Machine Learning in Geophysics

Lecture 1 – Nomenclature and Basics

Outline

Definition

> *Machine Learning is the study of computer algorithms that improve automatically through experience. Applications range from datamining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests.*

Mitchell, 1997

Earliest examples already in the 1960s, e.g. "Learning Machines", Nilsson 1965

Machine learning setup

- Set of inputs x_{ij} with $i = 1 \ldots N$ samples and $j = 1 \ldots M$ features
- For supervised learning we have associated expected outputs y_i
- Machine learning algorithm with parameters θ_k and hyper-parameters λ_l
- Choose hyper-parameters λ_l (regularization, iterations, target precision, etc.)
- Adjust parameters θ_k through training
- Can then make predictions for other input

Types of training

Classes of machines learning algorithms

Supervised learning Learning by example, provide algorithm with input and expected outcome (training), then make predictions with unknown input

Unsupervised learning No examples, algorithm tries to identify patterns or cluster in data

Types of data

Output data

Classification: Output is discrete

Regression: Output is continuous

Types of learning

Instance based: Compare new data with training data

Model based: Create an abstracted representation of data (model) and make predictions based on model.

Some examples

- Straight line fitting (model based, supervised learning, regression)
- Support vector machines (model based, supervised learning, classification/regression)
- Decision Trees (model based, supervised learning, classification/regression)
- Random Forests (model based, supervised learning, classification/regression)
- k-means Clustering (model based, unsupervised learning, classification)
- Neural Networks (model based, supervised learning, classification/regression)

Machine learning is data analysis

First principle of data analysis

GIGO – Garbage In = Garbage Out

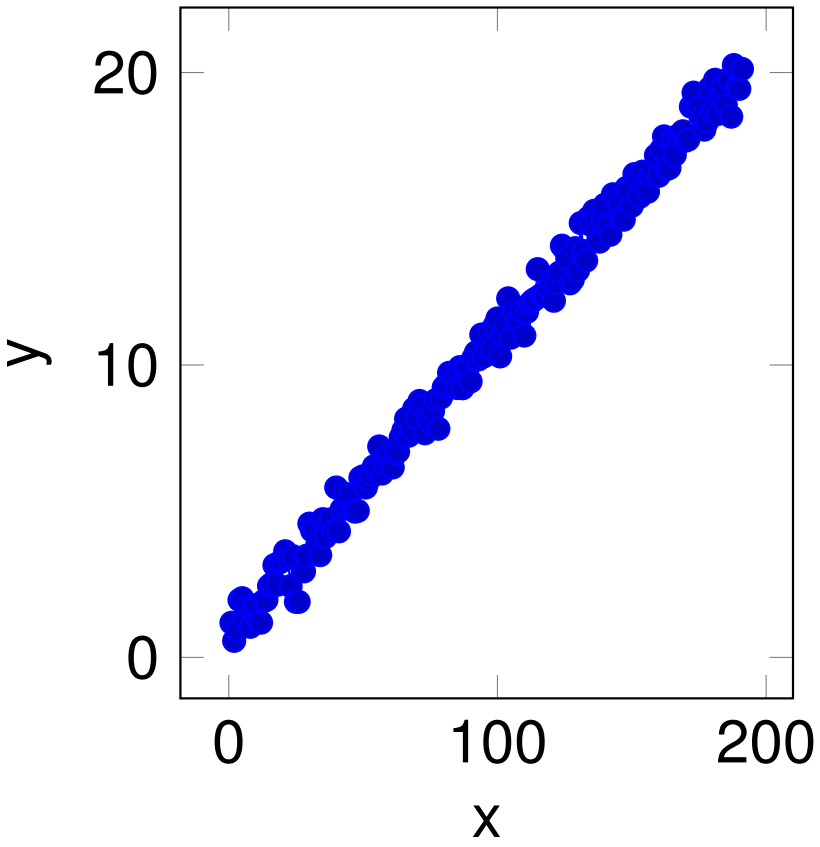
- If your assumptions are wrong,
- If your data is bad quality,
- If you don't really know what it is you want to achieve,

You can use very sophisticated mathematical methods to give you completely wrong/meaningless results without ever knowing it.

Challenges

- Formulate meaningful questions and find the right data to answer them
- For supervised learning, get representative training data
- Get data in a useable form
- Identify appropriate method and parameters
- Understand what the predictions are telling you

