POLITECNICO
MILANO 1863

Quality Data Analysis

Linear models - part 1

Bianca Maria Colosimo – biancamaria.colosimo@polimi.it

Sources: * chapter 3 Alwan + chapter 10 "Applied Statistics and Probability for Engineers" –D.C. Montgomery and G.C. Runger - John Wiley & Sons 2nd edition

How can we decide whether the standard model is appropriate?

Assumptions	Hypothesis test (to check the assumption)	Remedy in case of violation
"independence" (random pattern)	- Runs test - Bartlett's test - LBQ's test	**-gapping** **-batching** **-(Linear) regression** -Time series (ARIMA)
Normal distribution	Normality test	Transform data

Gapping

A first (trivial) method to deal with autocorrelated data consists of reducing the sampling frequency

– With reference to previous data – let's take one data out of 10.

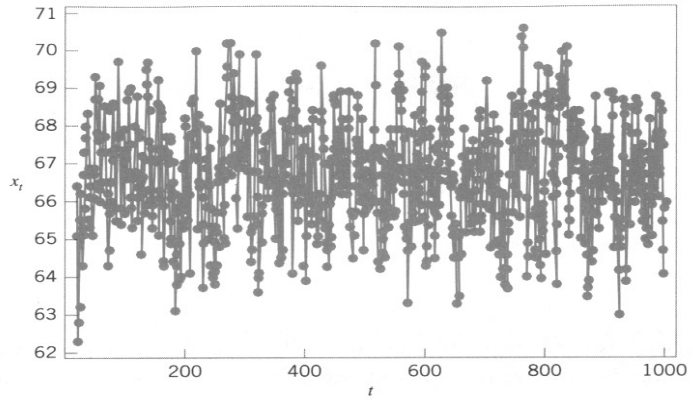


Figure 9-6 A process variable with autocorrelation.

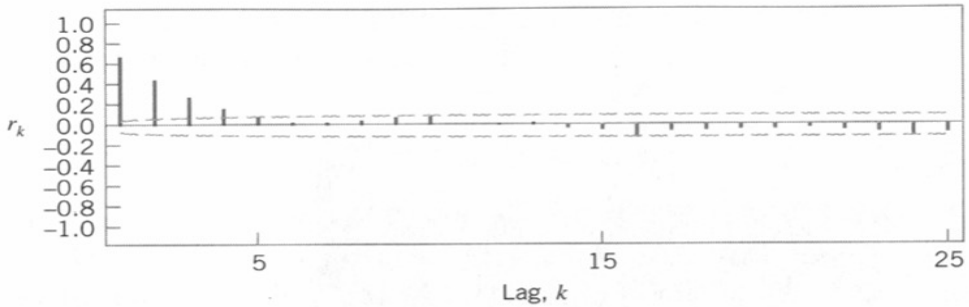
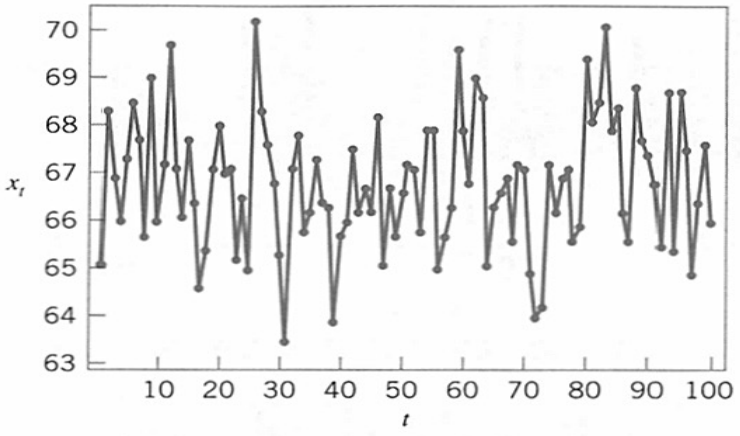
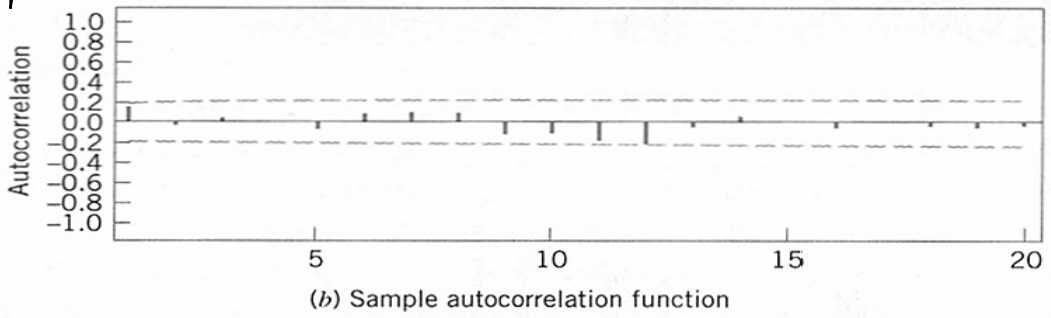


Figure 9-8 Sample autocorrelation function for the data in Fig. 9-6.



(*a*) Data



(*b*) Sample autocorrelation function

Figure 9-10 Plots for every 10th observation from Fig. 9-6.

Batching

A second approach to "remove" the autocorrelation is batching *:

- This is a model-free approach.
- The sequence of data is organized in sequential batches (not overlapped) and it is considered the average of values in each batch.

$$\bar{x}_j = \frac{1}{b}\sum_{i=1}^{b} x_{(j-1)b+i} \qquad j = 1,2,...$$

*in control charting: Unweighted Batch Mean (UBM Control chart 1996)

Ex:
1000 data from a chemical process

$$b = 10$$

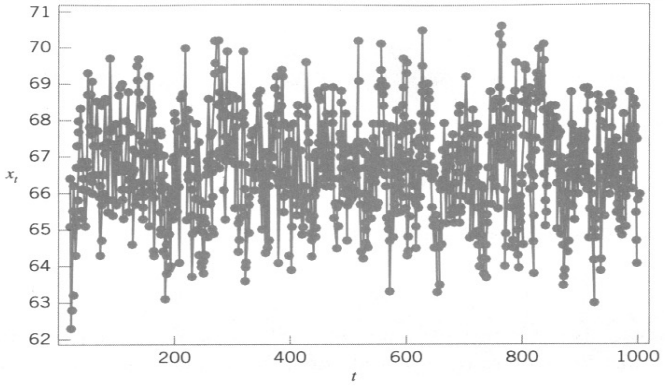


Figure 9-6 A process variable with autocorrelation.

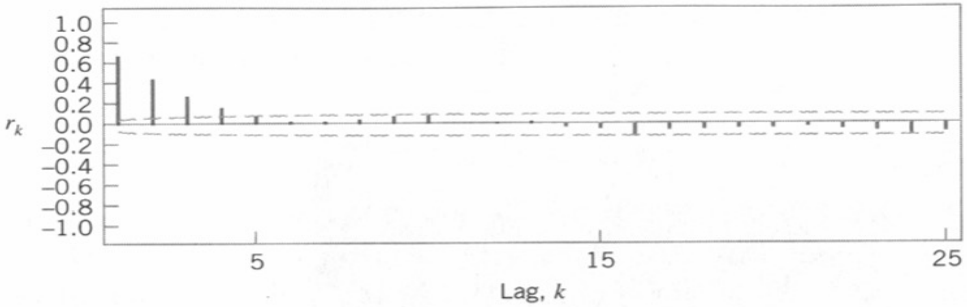
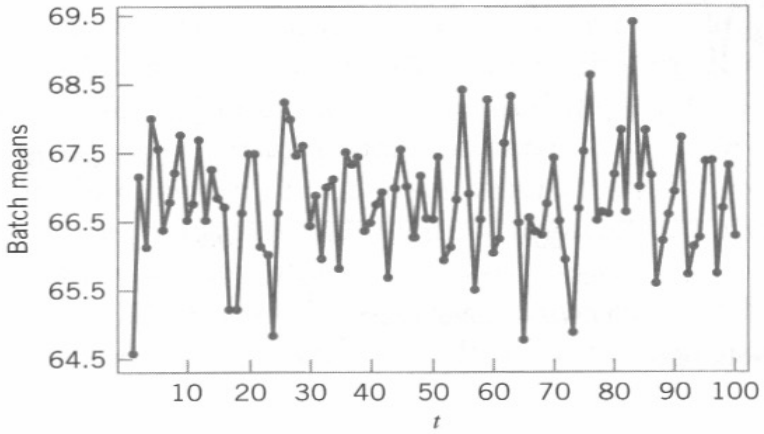
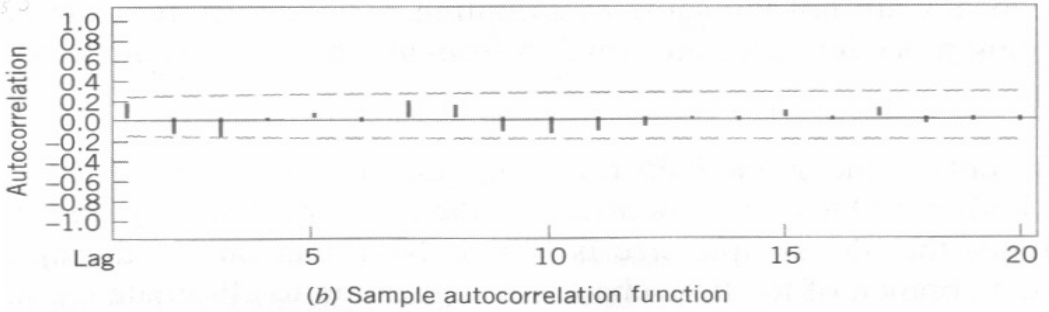


Figure 9-8 Sample autocorrelation function for the data in Fig. 9-6.



