

Simple Linear Regression

Setting:

Collect random sample of pairs (x_i, y_i) of size n

Questions:

- Is there a linear relationship between X and Y?
- If a value for X is known ($X=x^*$) what Y value do you predict?
- How strong is the relationship between X and Y?

Mathematical Model:

$$y_i = \mathbf{b}_0 + \mathbf{b}_1 \cdot x_i + \tilde{\mathbf{e}}$$

Two numbers summarize the relationship:

\mathbf{b}_0 : the intercept (a random variable) and

\mathbf{b}_1 : the slope coefficient (also a random variable)

The above random variables follow a known distribution
Student's-t (with $n-2$ Degrees of Freedom)

Assumptions:

\tilde{e} : The error terms are Independent and Identically Normally distributed

Therefore observed residual values are:

- 1) independent, thus
 - a. no serial correlation
 - b. no pattern in the residual plot

- 2) their distribution is normal with mean 0 and standard deviation that is constant

Hypothesis Testing for regression parameters

$H_0: \text{Parameter} = 0$ vs $H_a: \text{Parameter} \neq 0$

P-value is given by computer software

Also confidence interval is given by computer software

Prediction

For any X value two Confidence Intervals (CIs) are given:

- CI for mean response at X value (i.e. take 10 pairs for fixed X and find their average Y, it will lie within the CI for mean response)

- CI for individual observation at X value (i.e. take 1 (X,Y) pair at fixed X value, it will lie within this CI – also known as Prediction Interval (PI))

PI is wider than the CI for mean response

ANOVA TABLE

DFR = 1	SSR	MSR	F	P-value
DFE = n-2	SSE	MSE		
DFT = n-1	SST			

DF: Degrees of Freedom

SS: Sum of Squares

MS: Mean Square (= SS / DF)

F: Value of F-statistic (= MSR / MSE)

P-value: corresponds to observed F-statistic

R: Regression

E: Error

T: Total

General Rule

Total = Regression + Error

MSE is the variance of residuals around the regression line

Thus: Standard Error = SQRT (MSE)

Coefficient of Determination

(measures strength of relationship between X and Y)

$$\text{R-Square} = \text{SSR} / \text{SST}$$

Correlation Coefficient r

$$r = \text{sqrt}(\text{R-square}) * \text{sign}(\beta_1)$$

Calculation of t-stat(β) for a parameter β

$$\text{t-stat}(\beta) = \text{coef}(\beta) / \text{st.error}(\beta)$$

Visual Specification Tests

Check Residuals Follow the Normal distribution

Examine: Normal Probability Plot of Residuals
Histogram of Residuals
Box-Plot of Residuals

Check Residuals are Identically Distributed

Examine: Residual Plot against X values
Residual Plot against Y-predicted values
Residual Plot against Y values