Fitting a line

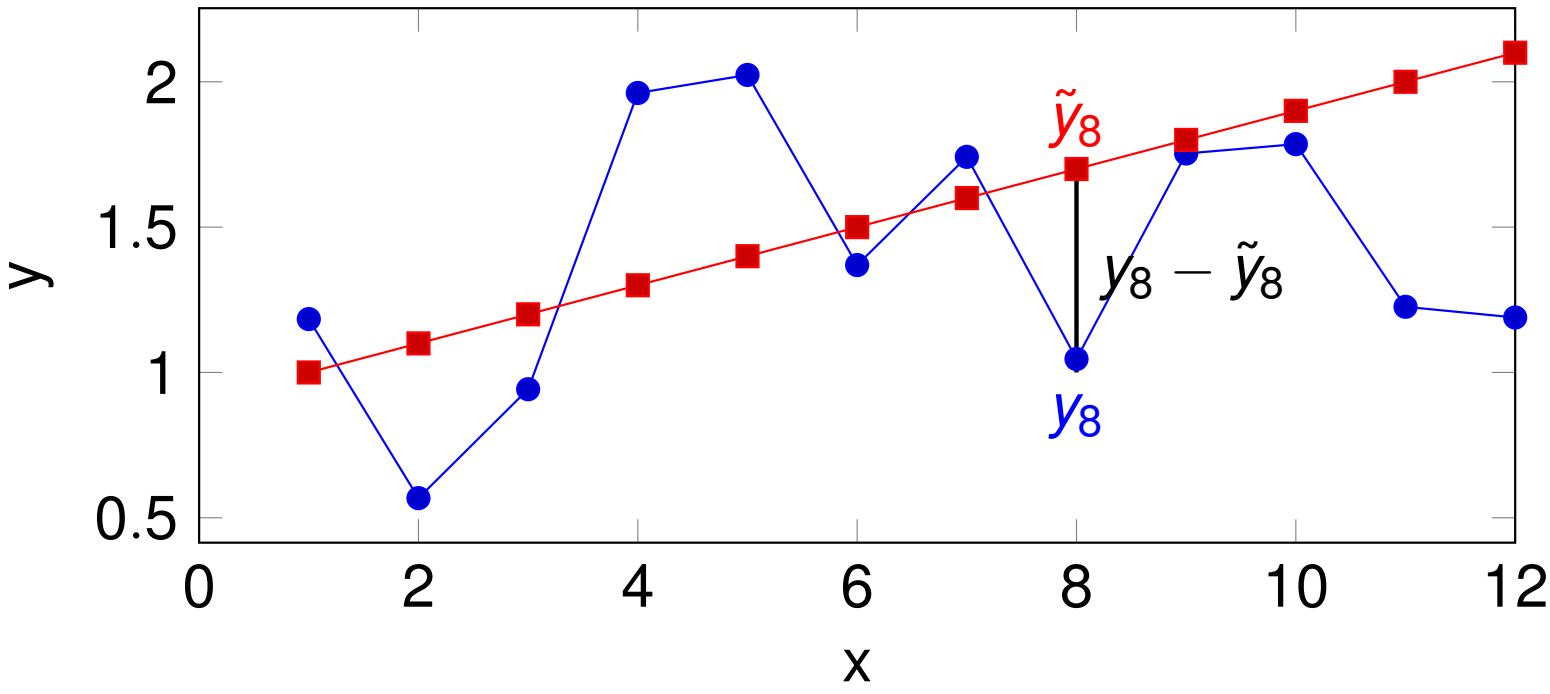


If we have some noisy measurements and we assume a relationship

$$y = ax + b,$$

we can use linear regression to estimate a and b.

Least-squares



We have n pairs of values (x_i, y_i) and we want to find the values a and b for which the overall squared difference is minimal

$$\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{N} (y_i - ax_i - b)^2 \to \min$$

Minimization

So we have to solve

$$\frac{\partial}{\partial a}\sum_{i=1}^{N}(y_i - ax_i - b)^2 = 0$$

$$\frac{\partial}{\partial b}\sum_{i=1}^{N}(y_i - ax_i - b)^2 = 0$$

Linear regression

We can write direct equations for a and b

$$a = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{\mathrm{Cov}(x,y)}{\sigma_x^2}$$

$$b = \frac{1}{N}\left(\sum_{i=1}^{N} y_i - b\sum_{i=1}^{N} x_i\right) = \bar{y} - a\bar{x}.$$

Goodness of fit

We define the *total sum of squares* as

$$SS_T = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

and the *sum of squares due to regression*

$$SS_R = \sum_{i=1}^{N} (\tilde{y}_i - \bar{y})^2 .$$

The *error sum of squares*

$$SS_E = SS_T - SS_R = \sum_{i=1}^{N} (\tilde{y}_i - y_i)^2$$

shows the scatter of the data around the line.

Goodness of fit

$$R = \sqrt{\frac{SS_R}{SS_T}}$$

is very similar to the correlation coefficient for the regression.

- We can show that $SS_R \leq SS_T$ and thus $0 \leq R \leq 1$.
- If our estimated line matches all datapoints, we have $SS_R = SS_t$ and thus $R = 1$.

Another look at linear regression

We can write the calculation of the values predicted by the linear relationship in matrix-vector form

$$\tilde{\mathbf{y}} = \mathbf{G}\mathbf{m}$$

or in long form

$$\begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_N \end{pmatrix} = \underbrace{\begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix}}_{\mathbf{G}} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\mathbf{m}} = \begin{pmatrix} ax_1 + b \\ \vdots \\ ax_N + b \end{pmatrix}$$

Generalized inverse

We want to solve

$$|\mathbf{y} - \tilde{\mathbf{y}}|^2 = |\mathbf{y} - \mathbf{Gm}|^2 = (\mathbf{y} - \mathbf{Gm})^T (\mathbf{y} - \mathbf{Gm}) \rightarrow \min,$$

