Simple Regression

Prof. Daniel A. Menascé

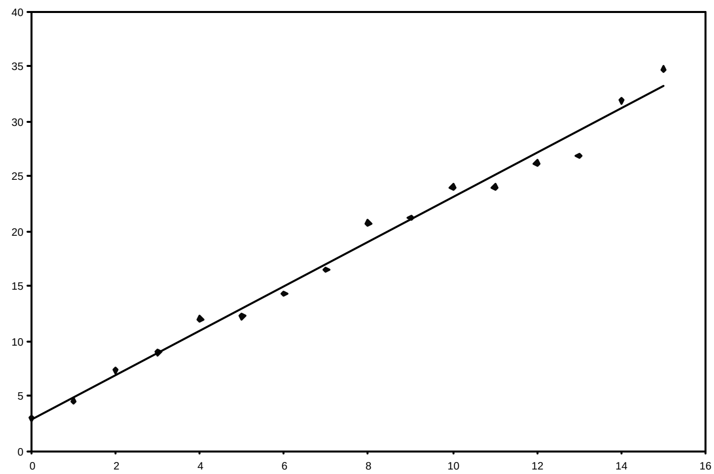
Dept. of Computer Science

George Mason University

Basics

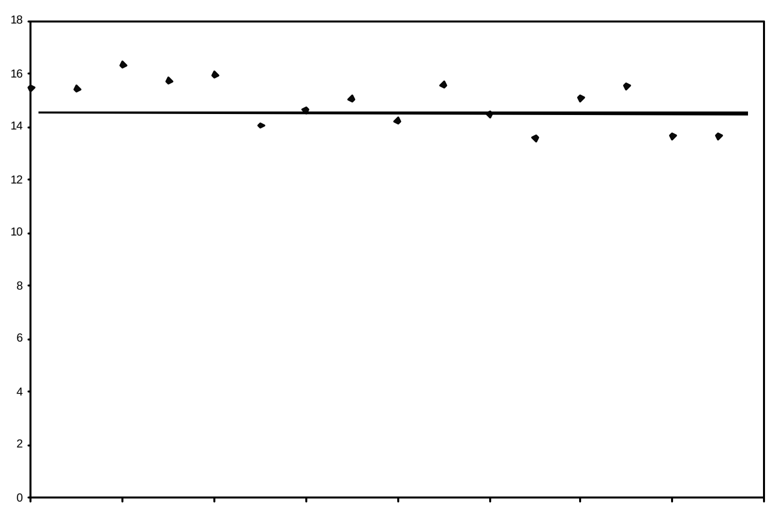
- Purpose of <u>regression analysis</u>: predict the value of a dependent or response variable from the values of at least one explanatory or independent variable (also called predictors or factors).
- Purpose of <u>correlation analysis</u>: measure the strength of the correlation between two variables.

