Simple Regression

Prof. Daniel A. Menasce
Dept. of Computer Science
George Mason University

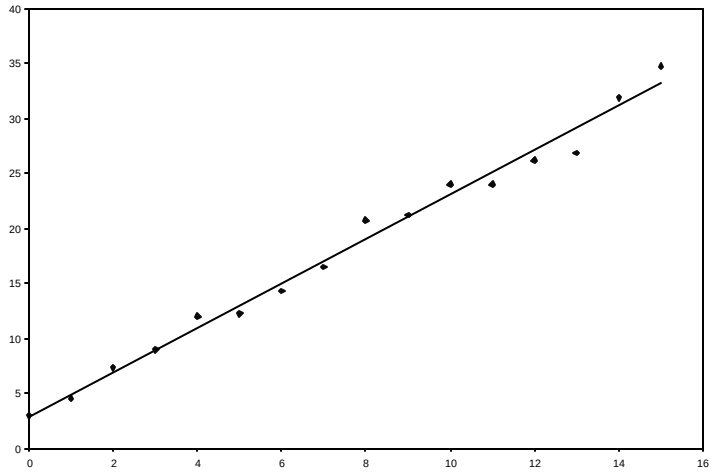
© 2001. D. A. Menascé. All Rights Reserved.

Basics

- Purpose of regression analysis: predict the value of a dependent or response variable from the values of at least one explanatory or independent variable.
- Purpose of correlation analysis: measure the strength of the correlation between two variables.

© 2001. D. A. Menascé. All Rights Reserved.

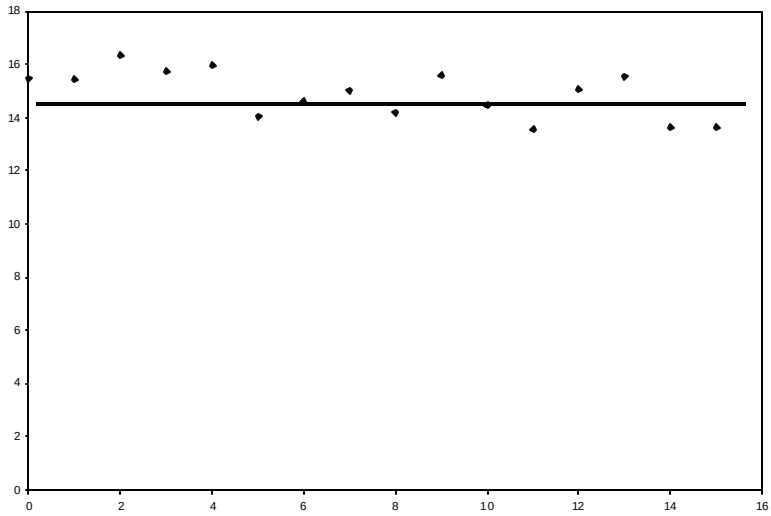
Linear Relationship



3

© 2001. D. A. Menascé. All Rights Reserved.

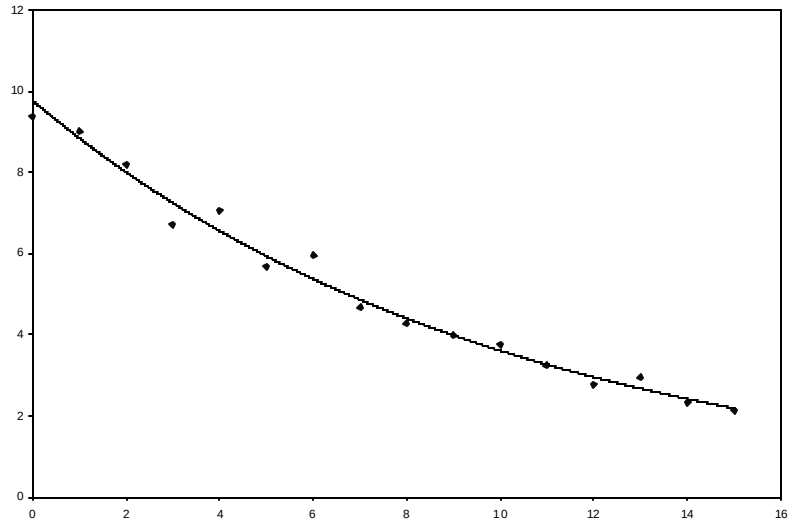
No Relationship



4

© 2001. D. A. Menascé. All Rights Reserved.

