0.510        0.996     -0.513      0.615     4.121      0.243
     waist     0.022      0.359        7.679      0.047      0.963     4.121      0.243
 Intercept     0.000    148.772

SOURCE       DF          SS          MS          F       Prob.>F
Regression    2    2564.420    1282.210      0.460      0.6390
Residual     17   47393.780    2787.869
Total        19   49958.200

R2 = 0.0513, F =      0.46, D.F. = 2 17, Prob>F = 0.6390
Adjusted R2 = -0.0603

Standard Error of Estimate =     52.80
F =  0.460 with probability =   0.639
Block 1 did not meet entry requirements
```

$$y' = -0.51x_1 + 0.359x_2 + 148.772$$

VALIDATION OF MLR MODEL

$$SS_{total} = SS_{reg} + SS_{error}$$

$$= \sum_{i=1}^{n} \left(y - \overline{y}\right)^2$$

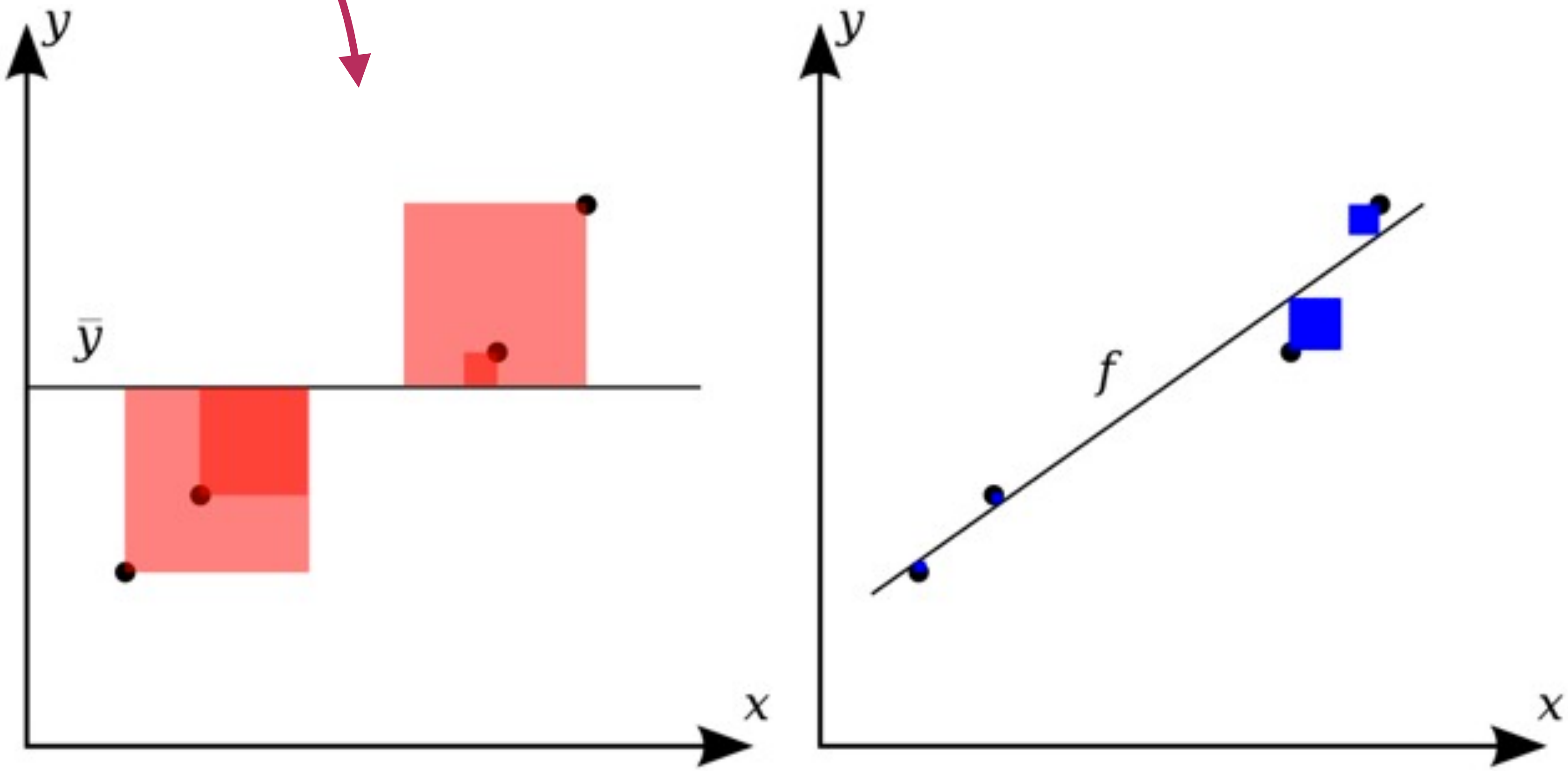
Total sum of squares