Linear Relationship



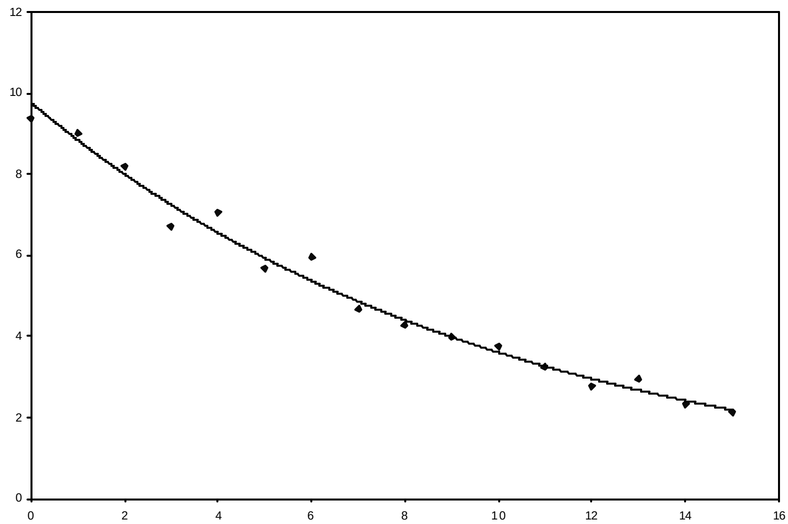
3

No Relationship

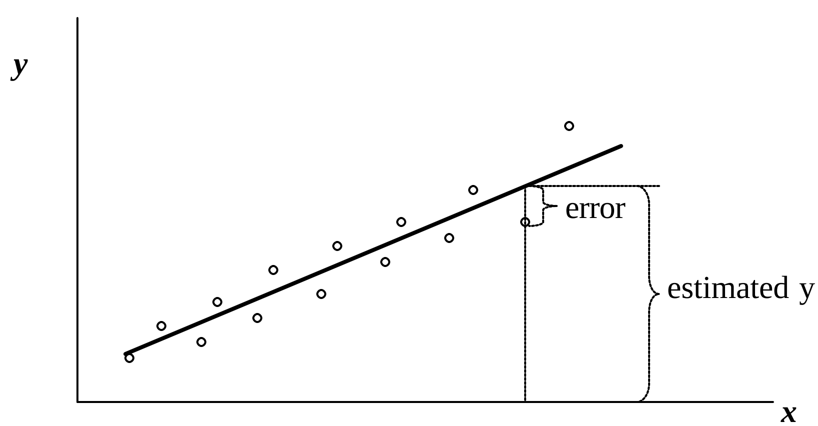


4

Negative Curvilinear

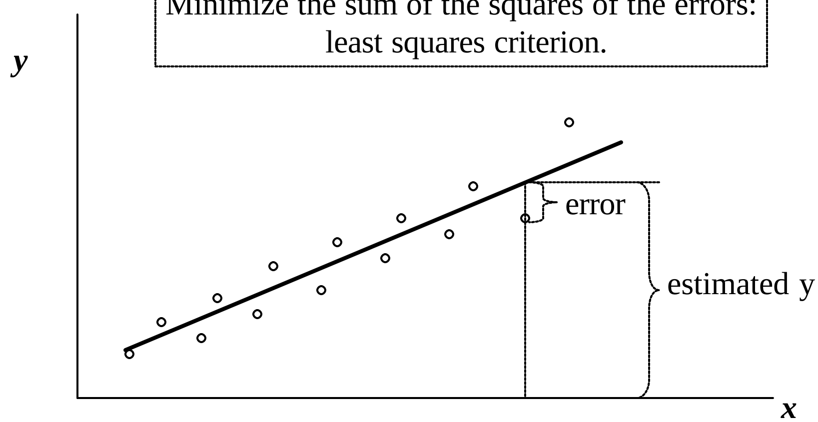


Simple Linear Regression Residual Error



Simple Linear Regression Selecting the "best" line

Minimize the sum of the squares of the errors: least squares criterion.



7

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:

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

Subject to: $\sum_{i=1}^{n} e_i = 0$

8

Method of Least Squares

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\left(\overline{X}\right)^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

9

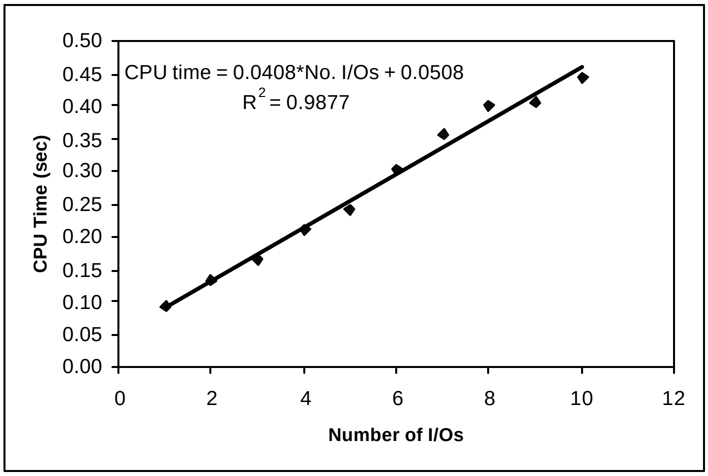
Linear Regression Example

Number of I/Os (x)	CPU Time (y)	Estimate (0.0408*x +0.0508)	Error	Error Squared
1	0.092	0.092	0.0005	0.00000
2	0.134	0.132	0.0013	0.00000
3	0.165	0.173	-0.0083	0.00007
4	0.211	0.214	-0.0026	0.00001
5	0.242	0.255	-0.0128	0.00016
6	0.302	0.295	0.0067	0.00005
7	0.357	0.336	0.0206	0.00042
8	0.401	0.377	0.0239	0.00057
9	0.405	0.418	-0.0131	0.00017
10	0.442	0.459	-0.0161	0.00026
				0.00171