(*a*) Plot of batch means using batch size $b = 10$



(*b*) Sample autocorrelation function

Batching

The main disadvantage of batching is the difficulty to define the appropriate value of *b* (batch size).

Some empirical approaches have been proposed.

Ex:

1. Initialize *b=1*
2. Compute the autocorrelation coefficient at the first lag
3. If the coefficient is smaller than 0.1 go to step 5
4. Set *b=2*b*, go to step 2
5. end

Gapping/Batching

1: Both the approaches are applicable to stationary processes (constant mean)

2: Both the approaches induce loss of information

These are approaches which do not tackle the autocorrelation issue instead of dealing with it.
How can we “identify” the appropriate model in case of nonrandom data?

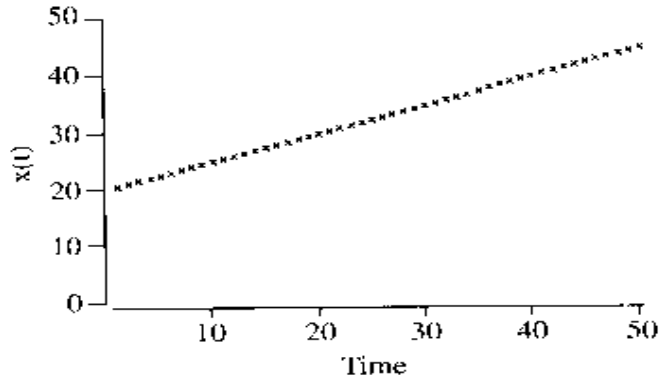
-Regression
-ARIMA

Process model

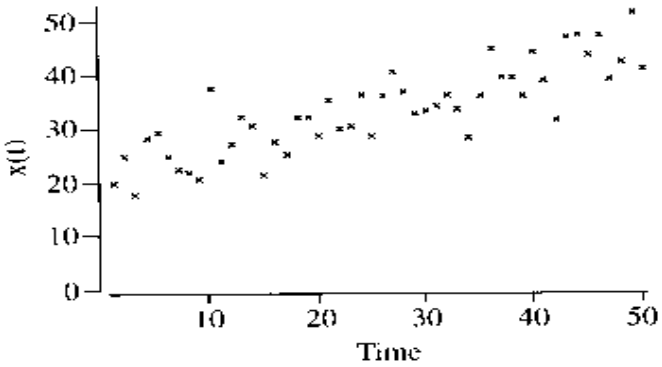
Basic model: $Y_t = \mu_t + \varepsilon_t$

with $-\mu_t = \mu$ ~~(cost)~~ $-\varepsilon_t \sim$ ~~NID~~$(0, \sigma_\varepsilon^2)$

Effect of ε_t



(a) "Ideal" data set revealing underlying true model.



(b) Real data set reflecting underlying true model and random errors.

Figure 3.17 Effects of random errors.

Linear regression

We can look for the model/coefficients to minimize the Sum of Squared Errors (Minimum Mean Squared Error – MSE – approach)

$$Y_t = \mu_t + \varepsilon_t$$

From this model, we take new data y_t $t = 1,\ldots,n$

$$SSE = \sum_{t=1}^{n} (y_t - \mu_t)^2 = \sum_{t=1}^{n} \varepsilon_t{}^2$$

Observed data

Deterministic (assumed known)

Linear regression

In order to identify the model:

1. we have to assume the model "structure"

$$\text{ex}: \mu_t = \text{cost}, \qquad \mu_t = \text{a}t + b,$$

$$\mu_t = \text{a}t^2 + bt + c, \qquad \mu_t = \text{a}\ln t,$$

$$(\sigma_\varepsilon^2 = \text{cost})$$

For the sake of simplicity che can assume models linear with reference to the unknown parameters: LINEAR REGRESSION

$$\text{ex: non linear}: \mu_t = at + bt^c$$

$$\text{if it were } \mu_t = bt^c \rightarrow \ln \mu_t = \ln b + c \ln t \text{ (linear)}$$

2. We have to estimate the unknown parameters by minimizing SSE

$$SSE = \sum_{t=1}^{n} (y_t - \mu_t)^2 = \sum_{t=1}^{n} \varepsilon_t^2$$

Observed data — Unknow mean (deterministic)

$$ex: \hat{\mu}_t = \hat{\text{c}}\text{ost}, \quad \hat{\mu}_t = \hat{a}t + \hat{b}, \quad \hat{\mu}_t = \hat{a}t^2 + \hat{b}t + \hat{c}, \quad \hat{\mu}_t = \hat{a}\ln t$$

Alwan: $Y_t = \mu_t + \varepsilon_t \longrightarrow \hat{y}_t = \hat{\mu}_t$ Estimate of the deterministic part of the model

It is a general model, where $\mu_t = \mu$ (cost) is a special case:

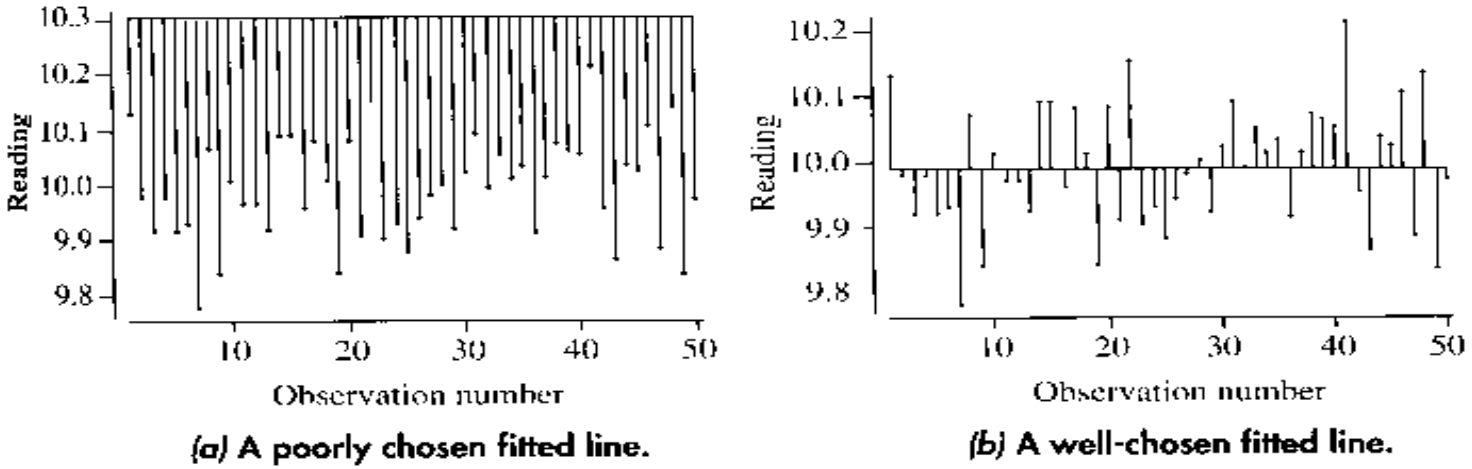
Assume the "true" model is $\mu_t = \beta_0$

- From this model we observe data $y_t \qquad t = 1,\ldots,n$

Assumed model: $Y_t = \mu + \varepsilon_t = \beta_0 + \varepsilon_t$

Let find $\hat{\beta}_0 = b_0$ so to minimize *SSE*:

$$SSE = \sum_{t=1}^{n} (y_t - \mu_t)^2 = \sum_{t=1}^{n} \varepsilon_t{}^2$$



(a) **A poorly chosen fitted line.** *(b)* **A well-chosen fitted line.**

Figure 3.18 Different fitted lines for the stopwatch data series.

observed

assumed

$$SSE = \sum_{t=1}^{n} (y_t - \beta_0)^2$$

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{t=1}^{n} (y_t - b_0) = 0 \qquad \sum_{t=1}^{n} y_t - n b_0 = 0$$

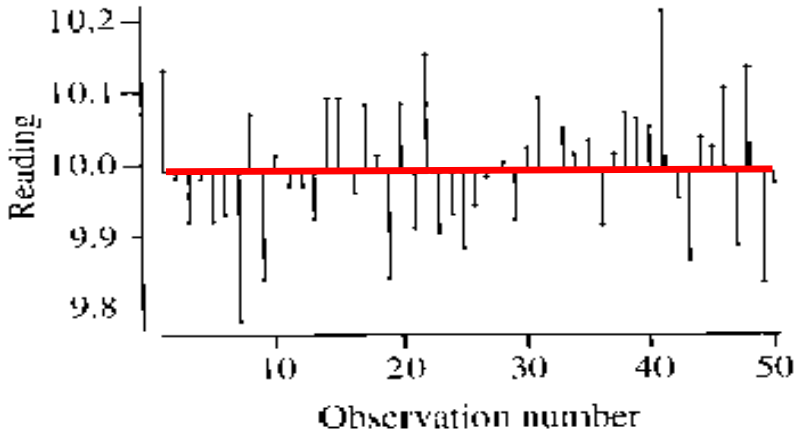
$$\Rightarrow b_0 = \frac{1}{n} \sum_{t=1}^{n} y_t = \bar{y}$$

$$SSE(b_0) = SSE = \sum_{t=1}^{n} (y_t - \hat{y})^2 = \sum_{t=1}^{n} (y_t - b_0)^2$$