$$SS_{reg} = \sum_{i=1}^{n} \left(y' - \overline{y}\right)^2$$

Regression sum of squares

$$SS_{error} = \sum_{i=1}^{n} \left(y - y'\right)^2$$

Residual sum of squares



http://en.wikipedia.org/wiki/Coefficient_of_determination

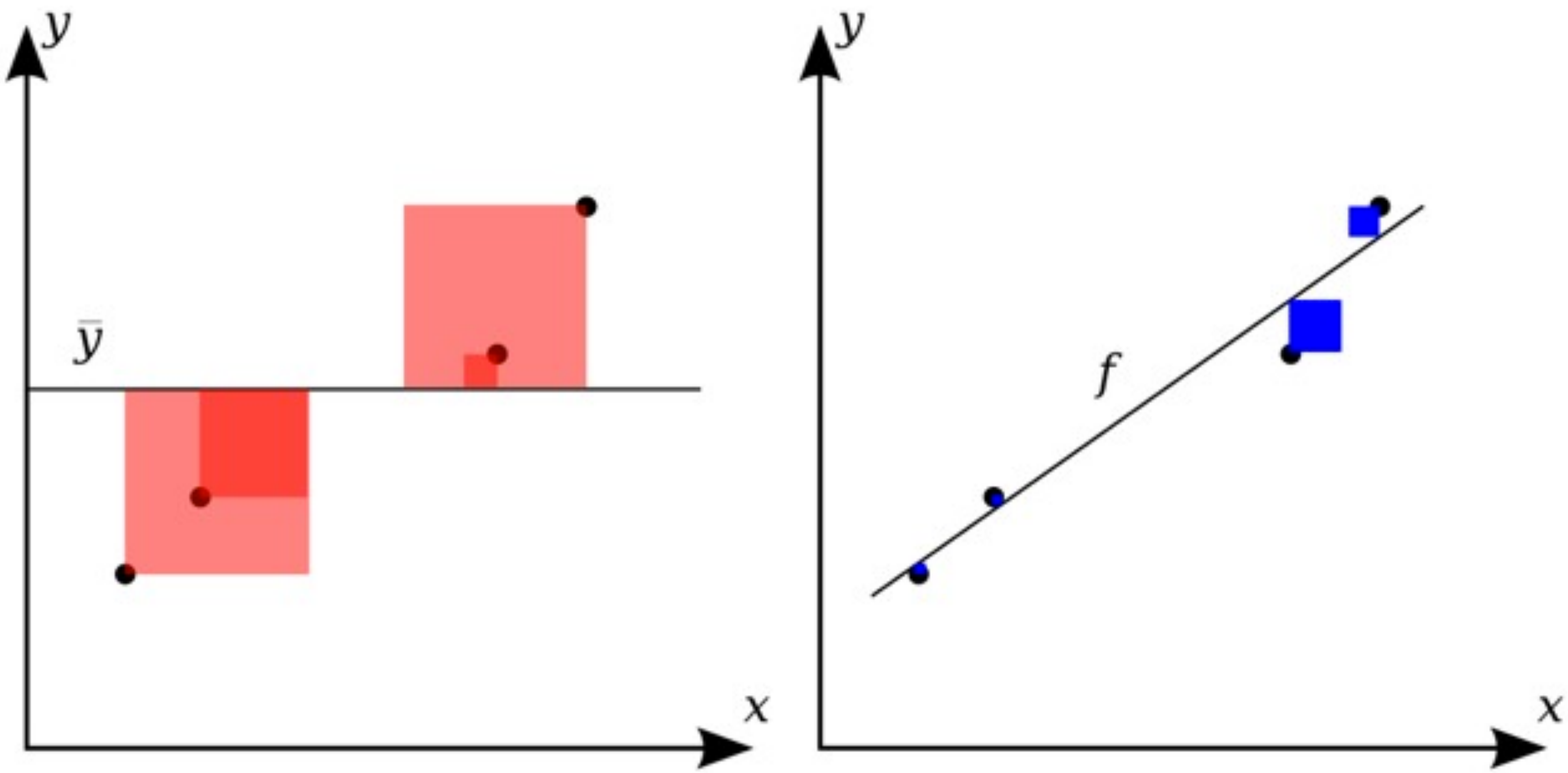
VALIDATION OF MLR MODEL

$$R^2 = 1 - \frac{\boxed{SS_{error}}}{\boxed{SS_{tot}}}$$

$$= \frac{SS_{reg}}{SS_{tot}}$$

Coefficient of determination

worst $0 \leq R^2 \leq 1$ best



http://en.wikipedia.org/wiki/Coefficient_of_determination

VALIDATION OF MLR MODEL

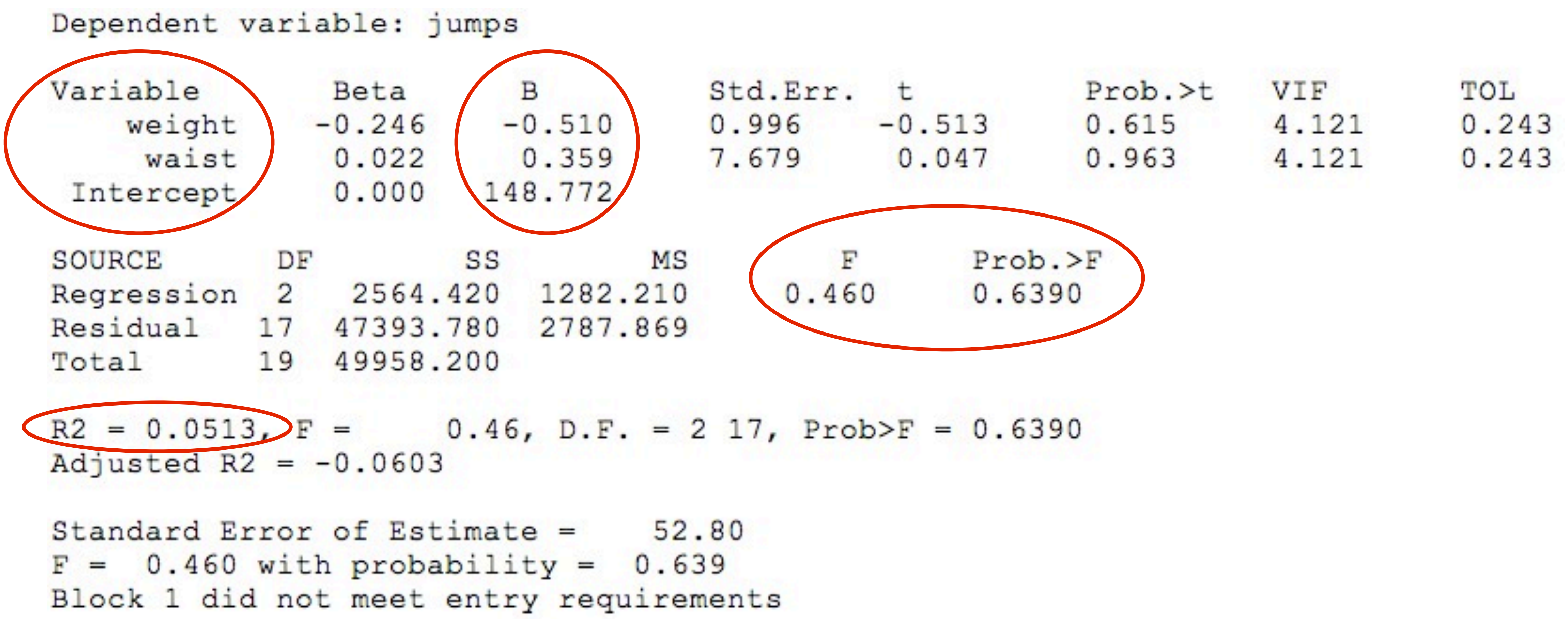
- F ratio test: test the regression model fits data well or not

$$F = \frac{\text{explained variance}}{\text{unexplained variance}} = \frac{SS_{reg}/K}{SS_{error}/(N-K-1)}$$

K : number of independent variables

N : number of samples

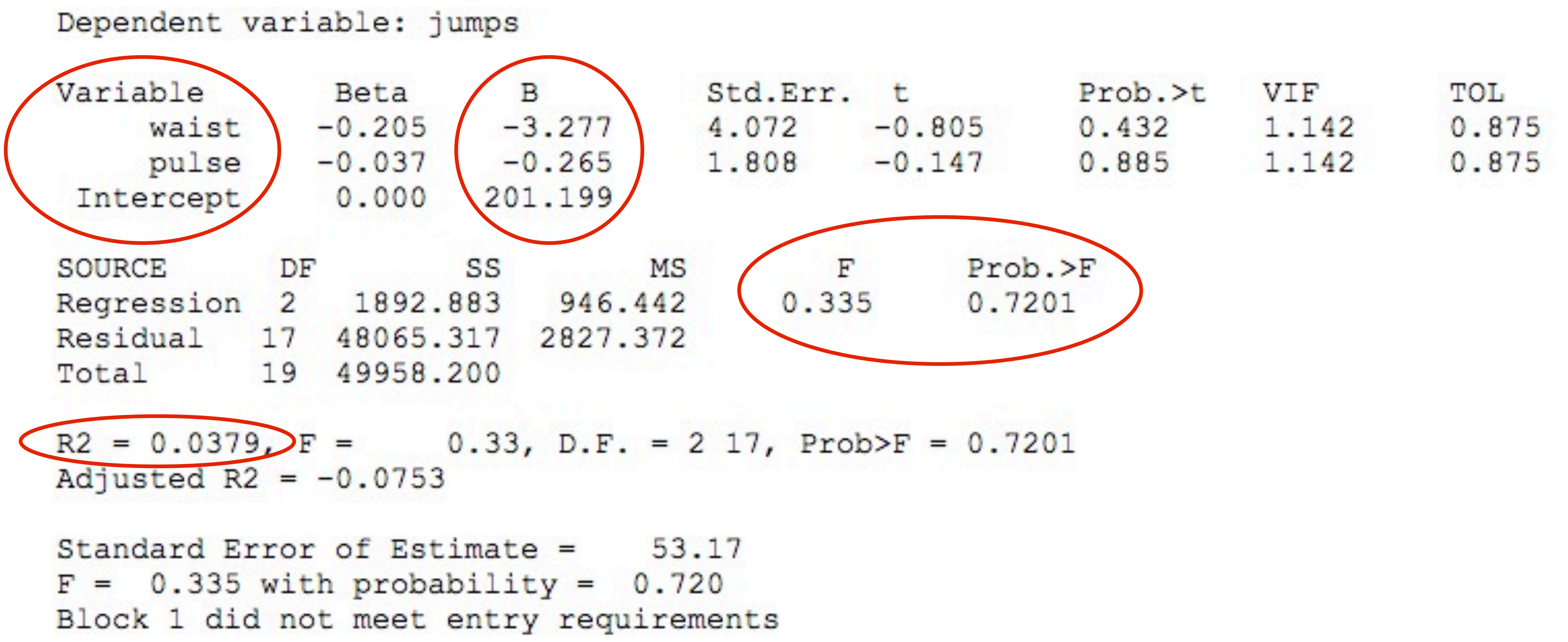