Xbar	5.5
Ybar	0.275
Sum x2	385
Sum xy	18.494616
b1	0.0408
b0	0.0508

10

Linear Regression Example



Allocation of Variation

- No regression model: use mean as predicted value. SSE is:

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$ Sum of squares total

$$SSR = SST - SSE$$ Sum of squares explained by the regression.

Variation not explained by regression

Allocation of Variation

- Coefficient of determination (R^2): fraction of variation explained by the regression.

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

The closer R^2 is to one, the better is the regression model.

13

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

SSE SSY

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

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

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

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

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

14

Standard Deviation of Errors

- Variance of errors: divide the sum of squares (SSE) by the number of degrees of freedom (n-2 since two regression parameters need to be computed first).

$$s_e^2 = \frac{SSE}{n-2}$$ —— Mean squared error (MSE)

15

Degrees of freedom of various sum of squares.

SST	n-1	Need to compute $\bar{Y}$
SSY	n	Does not depend on any other parameter
SS0	1	Can be computed from $\bar{Y}$
SSE	n-2	Need to compute two regression parameters
SSR	1	=SST-SSE

Degrees of freedom add as sum of squares do.

16

Confidence Interval for Regression Parameters

- b_o and b_1 were computed from a sample. So, they are just estimates of the true parameters β_0 and β_1 for the true model.
- Standard deviations for b_o and b_1.

$$s_{b_0} = s_e \sqrt{\frac{1}{n} + \frac{(\overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2}}$$

$$s_{b_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2}}$$

17

Confidence Interval for Regression Parameters

100(1-α)% confidence interval for b_o and b_1

$$b_0 \pm t_{[1-a/2;n-2]} s_{b_0}$$
$$b_1 \pm t_{[1-a/2;n-2]} s_{b_1}$$

18

Confidence Interval Example

Number of I/Os (x)	CPU Time (y)	Estimate (0.0408*x +0.0508)	Error	Error Squared
1	0.092	0.092	0.0005	0.00000
2	0.134	0.132	0.0013	0.00000
3	0.165	0.173	-0.0083	0.00007
4	0.211	0.214	-0.0026	0.00001
5	0.242	0.255	-0.0128	0.00016
6	0.302	0.295	0.0067	0.00005
7	0.357	0.336	0.0206	0.00042
8	0.401	0.377	0.0239	0.00057
9	0.405	0.418	-0.0131	0.00017
10	0.442	0.459	-0.0161	0.00026
			SSE:	0.00171

Xbar	5.5		
Ybar	0.275		
Sum x2	385		
Sum xy	18.494616		
b1	0.0408		
b0	0.0508		
se^2	0.0002144	Lower bo	0.027772
se	0.0146411	Upper bo	0.073900
sb0	0.0100017		
sb1	0.0016119	Lower b1	0.037058576
95% confidence level		Upper b1	0.044492804
alpha	0.05		
t[1-alpha/2;n-2]	2.3060056		
SST	0.1388841		
SSR	0.13717		
R2	0.9876524		

19

Confidence Interval for the Predicted Value

- The standard deviation of the mean of a future sample of *m* observations at X = X_p is

$$s_{\hat{y}_{mp}} = s_e \left[\frac{1}{m} + \frac{1}{n} + \frac{\left(X_p - \overline{X}\right)^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \right]^{1/2}$$

As the future sample size (*m*) decreases, the standard deviation for predicted value increases.