Difference:

S(β_0) function to be minimized

$$SSE = \min_{\beta_0} S(\beta_0)$$



$\hat{y} = \bar{y}$

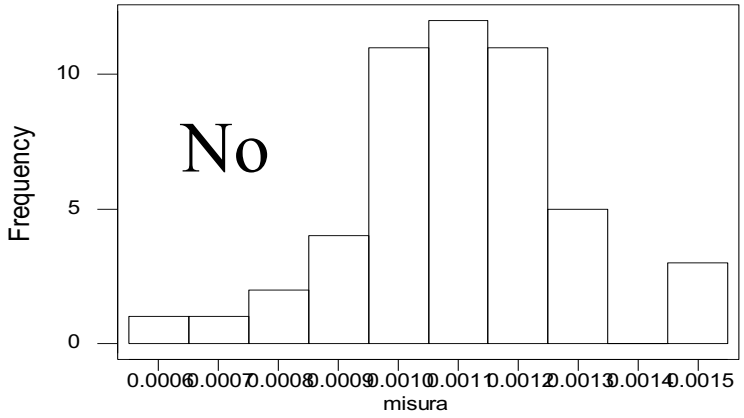
(b) **A well-chosen fitted line.**

TREND

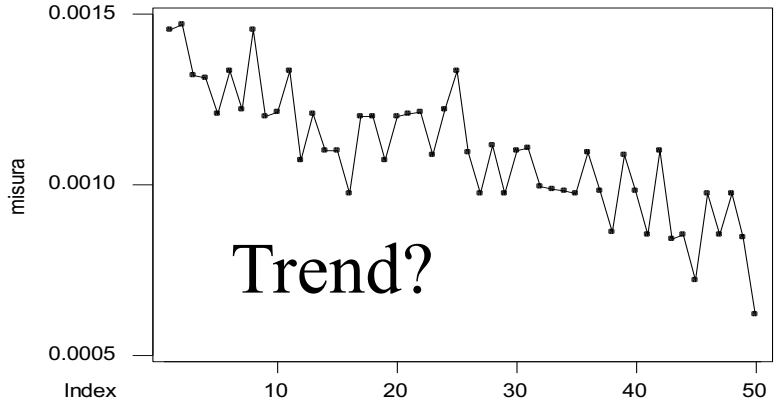
50 sequentially produced items: elongation of a spring subject to a force of 20 g

(Deming 1986: deming.dat)

t	misura	t	misura	t	misura	t	misura	t	misura
1	0.001454	11	0.001334	21	0.001207	31	0.001108	41	0.000854
2	0.001468	12	0.001073	22	0.001214	32	0.000995	42	0.001101
3	0.00132	13	0.001207	23	0.001087	33	0.000988	43	0.00084
4	0.001313	14	0.001101	24	0.001221	34	0.000981	44	0.000854
5	0.001207	15	0.001101	25	0.001334	35	0.000974	45	0.00072
6	0.001334	16	0.000974	26	0.001094	36	0.001094	46	0.000974
7	0.001221	17	0.0012	27	0.000974	37	0.000981	47	0.000854
8	0.001454	18	0.0012	28	0.001115	38	0.000861	48	0.000974
9	0.0012	19	0.001073	29	0.000974	39	0.001087	49	0.000847
10	0.001214	20	0.0012	30	0.001101	40	0.000981	50	0.000621



ering



Trend? "True" model

- In the example :
- Observations:

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad (1)$$

$x_t = t$ regressor or predictor

$y_t \qquad t = 1, \ldots, n$

Estimated model

$$\hat{y}_t = b_0 + b_1 x_t$$

where $b_0 = \hat{\beta}_0 \quad b_1 = \hat{\beta}_1$

Find b_0, b_1 to minimize $SSE = \sum_{t=1}^{n} \left(y_t - \beta_0 - \beta_1 x_t \right)^2$

$$SSE = \sum_{t=1}^{n} \left(y_t - \beta_0 - \beta_1 x_t \right)^2 \rightarrow \begin{cases} \dfrac{\partial SSE}{\partial \beta_0} = 0 \\ \dfrac{\partial SSE}{\partial \beta_1} = 0 \end{cases} \Rightarrow \begin{array}{l} \hat{\beta}_0 = b_0 \\ \hat{\beta}_1 = b_1 \end{array}$$

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{t=1}^{n} \left(y_t - b_0 - b_1 x_t \right) = 0 \quad n\bar{y} - nb_0 - nb_1 \bar{x} = 0 \Rightarrow b_0 = \bar{y} - b_1 \bar{x}$$

$$SSE = \sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 x_t)^2 \qquad b_0 = \bar{y} - b_1 \bar{x}$$

$$\frac{\partial SSE}{\partial \beta_1} = 0 \qquad \frac{\partial SSE}{\partial \beta_1} = -2 \sum_{t=1}^{n} x_t (y_t - b_0 - b_1 x_t) = -2 \sum_{t=1}^{n} \left(x_t y_t - x_t \bar{y} + b_1 x_t \bar{x} - b_1 x_t^2 \right) = 0$$

observe that

$$\sum_{t=1}^{n} (x_t - \bar{x}) = \sum_{t=1}^{n} (y_t - \bar{y}) = 0$$

$$\sum_{t=1}^{n} (x_t y_t - x_t \bar{y}) + b_1 \sum_{t=1}^{n} \left(x_t \bar{x} - x_t^2 \right) = 0$$

$$\sum_{t=1}^{n} (x_t y_t - x_t \bar{y}) - \bar{x} \sum_{t=1}^{n} (y_t - \bar{y}) + b_1 \sum_{t=1}^{n} \left(x_t \bar{x} - x_t^2 \right) + b_1 \bar{x} \sum_{t=1}^{n} (x_t - \bar{x}) = 0$$

$$\sum_{t=1}^{n} (x_t y_t - x_t \bar{y} - \bar{x} y_t + \bar{x} \bar{y}) + b_1 \sum_{t=1}^{n} \left(x_t \bar{x} - x_t^2 + \bar{x} x_t - \bar{x}^2 \right) = 0$$

$$\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y}) - b_1 \sum_{t=1}^{n} (x_t - \bar{x})^2 = 0$$

$$\Rightarrow b_1 = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n} (x_t - \bar{x})^2}$$

prop

Cont.

Therefore $Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \Rightarrow \hat{y}_t = b_0 + b_1 x_t$

$$\min SSE \Rightarrow \begin{cases} b_1 = \dfrac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \\ \\ b_0 = \bar{y} - b_1 \bar{x} \end{cases}$$

define $S_{xx} = \sum_{t=1}^{n}(x_t - \bar{x})^2$

$S_{xy} = \sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})$

$$\Rightarrow b_1 = \frac{S_{xy}}{S_{xx}}$$

Note that

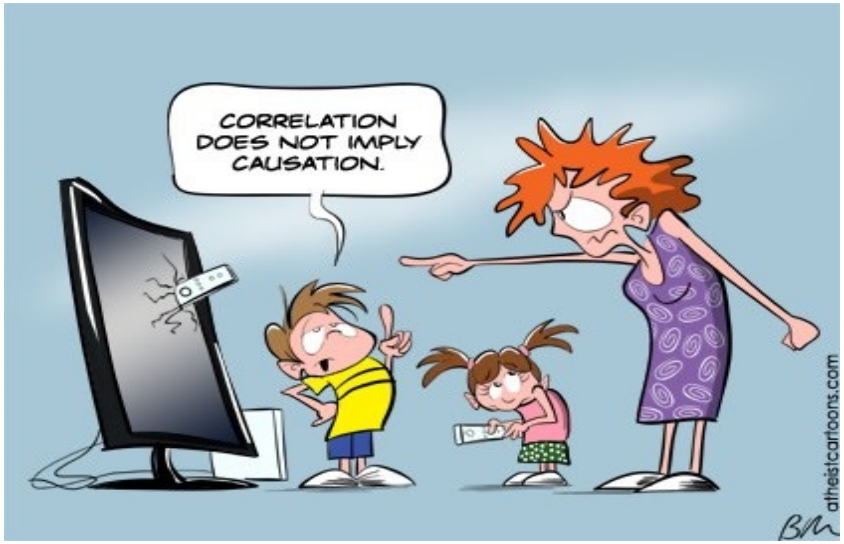
- Simple linear regression: first order model
  It is function of one single regressor (linear function of $x_t=t$)
- Pay attention:

β_0, β_1 deterministic values (unknown)