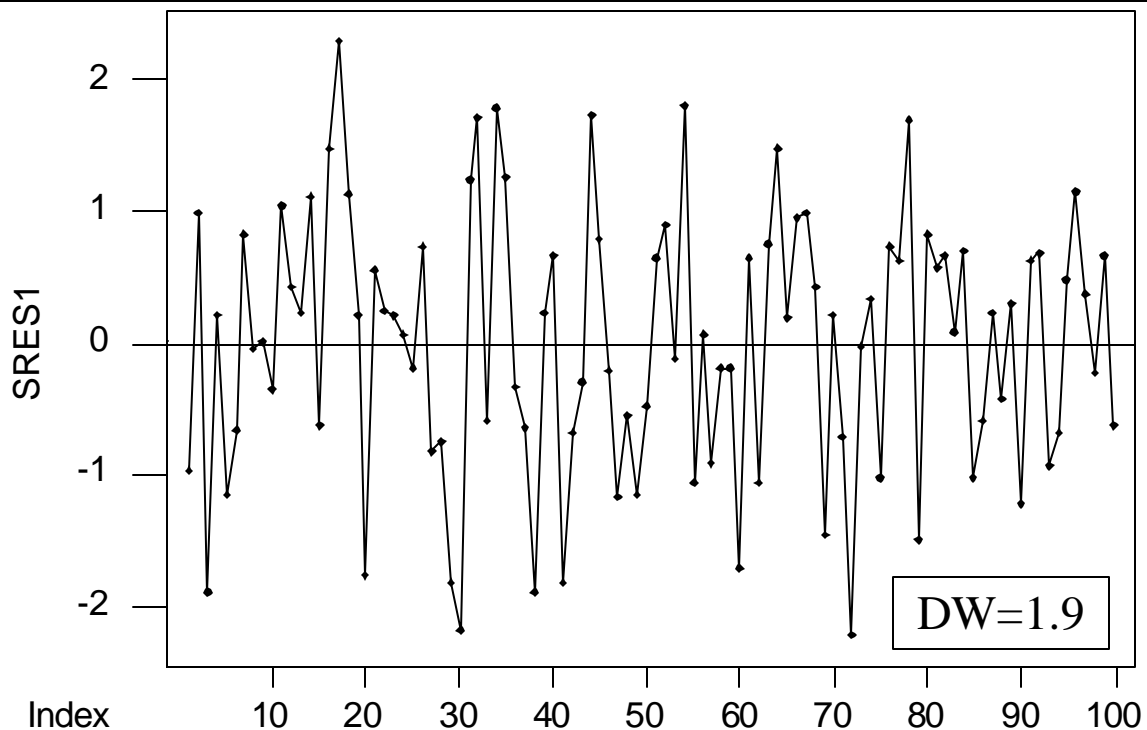
For Data in Time Series

Check Residuals are Independent to Each Other

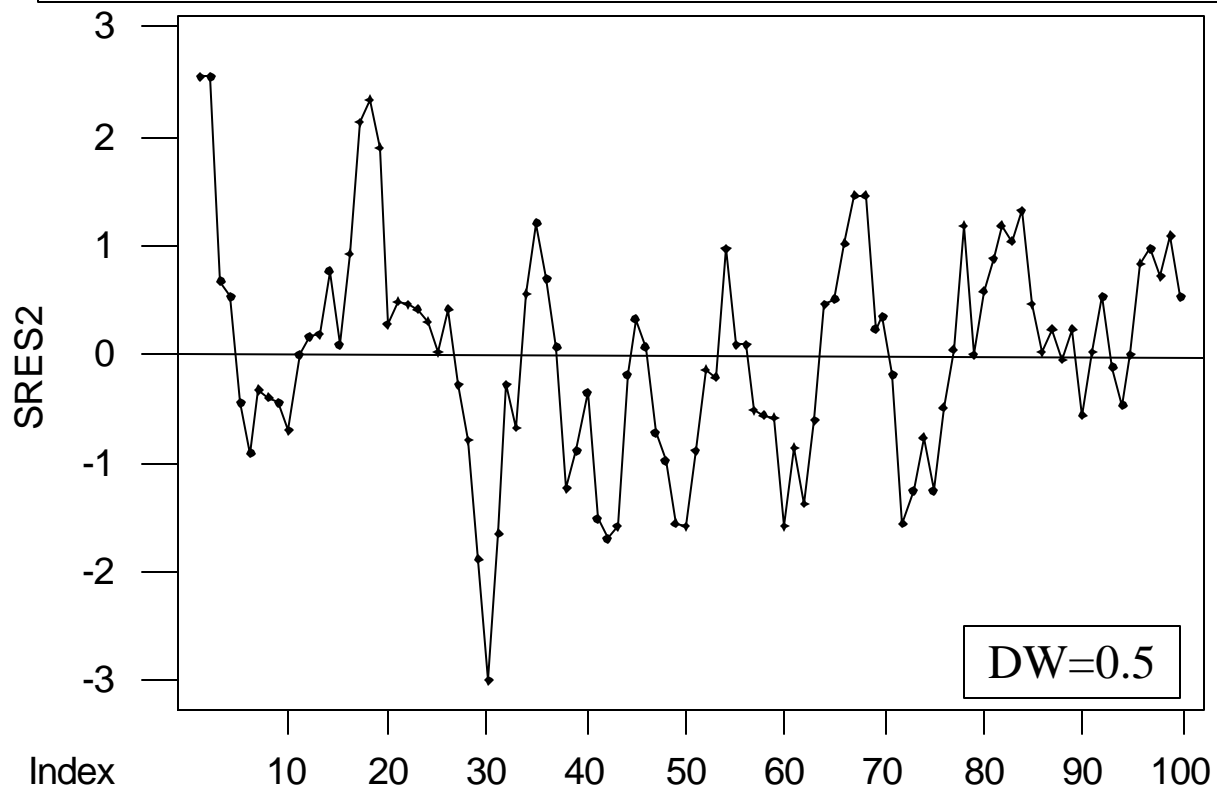
Examine: Residual Plot against time
(order of data collection)

Plot usually hard to read – check also DW statistic

Acceptable Residual Plot (Independent Residuals)



Autocorrelated Residuals (Serial Correlation)



Durbin-Watson Test

Obtain Durbin-Watson Statistic (PHStat, MINITAB)