Negative Curvilinear



5

© 2001. D. A. Menascé. All Rights Reserved.

Linear Regression

$$\hat{Y}_i = b_0 + b_1 X_i$$

$\hat{Y}_i$: predicted value of Y for observation i.

X_i : value of observation i.

b_0 and b_1 are chosen to minimize:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

6

© 2001. D. A. Menascé. All Rights Reserved.

Method of Least Squares

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2$$

7

© 2001. D. A. Menascé. All Rights Reserved.

Number of I/Os (x)	CPU Time (y)	Estimate (0.0408*x +0.0508)	Error	Error Squared	SSY
1	0.092	0.092	0.0005	0.00000	0.00848
2	0.134	0.132	0.0013	0.00000	0.017882
3	0.165	0.173	-0.0084	0.00007	0.027173
4	0.211	0.214	-0.0027	0.00001	0.044645
5	0.242	0.255	-0.0129	0.00017	0.058505
6	0.302	0.296	0.0066	0.00004	0.091331
7	0.357	0.336	0.0204	0.00042	0.127331
8	0.401	0.377	0.0238	0.00056	0.160771
9	0.405	0.418	-0.0133	0.00018	0.163795
10	0.442	0.459	-0.0163	0.00027	0.195783
	0.275			0.00172	0.89570

SST	0.1388841
SSR	0.1371690
R2	0.9876514

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \left(\sum_{i=1}^{n} Y_i^2 \right) - n\overline{Y}^2 = SSY - SS0$$

SSE SSY

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y} \right)^2 = SST - SSE$$

The higher the value of R^2 the better the regression.

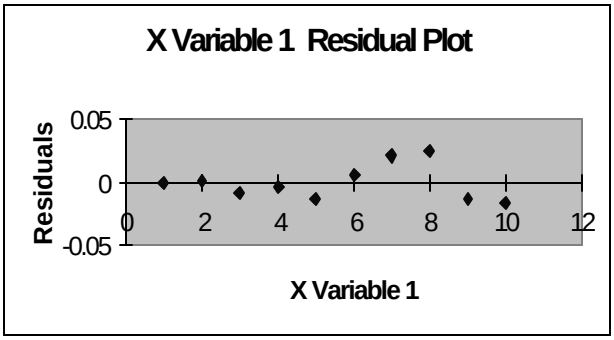
$R^2 = \frac{SSR}{SST}$ coefficient of determination.

8

© 2001. D. A. Menascé. All Rights Reserved.

Steps in Conduction Linear Regression

1. Do a scatter plot for X and Y.
2. Plot the residuals to check for apparent violations of equal variance at each level of X.

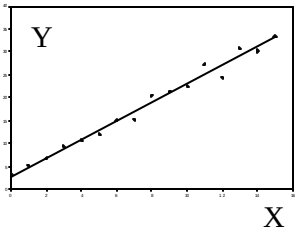


9

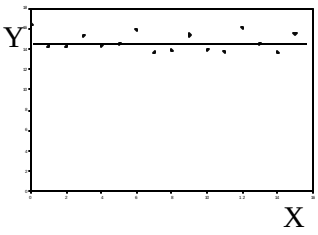
© 2001. D. A. Menascé. All Rights Reserved.

Correlation

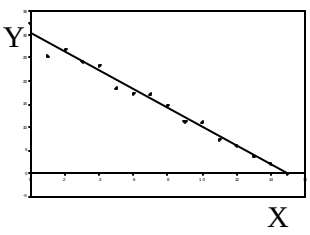
- Goal: measure the strength of the relationship between two variables measured by the coefficient of correlation ρ.



Perfect positive correlation (ρ=1)



No correlation (ρ=0)



Perfect negative correlation (ρ=-1)

10

© 2001. D. A. Menascé. All Rights Reserved.

Coefficient of Correlation

Sample coefficient of correlation: $R = \sqrt{R^2}$

The sign of R is the same as that of b_1.

$$R = \frac{SSXY}{\sqrt{SSX}\sqrt{SSY}}$$

$$SSXY = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

$$SSX = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$SSY = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

11

© 2001. D. A. Menascé. All Rights Reserved.

Testing for the existence of correlation

Hypotheses: H_0: $\rho=0$ (no correlation)
H_1: $\rho\neq0$ (there is correlation)

Test statistic t: $t = \dfrac{R - r}{\sqrt{\dfrac{1-R^2}{n-2}}}$

follows a t distribution with n-2 degrees of freedom.

12

© 2001. D. A. Menascé. All Rights Reserved.