VALIDATION OF MLR MODEL



```
Dependent variable: jumps

Variable      Beta        B        Std.Err.   t          Prob.>t   VIF        TOL
    weight   -0.246    -0.510     0.996     -0.513      0.615     4.121      0.243
     waist    0.022     0.359     7.679      0.047      0.963     4.121      0.243
 Intercept    0.000   148.772

SOURCE       DF         SS          MS          F         Prob.>F
Regression    2    2564.420    1282.210      0.460      0.6390
Residual     17   47393.780    2787.869
Total        19   49958.200

R2 = 0.0513, F =     0.46, D.F. = 2 17, Prob>F = 0.6390
Adjusted R2 = -0.0603

Standard Error of Estimate =    52.80
F =  0.460 with probability =  0.639
Block 1 did not meet entry requirements
```

$$y' = -0.51x_1 + 0.359x_2 + 148.772$$

VALIDATION OF MLR MODEL



```
Dependent variable: jumps

Variable       Beta       B        Std.Err.   t          Prob.>t   VIF       TOL
     waist    -0.205    -3.277    4.072     -0.805      0.432     1.142     0.875
     pulse    -0.037    -0.265    1.808     -0.147      0.885     1.142     0.875
 Intercept     0.000   201.199

SOURCE       DF        SS         MS          F          Prob.>F
Regression   2    1892.883    946.442      0.335       0.7201
Residual    17   48065.317   2827.372
Total       19   49958.200

R2 = 0.0379, F =      0.33, D.F. = 2 17, Prob>F = 0.7201
Adjusted R2 = -0.0753

Standard Error of Estimate =    53.17
F =  0.335 with probability =  0.720
Block 1 did not meet entry requirements
```

$$y' = -3.277x_2 - 0.256x_3 + 201.199$$

VALIDATION OF MLR MODEL

- Root mean square deviation (RMSD)