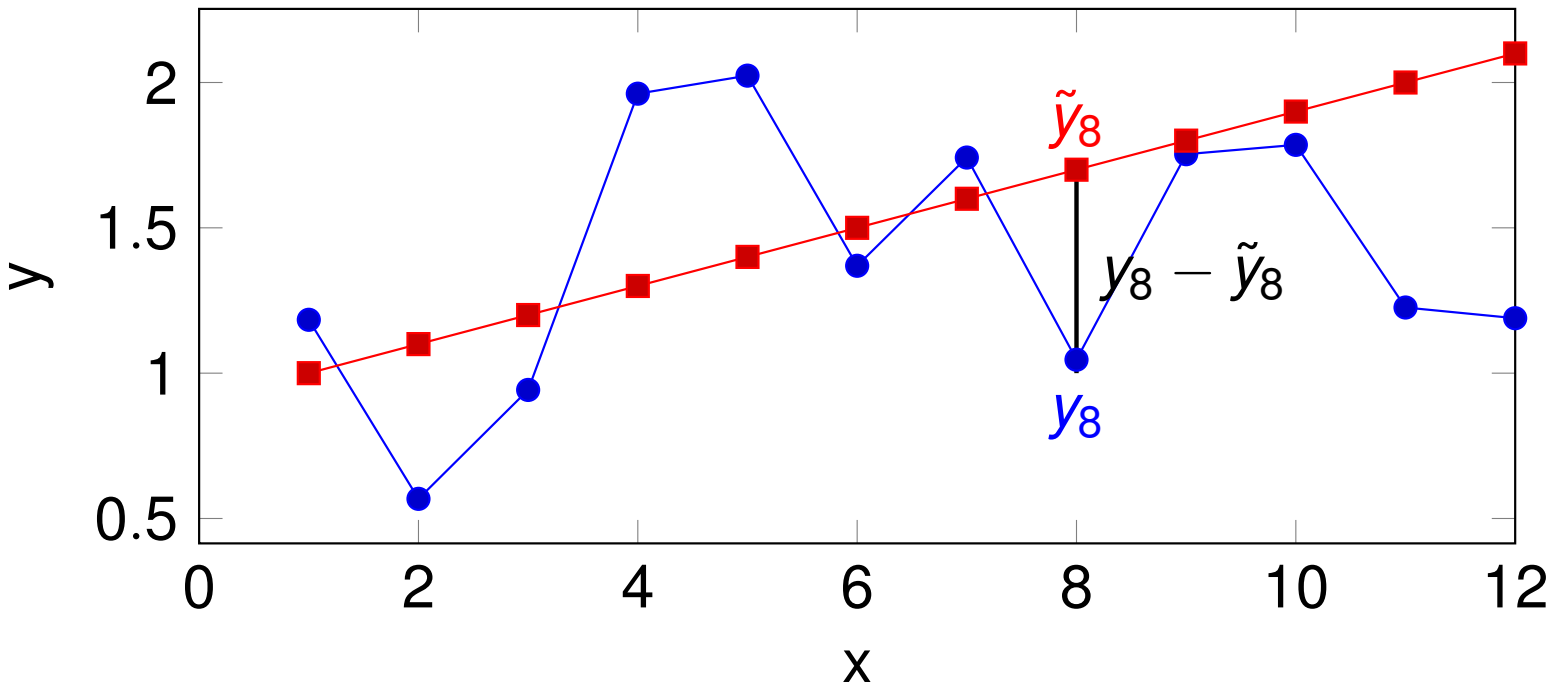
the solution is given by calculating the generalized inverse $\mathbf{G}^p$ of $\mathbf{G}$