20

Confidence Interval for the Predicted Value

100(1-α)% confidence interval for the predicted value for a future sample of size *m* at X_p:

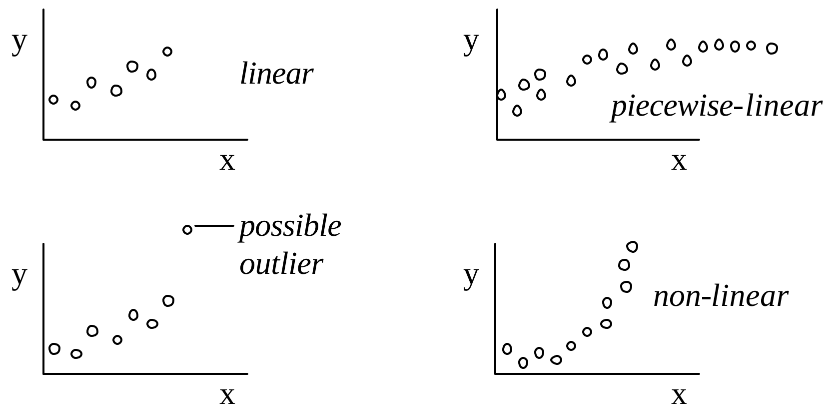
$$\hat{y}_p \pm t_{[1-a/2;n-2]} s_{\hat{y}_{mp}}$$

Linear Regression Assumptions

- Linear relationship between the response (y) and the predictor (x).
- The predictor (x) is non-stochastic and is measured without any error.
- Errors are statistically independent.
- Errors are normally distributed with zero mean and a constant standard deviation.

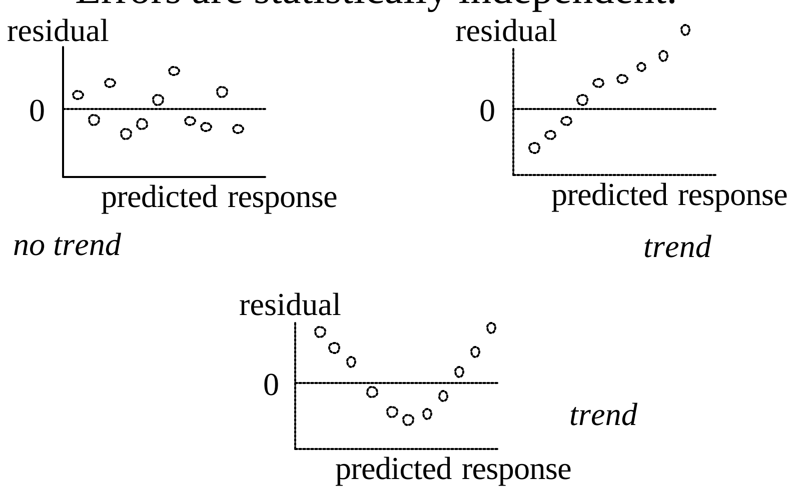
Linear Regression Assumptions

- Linear relationship between the response (y) and the predictor (x).



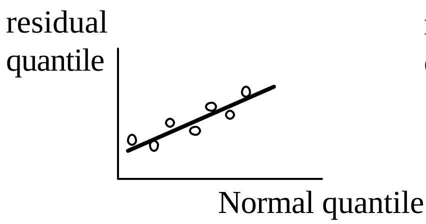
Linear Regression Assumptions

- Errors are statistically independent.

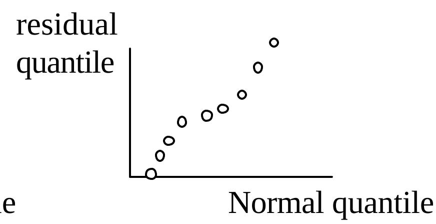


Linear Regression Assumptions

- Errors are normally distributed.



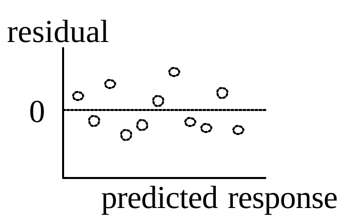
normally distributed errors



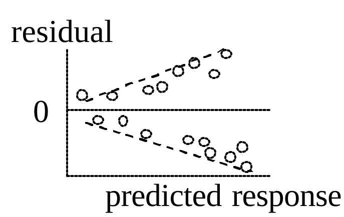
non-normally distributed errors

Linear Regression Assumptions

- Errors have a constant standard deviation.



no trend in spread



increasing spread