$$\mathrm{RMSD}\left(\hat{\theta}\right) = \sqrt{MSE\left(\hat{\theta}\right)} = \sqrt{E\left(\left(\hat{\theta} - \theta\right)^2\right)}$$

$\hat{\theta}$: values predicted from regression model

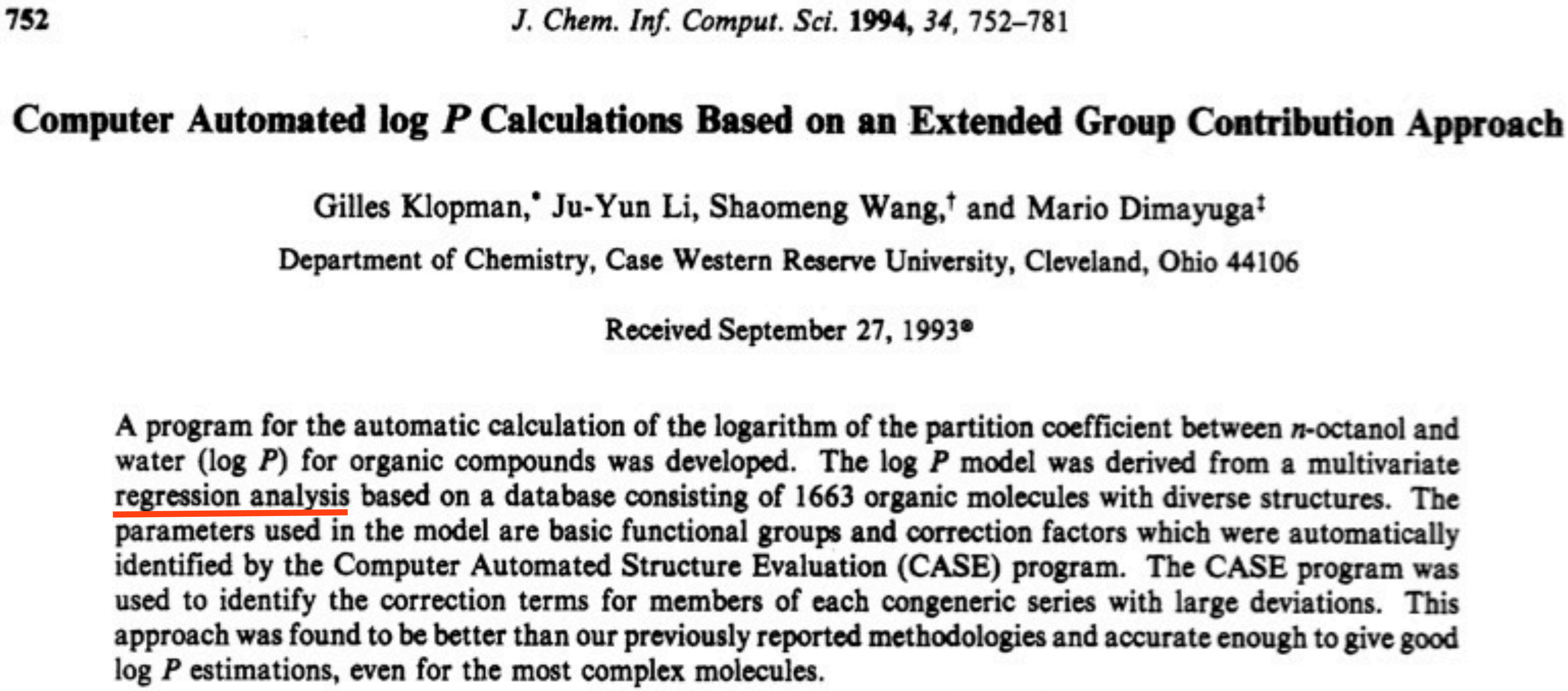
θ : values from observation

- Also used in structural bioinformatics

APPLICATION IN BIOINFORMATICS

- QSAR/QSPR equations
    - a tool for computer-based drug design if there is no structural information of target
    - form a quantitative relation between chemical structure and biological activity
    - need a huge number of measured data from experiments

APPLICATION IN BIOINFORMATICS



752 J. Chem. Inf. Comput. Sci. 1994, 34, 752–781

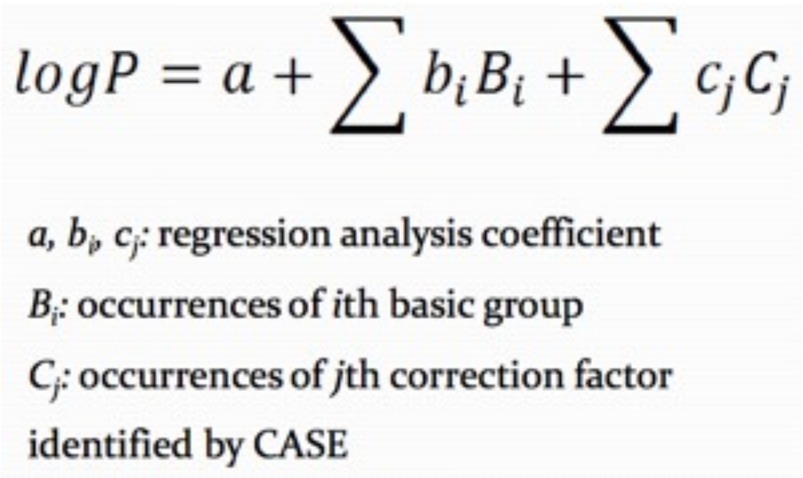
Computer Automated log *P* Calculations Based on an Extended Group Contribution Approach

Gilles Klopman,* Ju-Yun Li, Shaomeng Wang,† and Mario Dimayuga‡