$$\mathbf{G}^p = \left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T.$$

We then get an estimate for $\mathbf{m}$ from the observed values $\mathbf{y}$

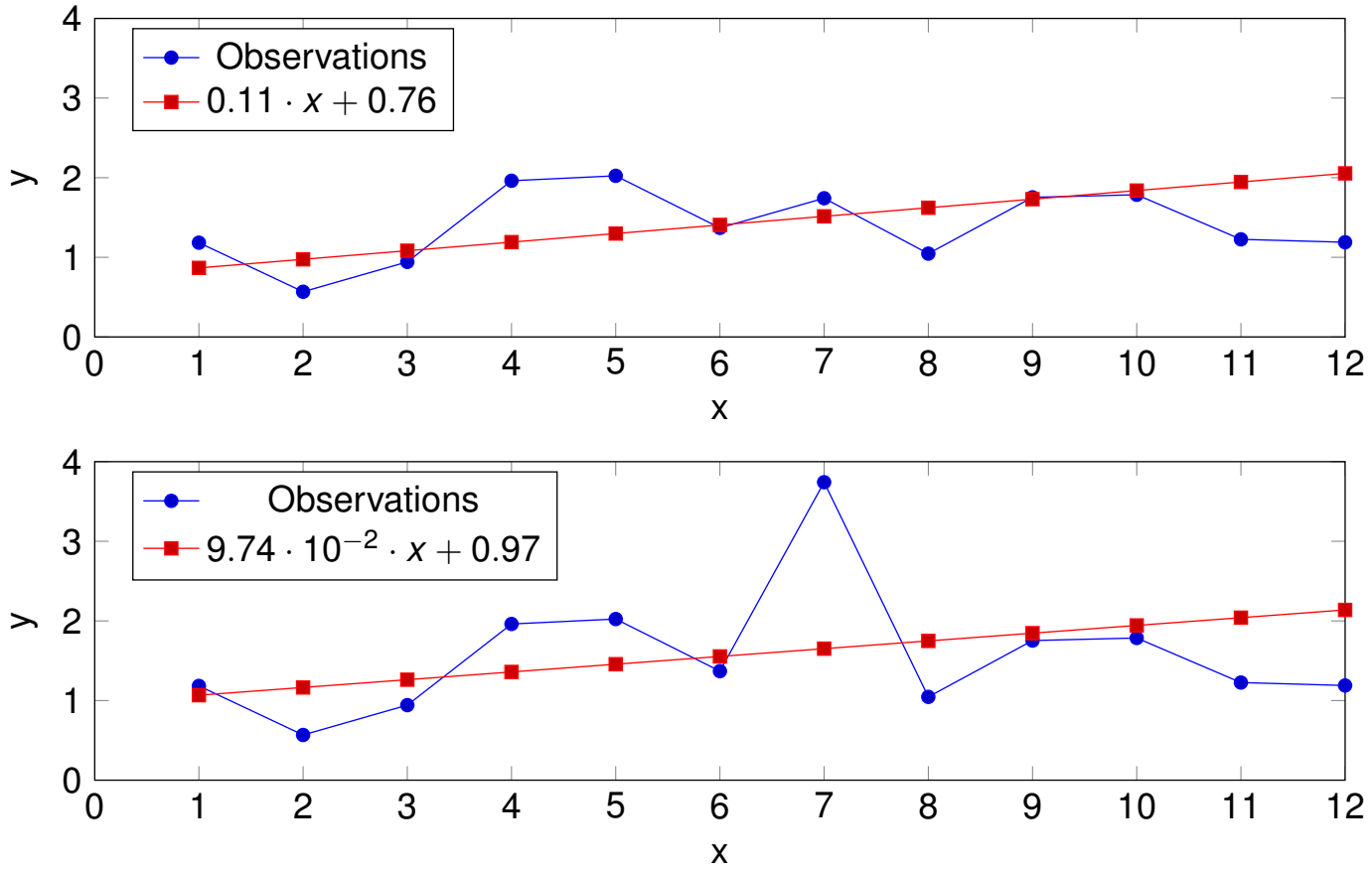
$$\mathbf{m}^* = \mathbf{G}^p\mathbf{y} = \left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\mathbf{y}$$

Least-squares



So far we have not considered data errors when fitting a line.
We assume all errors are identical.

Least-squares