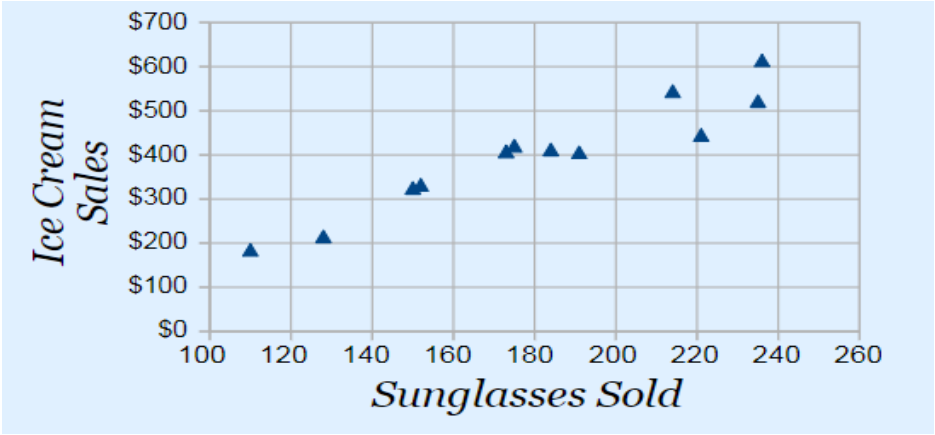
$b_0 = \hat{\beta}_0$, $b_1 = \hat{\beta}_1$ random variable (they are function of observed data)

- If the true model is equal to the assumed one:
  - Unbiased estimators: $E(b_0) = \beta_0 \quad E(b_1) = \beta_1$
  - Min variance estimators (among all the unbiased estimators)
- Pay attention to common errors in regression:
  - Strong relationship among variables does not mean casual relationship
  - The identified relationship is valid only in the explored interval of *x* pay attention to *extrapolation*

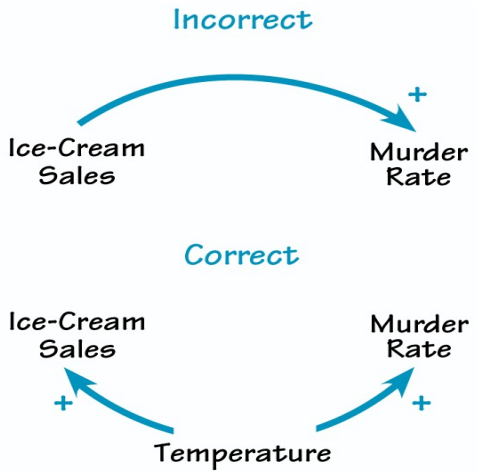
Correlation does not cause causation



Ice cream sales is correlated with homicides in New York (Study) As the sales of ice cream rise and fall, so do the number of homicides.



Does the consumption of ice cream causing the death of the people?



https://towardsdatascience.com/why-correlation-does-not-imply-causation-5b99790df07e

Multiple linear regression

When we have more than one regressor: multiple linear regression

A matrix-form notation is used:

$$\begin{bmatrix} y_1 \\ \\ y_i \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_i \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\Rightarrow \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

nx1 nxK Kx1 nx1

K=k+1

$$\Rightarrow \hat{\beta} = (X'X)^{-1}X'y$$

$(X'X)^{-1}$ exists if the regressors are linearly independent

(no column of X is linear combination of the other columns)

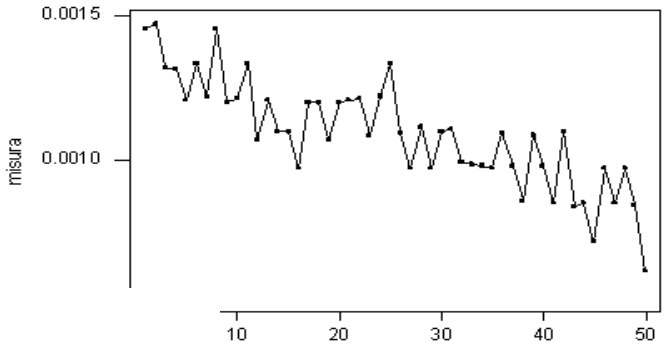
$$\mathrm{E}(\hat{\beta}) = \beta \qquad \mathrm{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

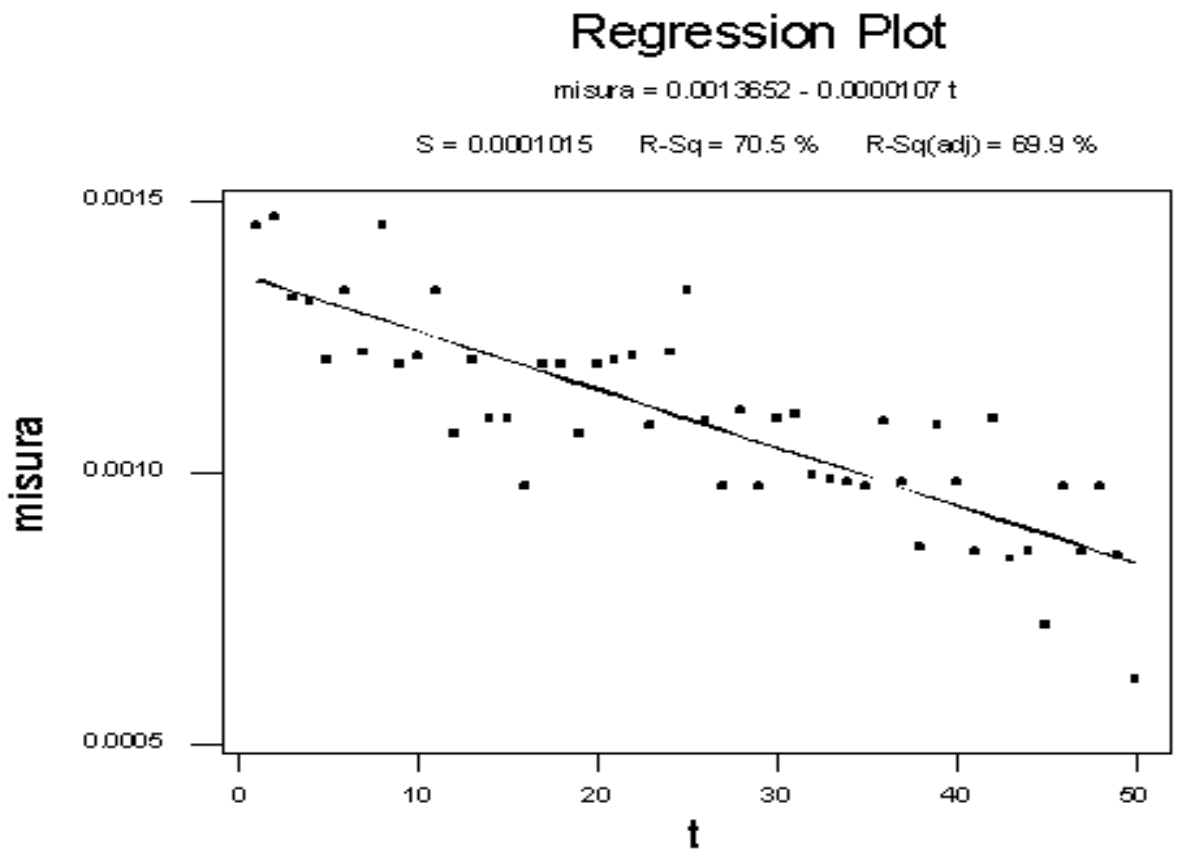
orthogonal columns: $X_2'X_1 = 0$

$$\Rightarrow \hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$$

$$\varepsilon \sim MN(0, \sigma^2 I) \quad \Rightarrow \quad \hat{\beta} \sim MN(\beta, (X^T X)^{-1}\sigma^2)$$

deming.dat





$$\hat{y}_t = b_0 + b_1 x_t$$

misura = 0.0013652 - 0.0000107 t

Trend: regression output

Regression Analysis: misura versus t

Regression equation

Hypothesis test

```
The regression equation is
misura = 0.00137 -0.000011 t

Predictor         Coef     SE Coef           T         P
Constant   0.00136522  0.00002914       46.84     0.000
t         -0.00001066  0.00000099      -10.72     0.000

S = 0.0001015   R-Sq = 70.5%   R-Sq(adj) = 69.9%
```

Coefficient of determination R^2

Estimate of σ_ε

```
Analysis of Variance
Source              DF           SS           MS          F       P
Regression           1 1.18428E-06 1.18428E-06     114.96   0.000
Residual Error      48 4.94479E-07 1.03017E-08
Total               49 1.67876E-06
```

Test of hypothesis

```
misura = 0.00137 -0.000011 t
```

≅0: is this a real trend? (slope is so close to 0)

b_0 e b_1 random variables:

- TEST 1

$$H_0 : \beta_i = 0 \qquad H_1 : \beta_i \neq 0 \qquad \text{for a given } i$$

- TEST 2:

$$H_0 : \quad \beta_1 = \beta_2 = ... = \beta_{K-1} = 0$$
$$H_1 : \quad \beta_j \neq 0 \quad \text{for at least one j}$$