Department of Chemistry, Case Western Reserve University, Cleveland, Ohio 44106

Received September 27, 1993®

A program for the automatic calculation of the logarithm of the partition coefficient between *n*-octanol and water (log *P*) for organic compounds was developed. The log *P* model was derived from a multivariate regression analysis based on a database consisting of 1663 organic molecules with diverse structures. The parameters used in the model are basic functional groups and correction factors which were automatically identified by the Computer Automated Structure Evaluation (CASE) program. The CASE program was used to identify the correction terms for members of each congeneric series with large deviations. This approach was found to be better than our previously reported methodologies and accurate enough to give good log *P* estimations, even for the most complex molecules.

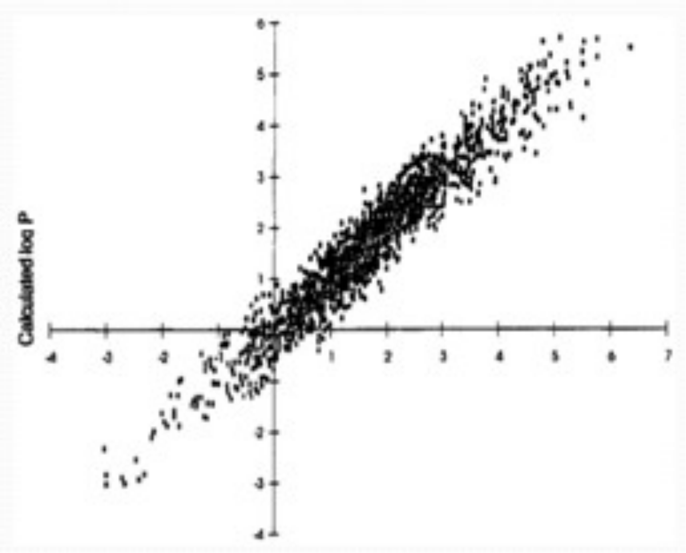


$$logP = a + \sum b_i B_i + \sum c_j C_j$$

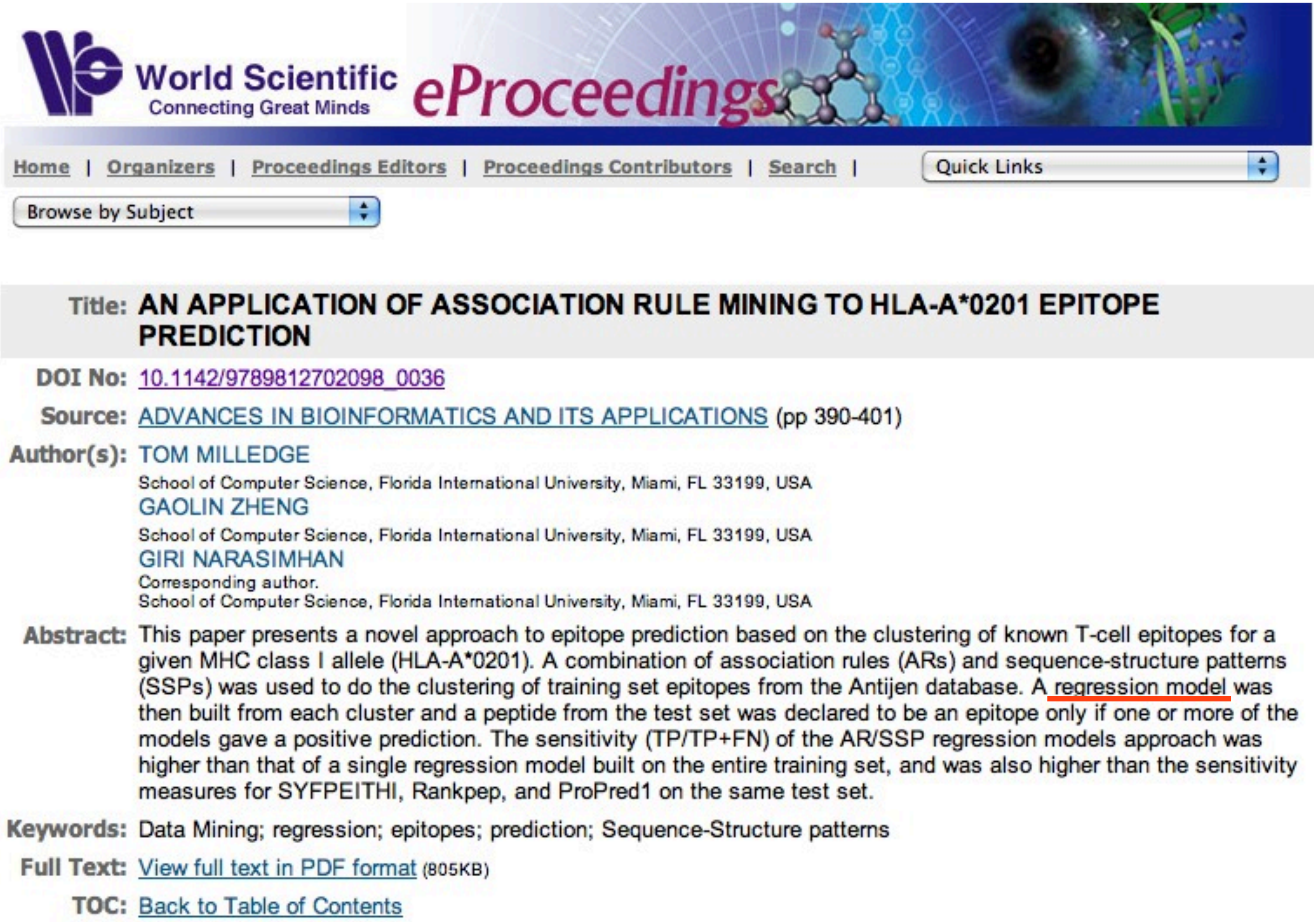
a, b_i, c_j: regression analysis coefficient

B_i: occurrences of ith basic group

C_j: occurrences of jth correction factor identified by CASE



APPLICATION IN BIOINFORMATICS



Title: AN APPLICATION OF ASSOCIATION RULE MINING TO HLA-A*0201 EPITOPE PREDICTION

DOI No: 10.1142/9789812702098_0036

Source: ADVANCES IN BIOINFORMATICS AND ITS APPLICATIONS (pp 390-401)

Author(s): TOM MILLEDGE
School of Computer Science, Florida International University, Miami, FL 33199, USA
GAOLIN ZHENG
School of Computer Science, Florida International University, Miami, FL 33199, USA
GIRI NARASIMHAN
Corresponding author.