So far we have not considered data errors when fitting a line.
We assume all errors are identical.

Errors

Question

How can we modify

$$\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{N} (y_i - ax_i - b)^2 \to \min$$

to take into account the errors σ_i?

Errors

Question

How can we modify

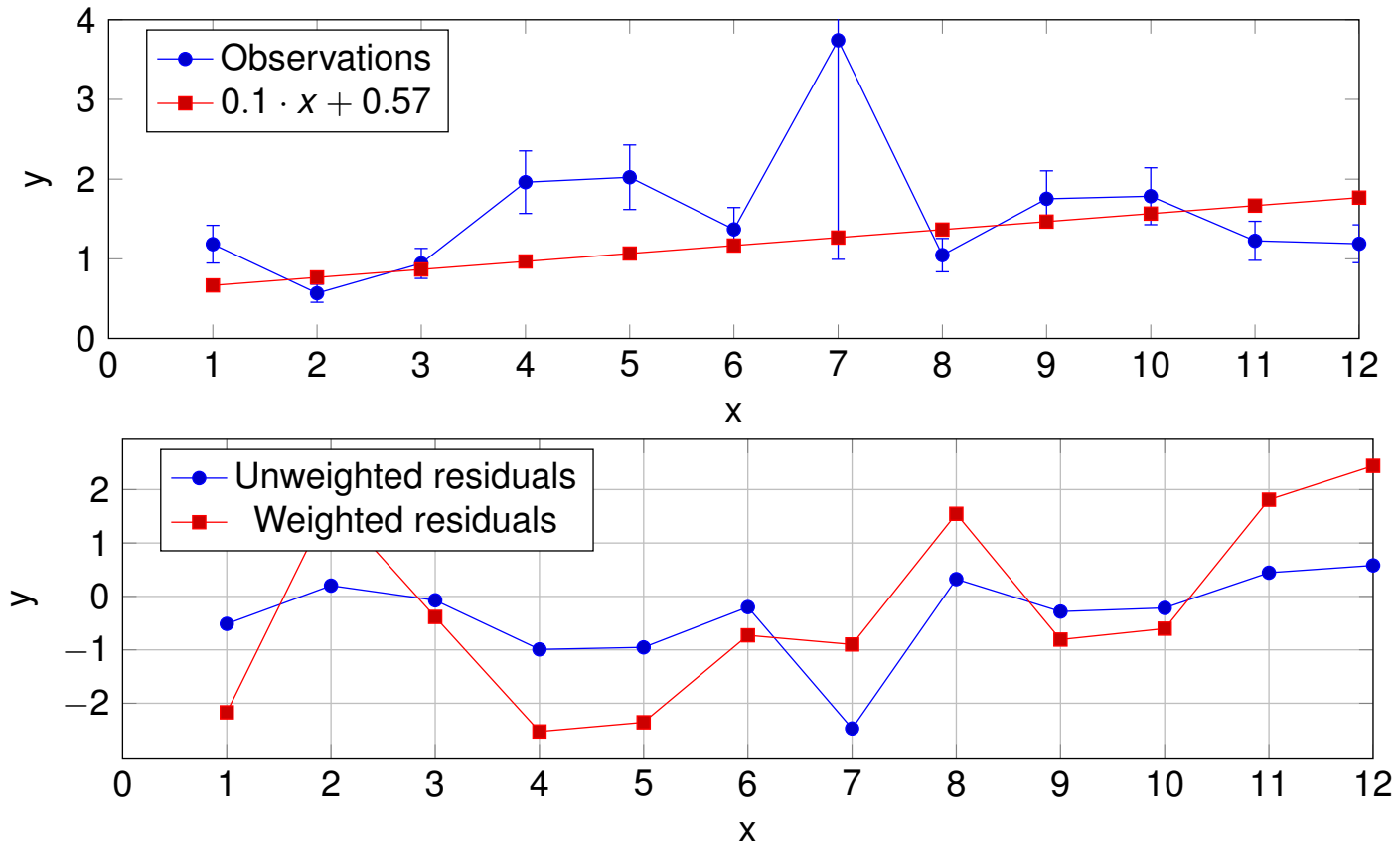
$$\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{N} (y_i - ax_i - b)^2 \to \min$$