Measures autocorrelation between successive residuals

Ranges between 0 – 4

0	d_L	d_U			2		4
---	-------	-------	--	--	---	--	---

If D-W is low there is positive autocorrelation in the residuals

n	d_L
15	1.08
30	1.35
60	1.55
100	1.65

If D-W is high there is negative autocorrelation in the residuals

Find d_L , d_U in Table E.11

MINITAB REGRESSION OUTPUT EXAMPLE

Regression Analysis: y versus x

The regression equation is
 $y = 1.23 + 0.202 x$

Predictor	Coef	SE Coef	T	P
Constant	1.2324	0.2860	4.31	0.004
x	0.20221	0.01145	17.66	0.000

S = 0.4345 R-Sq = 97.8% R-Sq(adj) = 97.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	58.907	58.907	311.97	0.000
Residual Error	7	1.322	0.189		
Total	8	60.229			

Unusual Observations

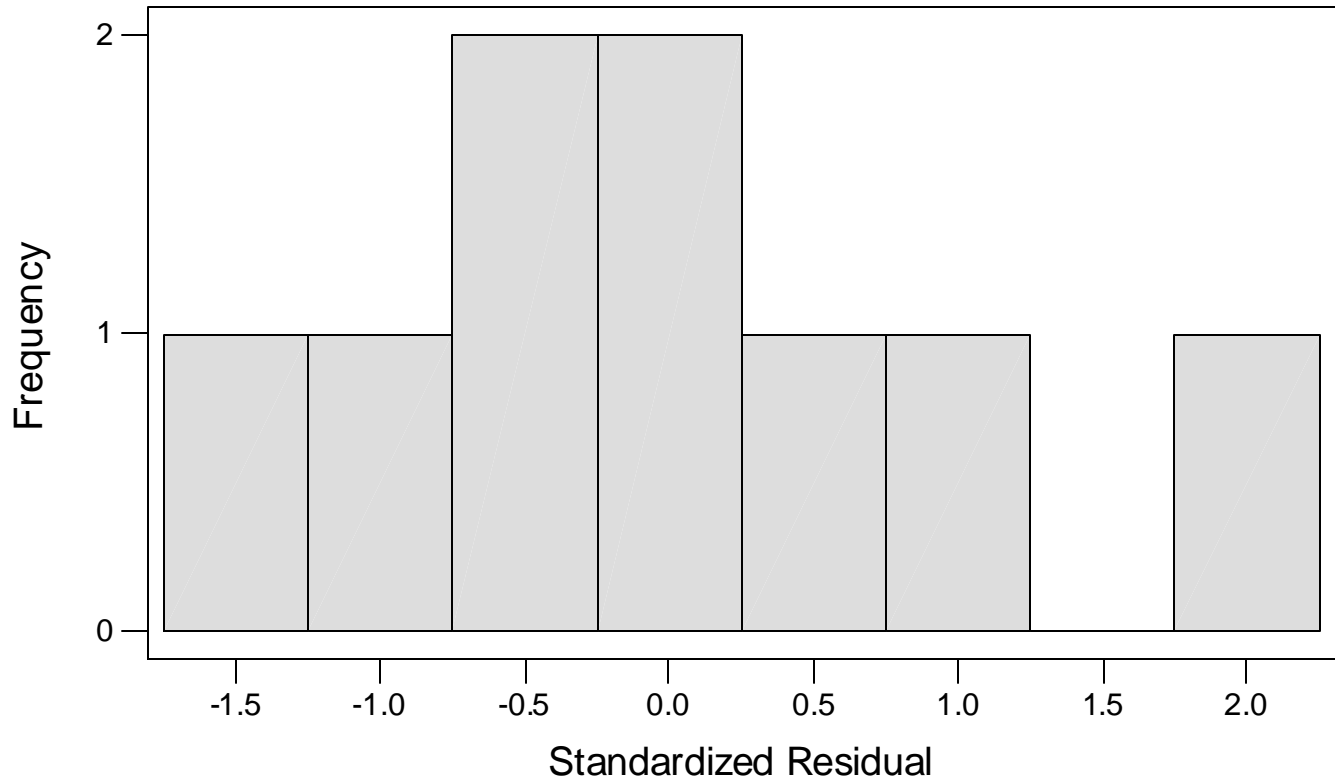
Obs	x	y	Fit	SE Fit	Residual	St Resid
1	15.6	5.200	4.387	0.160	0.813	2.01R

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 2.15

Histogram of the Residuals

(response is y)



Residuals Versus the Fitted Values

(response is y)