K=number of regressors (p)+1

Pay attention:
For these tests we need to assume normality of the ε_t's

$$\varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2)$$

TEST 1

$$H_0 : \beta_i = 0 \qquad H_1 : \beta_i \neq 0$$

$$\varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \Rightarrow Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \sim NID(\beta_0 + \beta_1 x_t, \sigma_\varepsilon^2)$$

$$b_1 = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \quad \text{e} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Are linear combination of data y_t's

$$\Rightarrow b_i \ (i = 0,1) \ \text{normally distributed}$$

We can show that:

$$E(\hat{\beta}_1) = E(b_1) = \beta_1 \qquad V(b_1) = \frac{\sigma_\varepsilon^2}{S_{xx}}$$

$$E(\hat{\beta}_0) = E(b_0) = \beta_0 \qquad V(b_0) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

Please note:

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma_\varepsilon^2 \frac{\bar{x}}{S_{xx}}$$

$$SS_E = \sum_{t=1}^{n} e_t^{\,2} = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

Error Sum of Squares

$$E(SS_E) = (n-2)\sigma_\varepsilon^{\,2} \Rightarrow \hat{\sigma}_\varepsilon^{\,2} = \frac{SS_E}{n-2}$$

```
•S = 0.0001015    R-Sq = 70.5%    R-Sq(adj) = 69.9%
```

In general: S is the standard deviation of residuals. $S = \hat{\sigma}_\varepsilon$

Mean Squared Error (MS_E)

$$\hat{\sigma}_\varepsilon^2 = MS_E = \frac{SS_E}{df_E} = \frac{SS_E}{n-K}$$

with *K*=number of regressors+1

• S = 0.0001015

$$S = \hat{\sigma}_{\varepsilon}$$

Estimated variance

Estimated std deviation

$$\sigma_{b_0}^2 = V(b_0) = \sigma_{\varepsilon}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \qquad \hat{\sigma}_{b_0}^2 = \hat{\sigma}_{\varepsilon}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \qquad \hat{\sigma}_{b_0} = s_{b_0} = \sqrt{\hat{\sigma}_{b_0}^2}$$

$$\sigma_{b_1}^2 = V(b_1) = \frac{\sigma_{\varepsilon}^2}{S_{xx}} \qquad \hat{\sigma}_{b_1}^2 = \frac{\hat{\sigma}_{\varepsilon}^2}{S_{xx}} \qquad \hat{\sigma}_{b_1} = s_{b_1} = \sqrt{\hat{\sigma}_{b_1}^2}$$

Predictor	Coef	SE Coef	T	P
Constant	0.00136522	**0.00002914**	46.84	0.000
t	-0.00001066	**0.00000099**	-10.72	0.000

TEST 1

$$H_0 : \beta_i = 0 \qquad H_1 : \beta_i \neq 0$$

$\Rightarrow b_i \ (i = 0,1)$ normally distributed with estimated standard deviation given by s_{b_i}

$$t_0 = \frac{b_i - \beta_i}{s_{b_i}} \sim t_{n-K}$$

t Student $n - K$ degree of freedom (dof)

K = number of regressors + 1 -in the example 1! regressor $(x_t = t) \Rightarrow$ K = 2

Ex

$H_0: \beta_1 = 0$

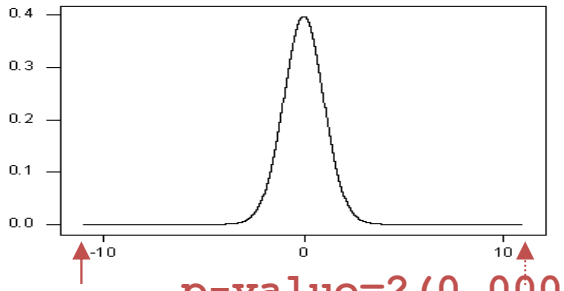
Predictor	Coef	SE Coef	T	P
Constant	0.00136522	0.00002914	46.84	0.000
t	-0.00001066	0.00000099	-10.72	0.000

$$t_0 = \frac{b_1 - 0}{s_{b_1}} = -10.7$$

Cumulative Distribution Function

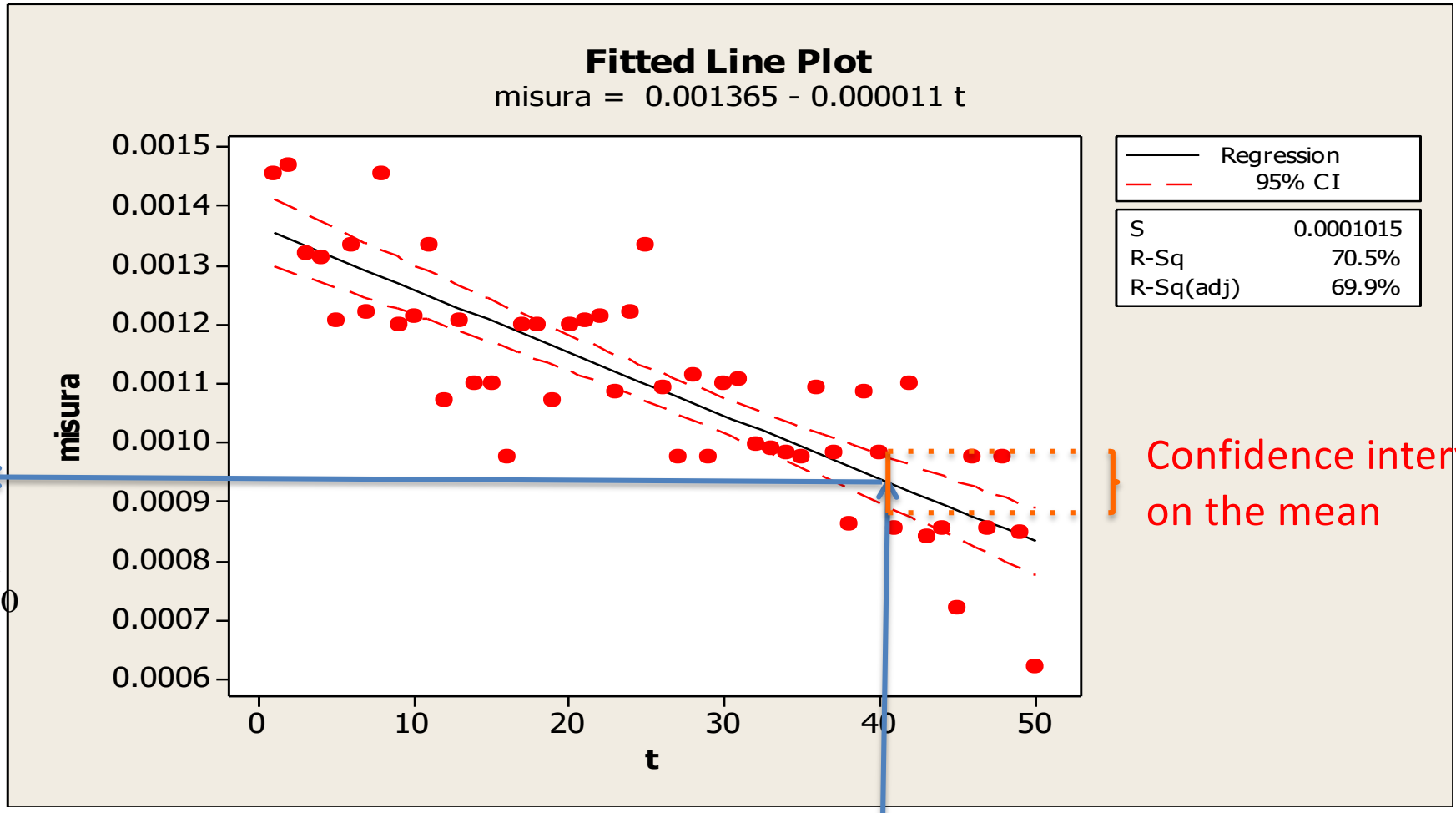
Student's t distribution with 48 DF

x	P(X <= x)
-10.7700	0.0000



p-value=2(0.0000)

Point estimate and confidence interval on the mean



$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Confidence interval on the mean

x_0

Confidence interval for the mean

$$E(Y|x_0) = \mu_{Y|x_0} = \mu_0$$

$$\hat{\sigma}^2_{b_0} = \hat{\sigma}^2_\varepsilon \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