to take into account the errors σ_i?

Answer

$$\sum_{i=1}^{N} \left(\frac{y_i - ax_i - b}{\sigma_i} \right)^2 \to \min$$

Errors



We compare the residuals $r_i = \frac{y_i - ax_i - b}{\sigma_i}$.

Weighted inverse

Remember the original problem

$$(\mathbf{y} - \mathbf{Gm})^T (\mathbf{y} - \mathbf{Gm}) \to \min,$$

and the unweighted generalized inverse

$$\mathbf{m}^* = \left(\mathbf{G}^T\mathbf{G}\right)^{-1} \mathbf{G}^T\tilde{\mathbf{y}}.$$

We can introduce a weighting matrix

$$\mathbf{W} = \begin{pmatrix} \frac{1}{\sigma_1^2} & & & \\ & \frac{1}{\sigma_2^2} & \mathbf{0} & \\ & & \ddots & \\ & \mathbf{0} & & \ddots \\ & & & \frac{1}{\sigma_N^2} \end{pmatrix}$$

Weighted inverse

And solve the weighted problem

$$(\mathbf{y} - \mathbf{Gm})^T \mathbf{W} (\mathbf{y} - \mathbf{Gm}) \to \min,$$

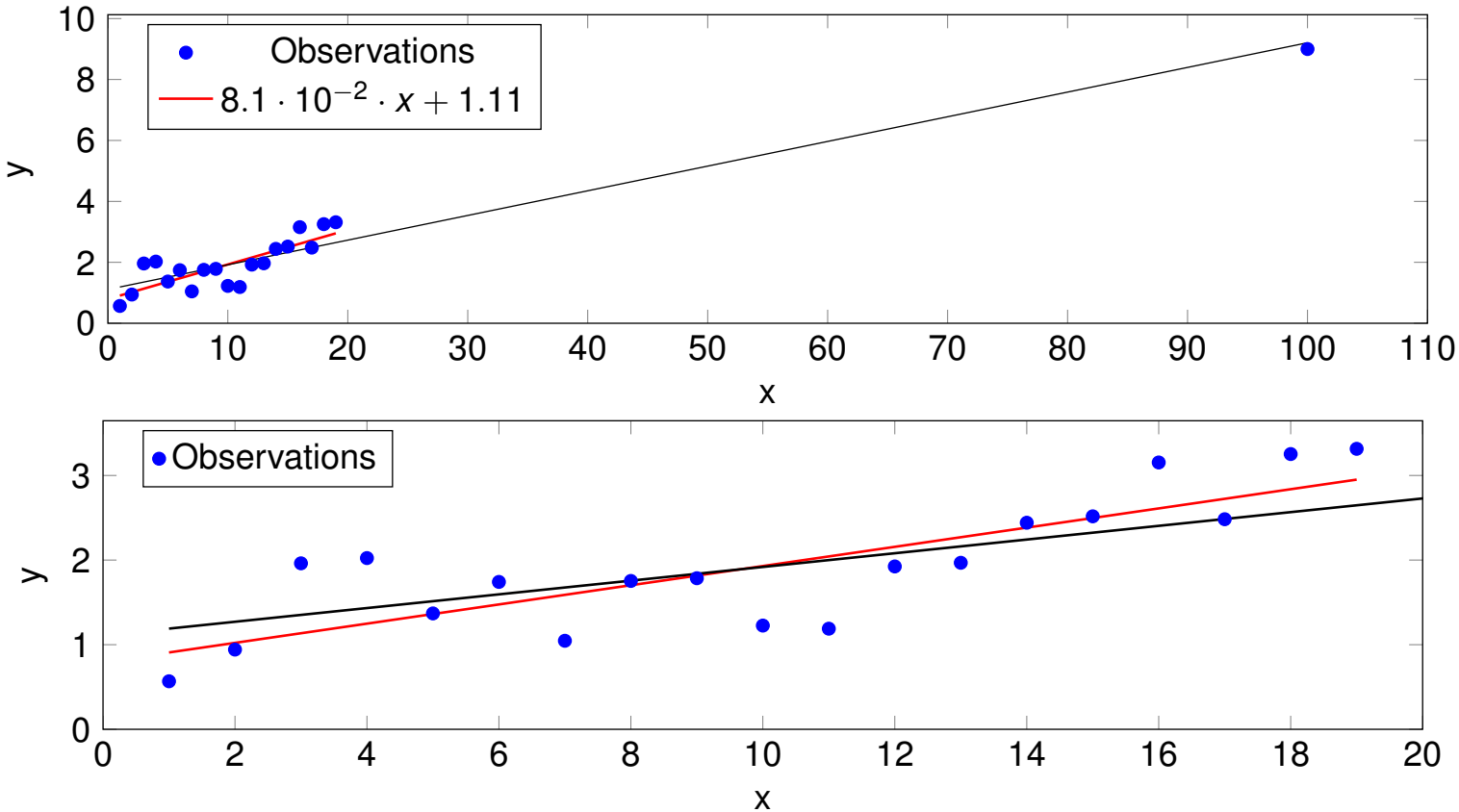
The weighted generalized inverse

$$\mathbf{m}^* = \left(\mathbf{G}^T\mathbf{WG}\right)^{-1} \mathbf{G}^T\mathbf{Wy}.$$

is very similar to the unweighted form

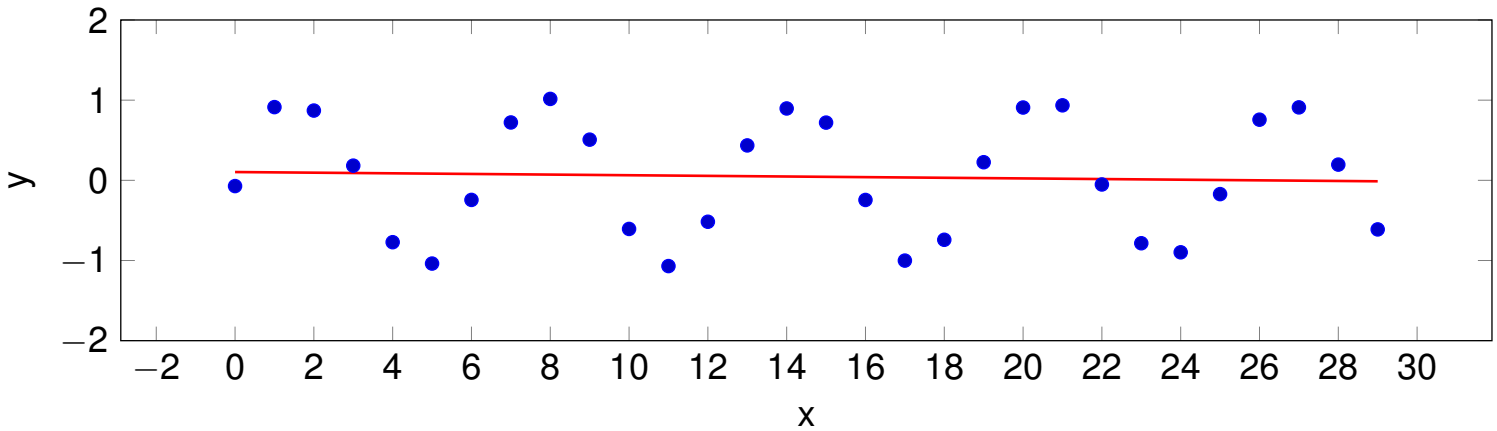
$$\mathbf{m}^* = \left(\mathbf{G}^T\mathbf{G}\right)^{-1} \mathbf{G}^T\mathbf{y}.$$

Leverage points



Single points with large values can strongly influence the result, even if they are not very unusual.