School of Computer Science, Florida International University, Miami, FL 33199, USA

Abstract: This paper presents a novel approach to epitope prediction based on the clustering of known T-cell epitopes for a given MHC class I allele (HLA-A*0201). A combination of association rules (ARs) and sequence-structure patterns (SSPs) was used to do the clustering of training set epitopes from the Antijen database. A regression model was then built from each cluster and a peptide from the test set was declared to be an epitope only if one or more of the models gave a positive prediction. The sensitivity (TP/TP+FN) of the AR/SSP regression models approach was higher than that of a single regression model built on the entire training set, and was also higher than the sensitivity measures for SYFPEITHI, Rankpep, and ProPred1 on the same test set.

Keywords: Data Mining; regression; epitopes; prediction; Sequence-Structure patterns

Full Text: View full text in PDF format (805KB)

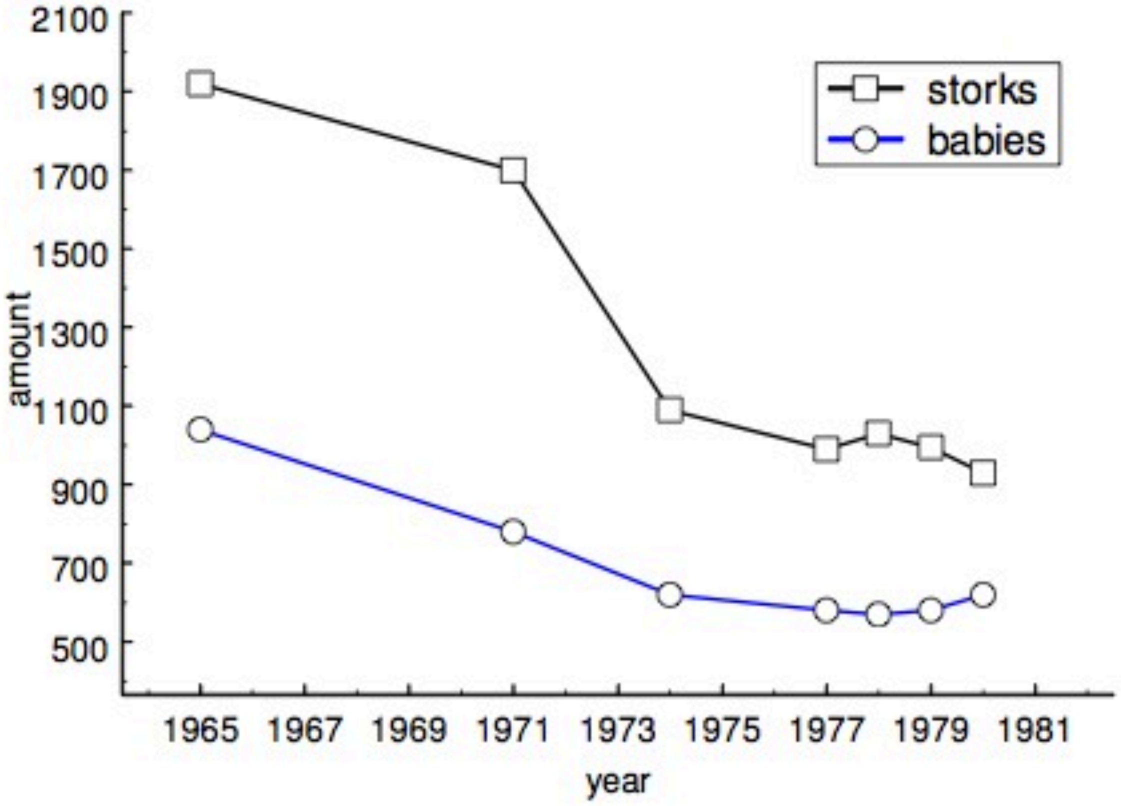
TOC: Back to Table of Contents

BEWARE!

- The dependent variable & independent variables should have reasonable relationship

Example: H. Sies *Nature* **332** (1988) 495.

Scientific proof that babies are delivered by storks



("Modern Methods in Drug Discovery" lecture notes)

SUMMARY

- RE is a statistical method which observes relation between a dependent variable & several independent variables

- The model is a linear function

- The dependent variable & independent variables should have reasonable relationship

- Can be validated by F-test & RMSD

- QSAR/QSPR is one of applications in bioinformatics

THANK YOU!

- http://statpages.org/miller/openstat/
- http://www.wikipedia.org/
- “Modern Methods in Drug Discovery” lecture notes