For a given x_t=x_0

$$\hat{\mu}_0 = b_0 + b_1 x_0$$

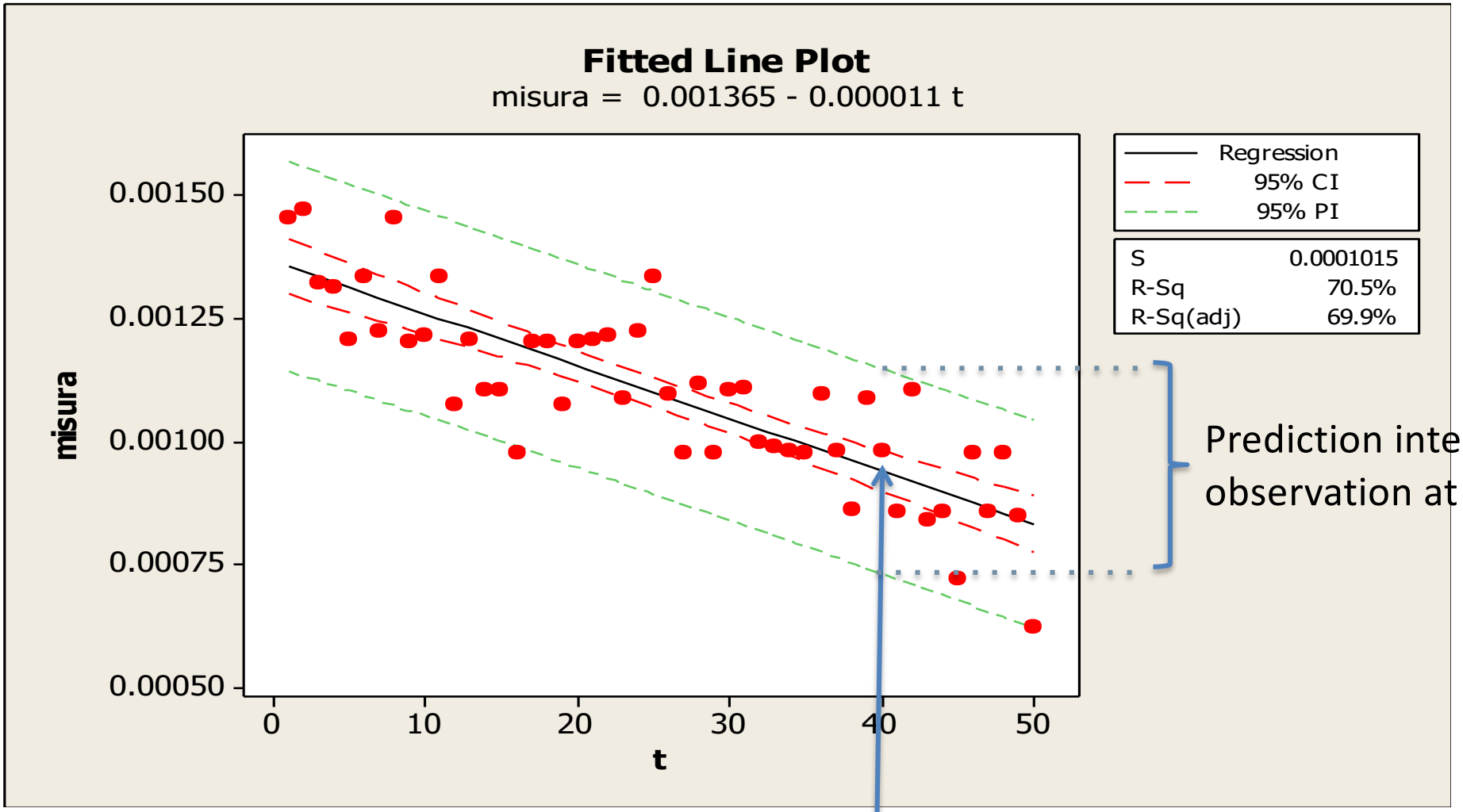
Point estimate

$$V(\hat{\mu}_0) = \hat{\sigma}^2_\varepsilon \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

$$\frac{\hat{\mu}_0 - \mu_0}{\sqrt{V(\hat{\mu}_0)}} \sim t_{n-2}$$

$$\hat{\mu}_0 \pm t_{\alpha/2,n-2}\sqrt{V(\hat{\mu}_0)}$$

Point prediction and preditcion interval on data



Prediction of a new observation

New observation at x_0 $\quad Y|x_0 = Y_0$

Point prediction at x_0: $\quad \hat{y}_0 = \hat{\mu}_0 = b_0 + b_1 x_0$

Prediction error $\quad Y_0 - \hat{y}_0$

$$V(Y_0 - \hat{y}_0) = \sigma_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

Note that in the confidence interval for the mean we used $V(\hat{\mu}_0) = \hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$

$$\hat{\mu}_0 \pm t_{\alpha/2,n-2} \sqrt{\hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

Minitab:

```
Predicted Values for New Observations

NewObs Fit    SE Fit        95.0% CI          95.0% PI

1   0.000821 0.000029 (0.000763,0.000880)(0.000609,0.001034)
```

Two sources of uncertainty: $Y = \mu_t + \varepsilon_t \rightarrow \overset{2}{V}(\hat{\mu}_t) + \overset{1}{\sigma}_\varepsilon^2$

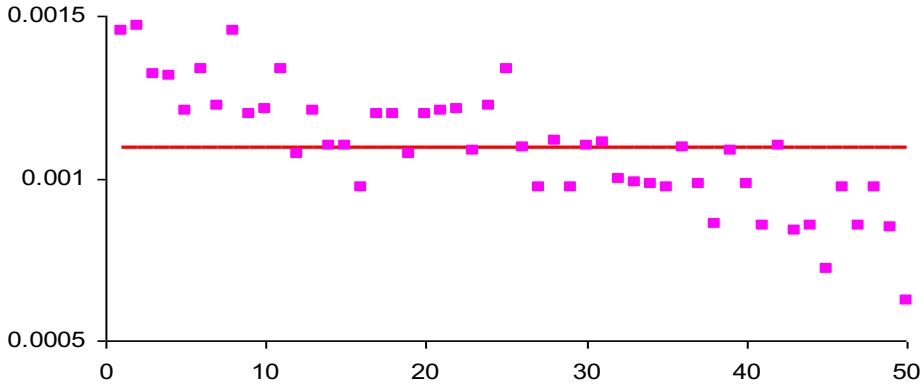
1. Variability of data about the estimated model (S)
2. Uncertainty of the estimated model with respect to the true one

For n reasonably large (n>=30), the second source of variability is small

$$H_0: \quad \beta_1 = \beta_2 = ... = \beta_{K-1} = 0$$
$$H_1: \quad \beta_j \neq 0 \quad \text{for at least one j}$$

NB:
- K=number of regressors (p)+1=2
- The constant β_0 is excluded

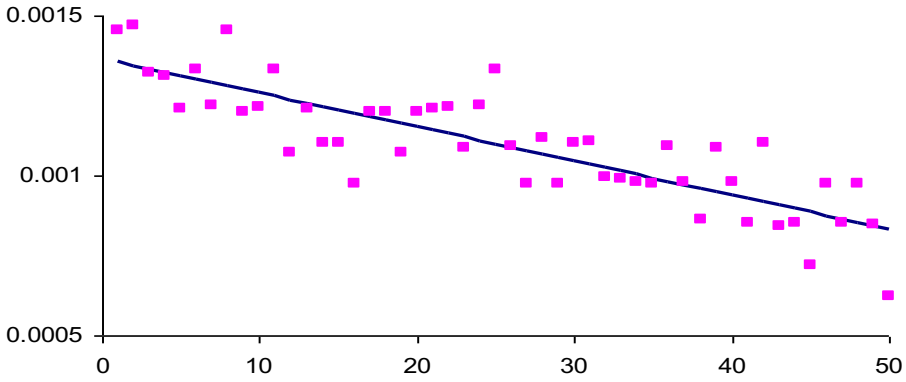
Based on AnOVa (Analysis of Variance) test

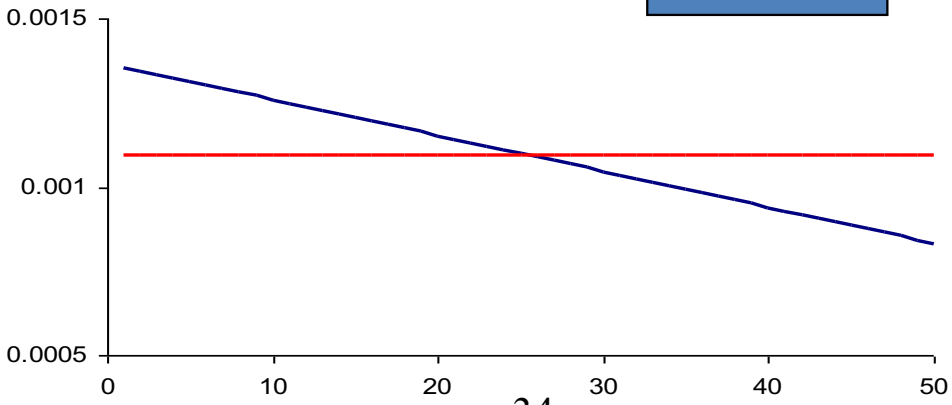


$$\sum_{t=1}^{n}(y_t - \bar{y})^2 =$$

$$= \sum_{t=1}^{n}(y_t - \hat{y}_t)^2 + \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2$$

DIM





$$\sum_{t=1}^{n}(y_t-\bar{y})^2=\sum_{t=1}^{n}(y_t-\hat{y}_t)^2+\sum_{t=1}^{n}(\hat{y}_t-\bar{y})^2$$

$$SS_T = SS_E + SS_R$$

Degree of freedom

$df_R = K-1$	Variability explained by the regression model	$SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$
$df_E = n-K$	Variability not explained by the regression model	$SS_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
$df_T = n-1$	Total variability	$SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2$