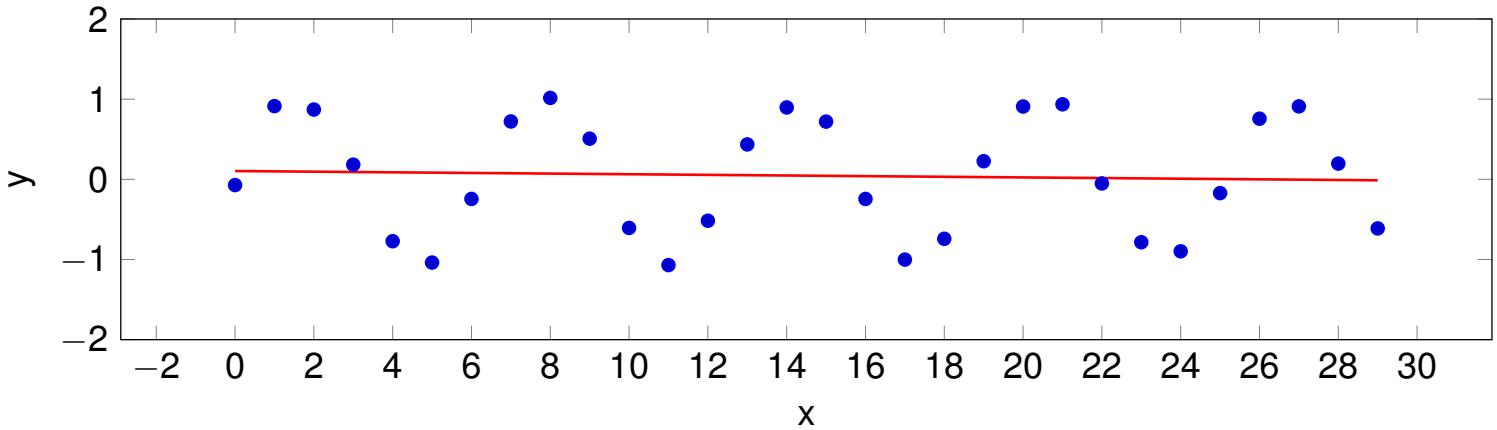
Model discrepancy



Question

Is the fitted red line an adequate representation of the data (blue dots)?

Model discrepancy



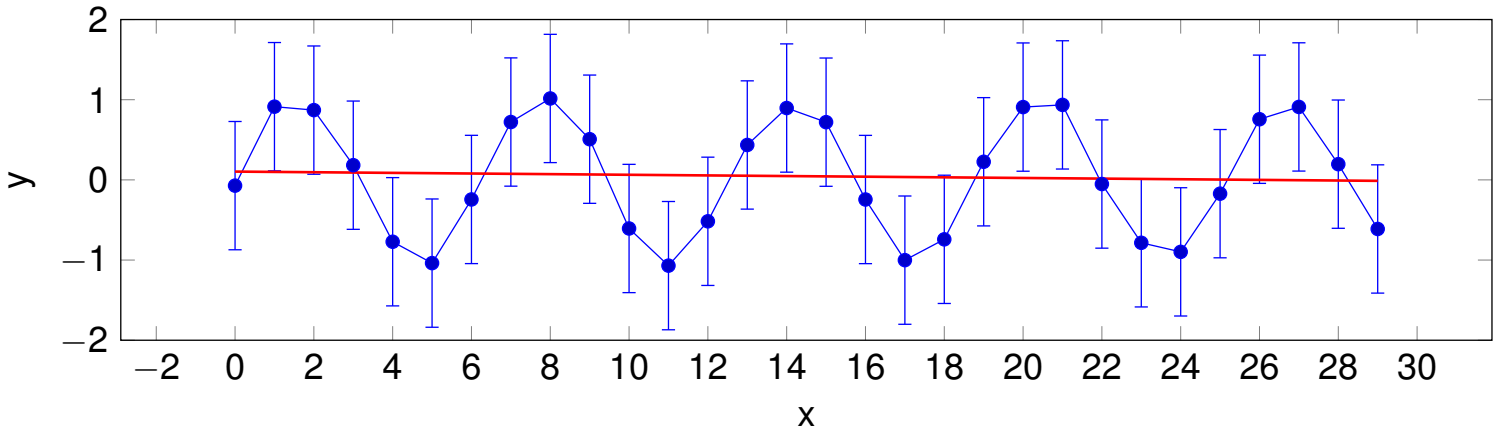
Question

Is the fitted red line an adequate representation of the data (blue dots)?

Answer

You can't answer this question without error bars !

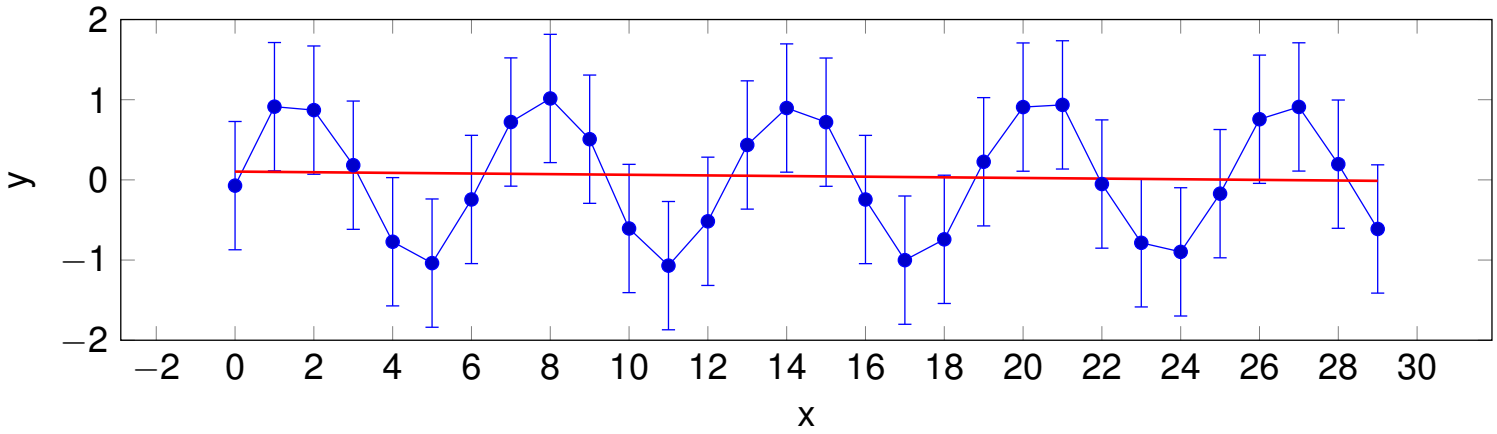
Model discrepancy



Question

Is the fitted red line an adequate representation of the data (blue dots)?

Model discrepancy



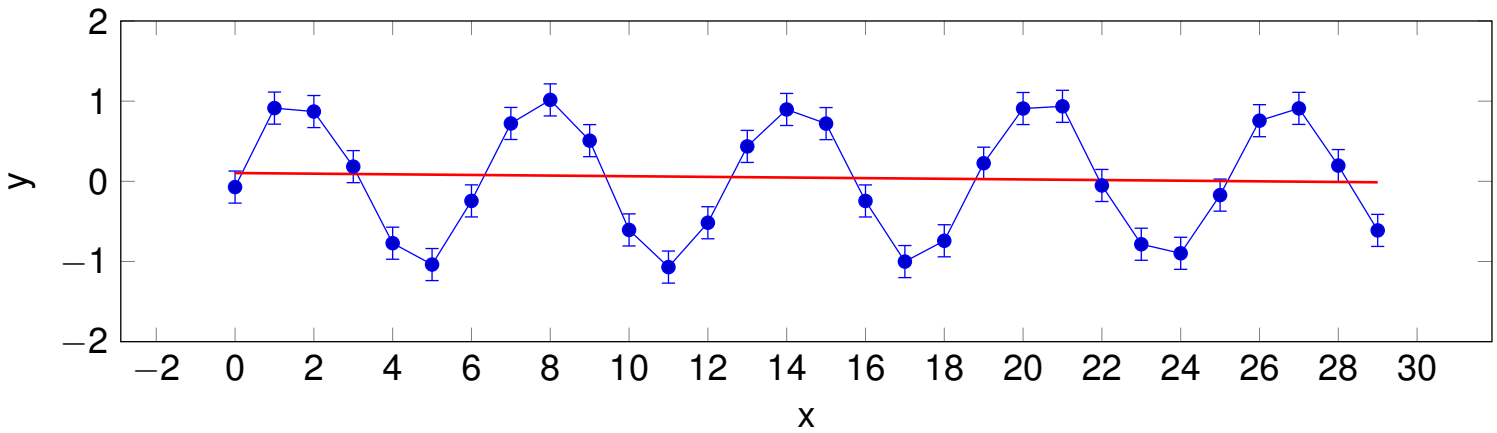
Question

Is the fitted red line an adequate representation of the data (blue dots)?

Answer

Statistically yes, but something strange is going on with the way you measure.

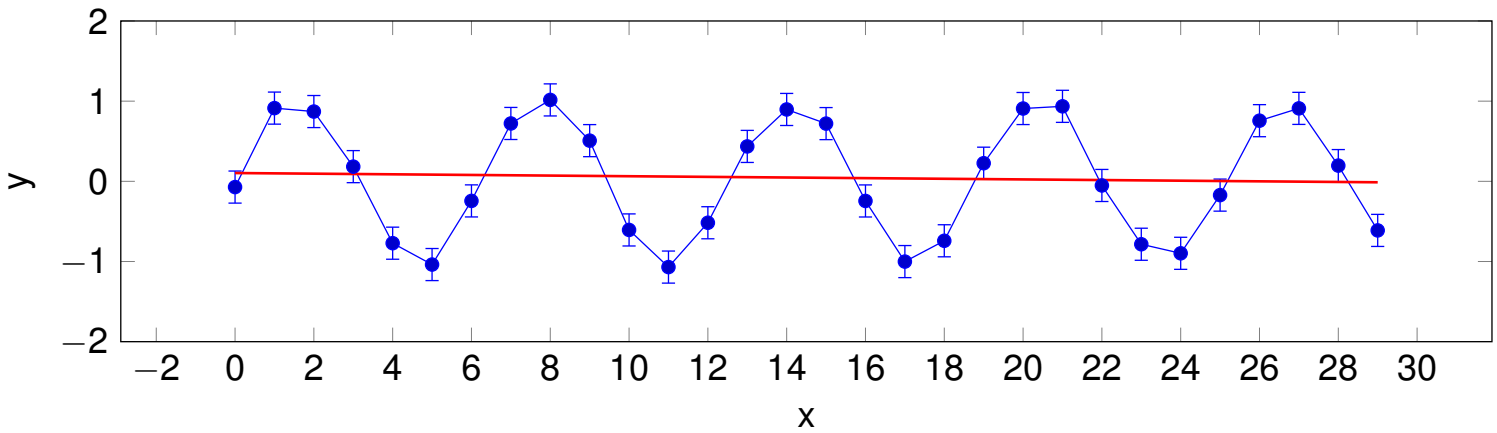
Model discrepancy



Question

Is the fitted red line an adequate representation of the data (blue dots)?

Model discrepancy



Question

Is the fitted red line an adequate representation of the data (blue dots)?

Answer

No, in fact all data were generated using $y = \sin(x)$ with added noise.