Analysis of Variance

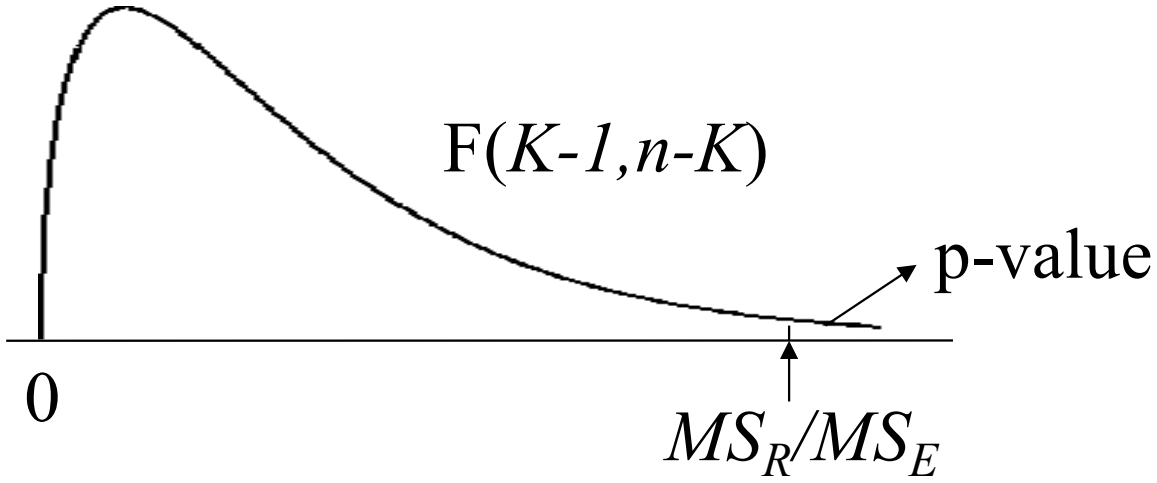
Source	DF	SS	MS	F	P
Regression	1	1.18428E-06	1.18428E-06	114.96	0.000
Residual Error	48	4.94479E-07	1.03017E-08		
Total	49	1.67876E-06			

$$MS_R = \frac{SS_R}{df_R} \qquad MS_E = \frac{SS_E}{df_E}$$

$$\text{se } H_0 \text{ è vera, allora } \frac{MS_R}{MS_E} \sim F(K-1, n-K)$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1.18428E-06	1.18428E-06	114.96	0.000
Residual Error	48	4.94479E-07	1.03017E-08		
Total	49	1.67876E-06			



1-1.0000

Cumulative Distribution Function

F distribution with 1 DF in numerator and 48 DF in denominator

x	P(X <= x)
114.9600	1.0000

1) Relationship between TEST 1 and TEST2. With one single coefficient

$$H_0: \ \beta_1 = 0$$
$$H_1: \ \beta_1 \neq 0$$

The two tests are identical. Indeed, it can be shown that:

$$t_0 = \frac{b_i - 0}{s_{b_i}} \sim t_{n\text{-}K} \Rightarrow (t_0)^2 = \left(\frac{b_i - 0}{s_{b_i}}\right)^2 = \frac{MS_R}{MS_E} \sim F(1, n\text{-K})$$

2) for TEST 1: In order to use the test on *N* coefficients at the same time, consider the family error rate *a* (it is actually better using a different approach - the extra sum of squares)

$$\alpha_i = \frac{\alpha'_{nom}}{N} \qquad \forall i = 1,\ldots,N \Rightarrow \alpha' \leq \sum_{i=1,..,N} \alpha_i = \alpha'_{nom}$$

3) There is another test (lack of fit) that requires more than one replicate for each value of the regressors

S = 0.0001015 R-Sq = 70.5% R-Sq(adj) = 69.9%

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ Variability explained by the regression model

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ Variability not explained by the regression model

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_R + SS_E$$

$$R^2 = \frac{SS_R}{SS_T}$$

It is a measure of the percentage of variability observed in the data that is explained by the estimated regression model

Comments on R^2

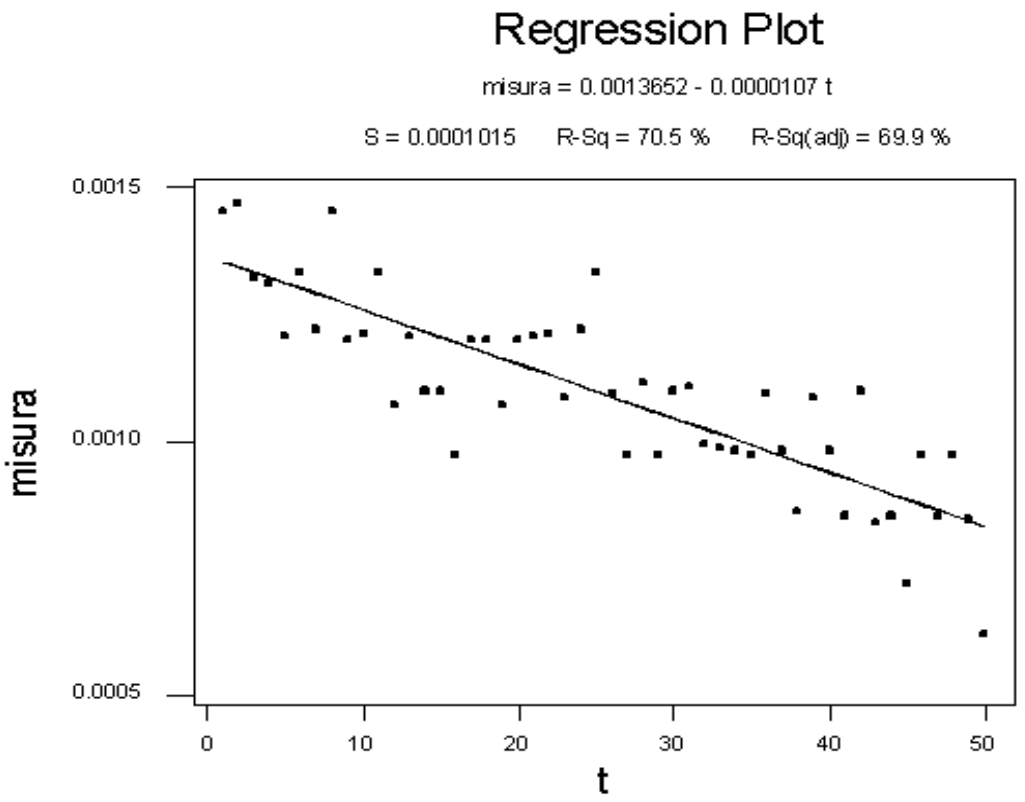
Pay attention to use R^2 as a measure of adequacy of the estimated model (check of residuals)

Please note that R^2 does never decrease when new regressors are included

$$R^2 = \frac{SS_R}{SS_T} = \frac{SS_T - SS_E}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

$$R_{adj}^{\ 2} = 1 - \frac{SS_E / n - K}{SS_T / n - 1}$$

trade-off between reduction of the SSE and reduction of n-K



$$\hat{y}_t = b_0 + b_1 x_t$$

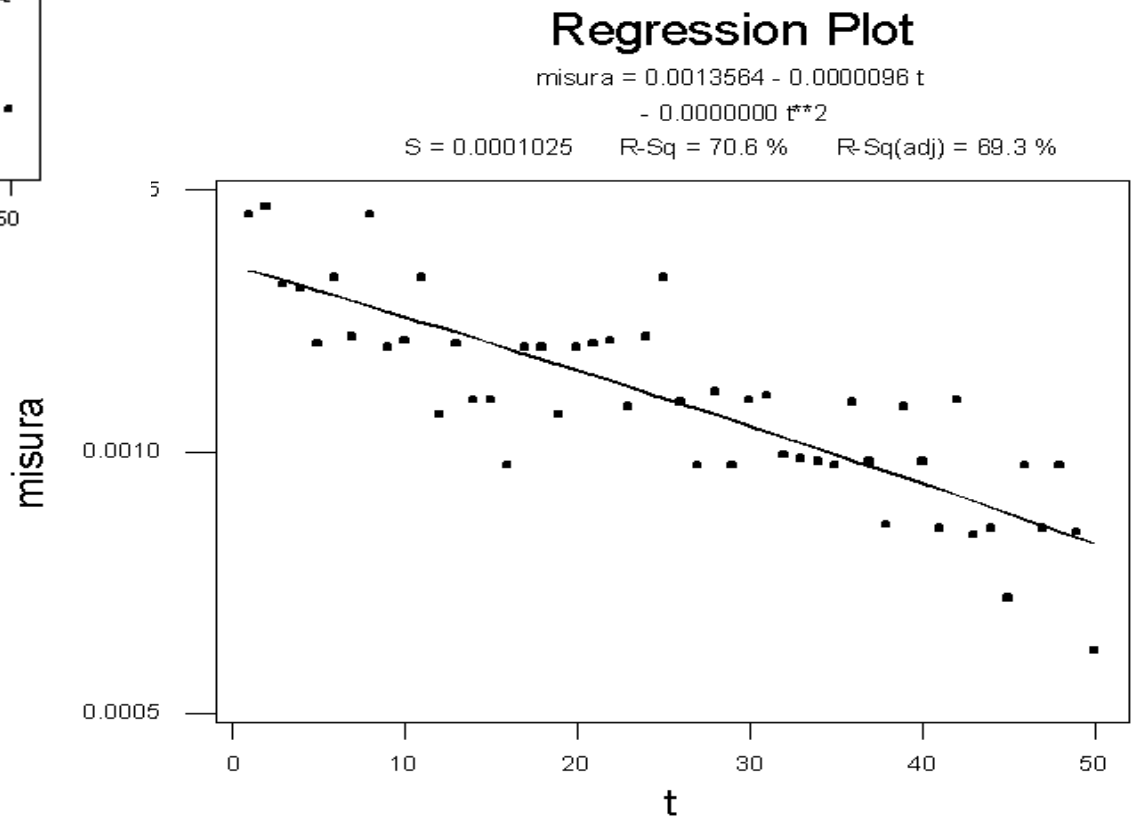
R^2=70.5%

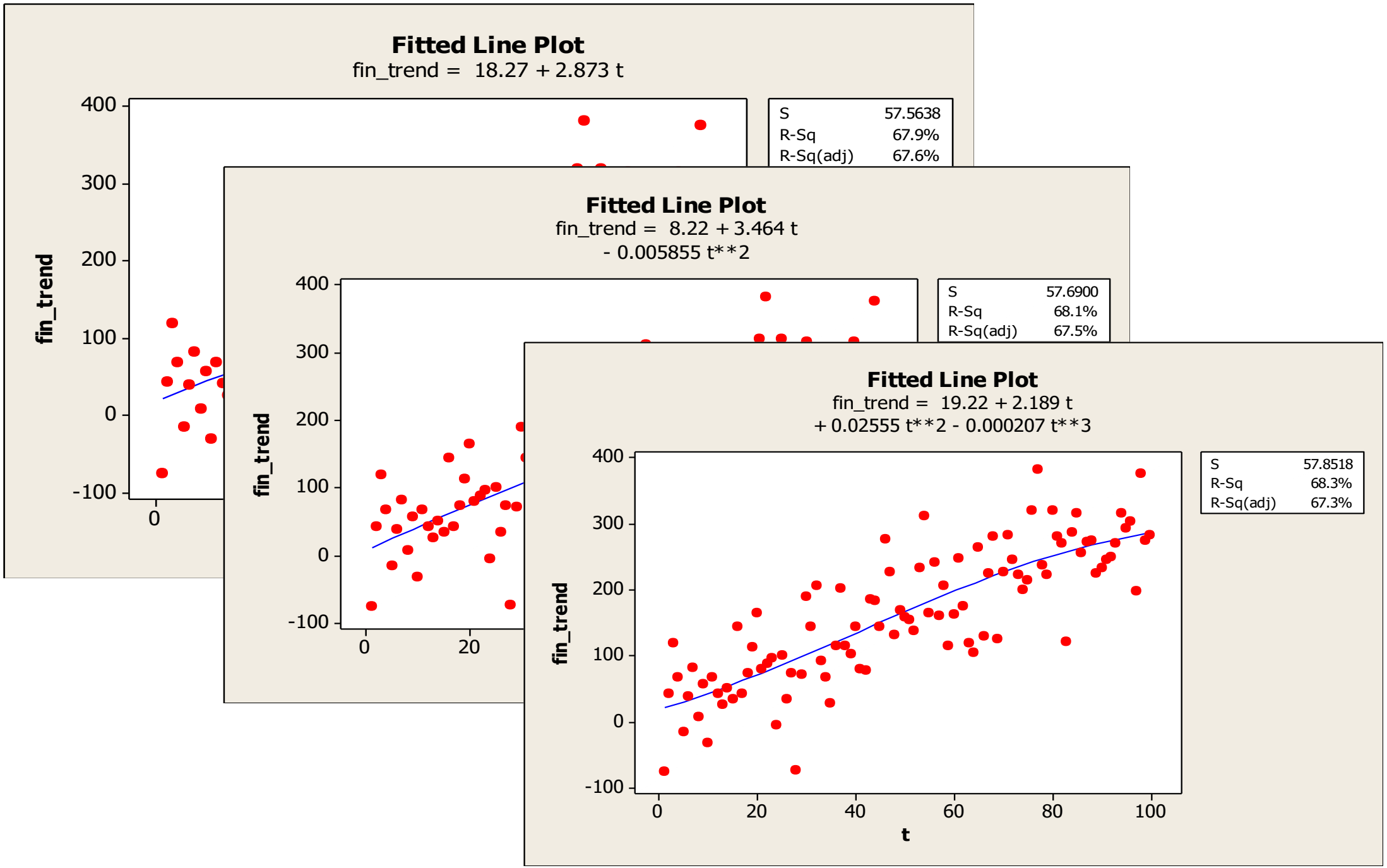
R^2_{adj}=69.9%

$$\hat{y}_t = b_0 + b_1 x_t + b_2 x_t^2$$

R^2=70.6%

R^2_{adj}=69.3%




Fitted Line Plot
fin_trend = 18.27 + 2.873 t
S 57.5638
R-Sq 67.9%
R-Sq(adj) 67.6%
Fitted Line Plot
fin_trend = 8.22 + 3.464 t
- 0.005855 t**2
S 57.6900
R-Sq 68.1%
R-Sq(adj) 67.5%
Fitted Line Plot
fin_trend = 19.22 + 2.189 t
+ 0.02555 t**2 - 0.000207 t**3
S 57.8518
R-Sq 68.3%
R-Sq(adj) 67.3%
fin_trend
t
400
300
200
100
0
-100
0
20
40
60
80
100

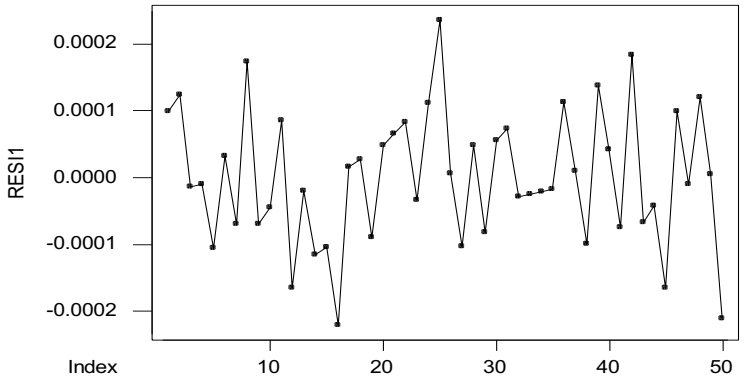
Residual check

1. Observed data = deterministic component + random error
2. Observed data = estimated model + residuals

Check the assumptions (errors are iid and normally distributed)

Randomness (before checking normality):

- Time series plot:
- Runs test
- Normality
- «Unocorrelation» (Bartlett)



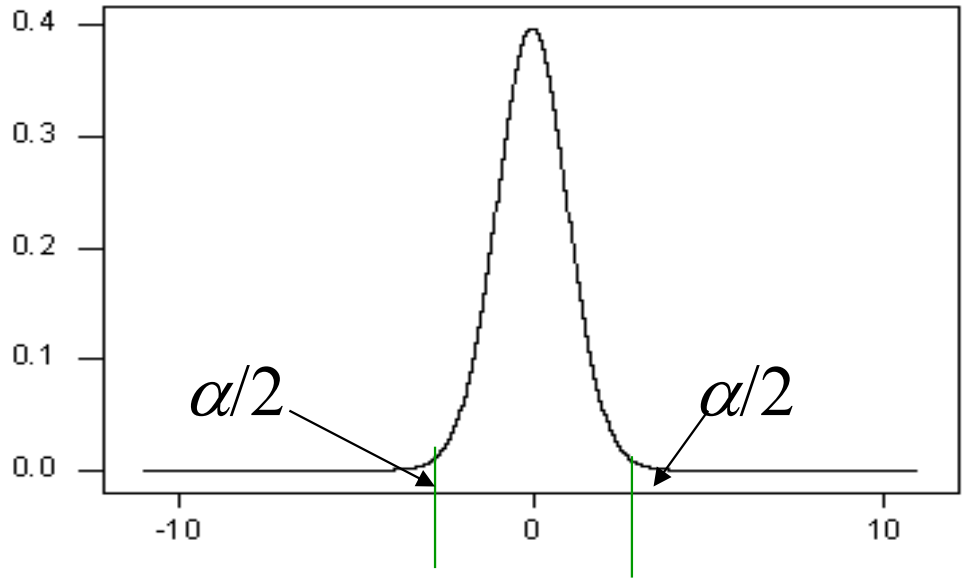
Additional Slides

Confidence interval (approx) for coefficients

$$t_0 = \frac{b_i - \beta_i}{s_{b_i}} \sim t_{n-K}$$

$$-t_{\alpha/2,n-K} \le \frac{b_i - \beta_i}{s_{b_i}} \le t_{\alpha/2,n-K}$$

$$\boxed{b_i - t_{\alpha/2,n-K} s_{b_i} \le \beta_i \le b_i + t_{\alpha/2,n-K} s_{b_i}}$$



for β_1 :

$$\alpha = 5\% \Rightarrow t_{0.025,48} = 2.0106$$

$$-1.066 \cdot 10^{-5} - 2.0106(0.099 \cdot 10^{-5}) \le \beta_1 \le -1.066 \cdot 10^{-5} + 2.0106(0.099 \cdot 10^{-5})$$

$$-1.265 \cdot 10^{-5} \le \beta_1 \le -0.867 \cdot 10^{-5}$$