Applied Statistics

Lecture 14+15

David Barton & Sabine Hauert

Department of Engineering Mathematics

Outline

- Linear regression
- Conditions for linear regression
- Residual plots
- Linear regression with more complicated functions

OpenIntro Statistics

Chapter 7

Linear regression

Data fitting is an important part of statistics.

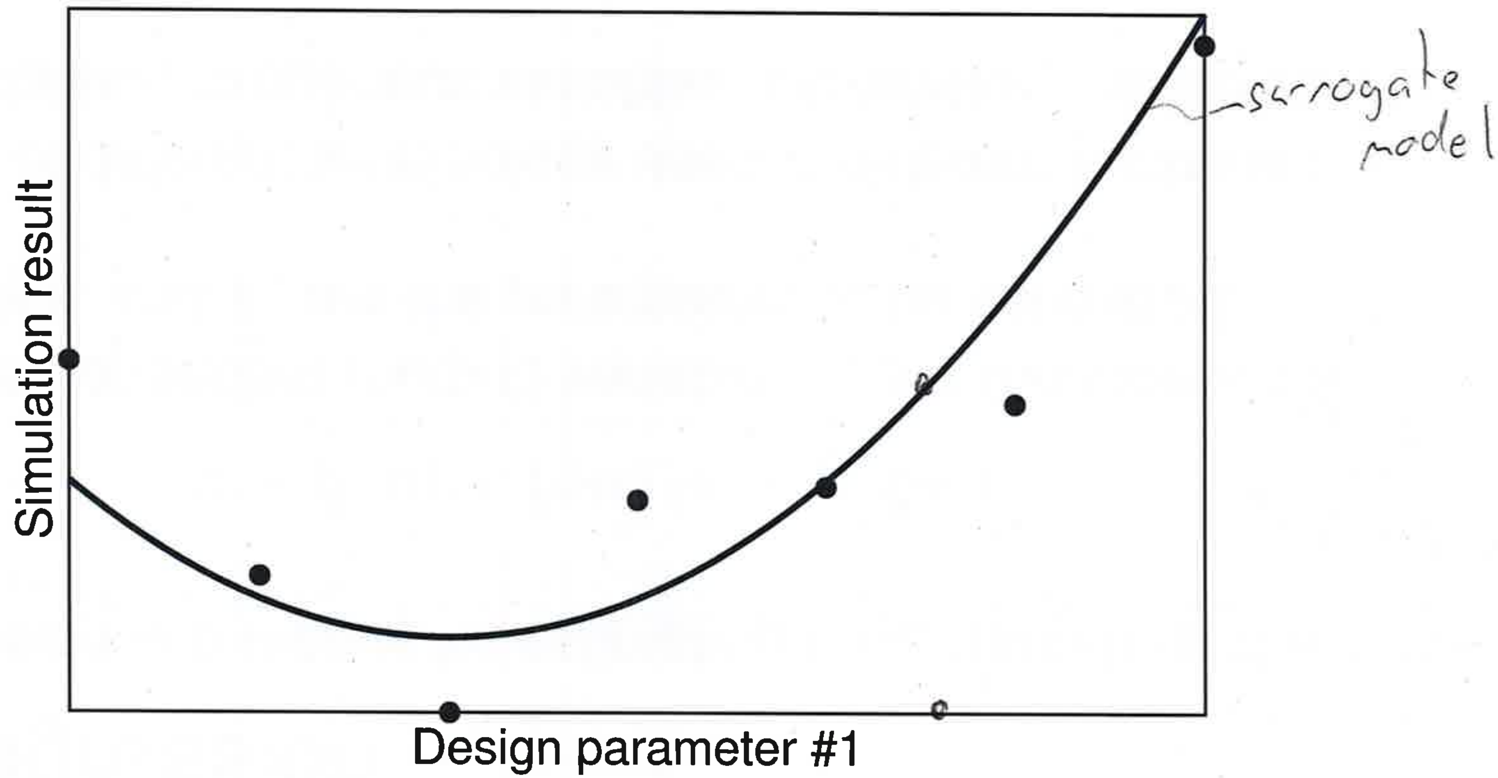
- Given a particular model, what parameters best fit the data?
  - Modal analysis is a big part of structural engineering!
- More data than parameters: regular data fitting ← focus
- More parameters than data: inverse problems!
  - Most of medical imaging
  - Source reconstruction in acoustics
  - Optics, radar, communications, nondestructive testing, . . .

quite challenging

How do you design large-scale structures?

only run costly simulations a few times and use the data points to build a model that allows you to make predictions about your system.

Finite element simulations are slow. . .



Use a *surrogate model* for quick computation and the full model for final testing. (Also known as a metamodel, reduced-order model or emulator.)

Linear regression

Linear regression is a means for estimating the parameters of a model of the form

approximation

inputs

general form

$$\boxed{y = \beta_1 u_1 + \beta_2 u_2 + \cdots + \beta_p u_p}$$

$y = \beta_1 + \beta_2 x$
$u_1 = 1$
$u_2 = 2$

where y is the dependent (output) variable, u_i are the independent (input) variables, and β_i are the parameters to be estimated.

(There is lots of different terminology used in different textbooks — regressor variables, exogenous variables, explanatory variables, etc.)

This model is *linear in the parameters*; it can be that the model is nonlinear in terms of the independent variables! For example,

linear in parameters

$$y = \boxed{\beta_1} x + \boxed{\beta_2} x^2 + \boxed{\beta_3} x^3$$

find the best "β" values to match inputs to the outputs

fits in the framework of linear regression.

Linear regression

With linear regression, y and x_i are known from n different samples.
Hence we have

$\begin{bmatrix} x \\ y \end{bmatrix} \begin{bmatrix} 1/2 \\ 1 \end{bmatrix} \begin{bmatrix} 3/4 \\ 1.2 \end{bmatrix} \begin{bmatrix} 1 \\ 1.3 \end{bmatrix}$

$y = \beta_1 + \beta_2 x$

$$y_1 = \beta_1 u_{1,1} + \beta_2 u_{1,2} + \cdots + \beta_p u_{1,p}$$
$$y_2 = \beta_1 u_{2,1} + \beta_2 u_{2,2} + \cdots + \beta_p u_{2,p}$$
$$\vdots$$
$$y_n = \beta_1 u_{n,1} + \beta_2 u_{n,2} + \cdots + \beta_p u_{n,p}$$

$$\left.\begin{aligned} 1 &= \beta_1 + \beta_2 \tfrac{1}{2} \\ 1.2 &= \beta_1 + \beta_2 \tfrac{3}{4} \\ 1.3 &= \beta_1 + \beta_2 1 \end{aligned}\right\}$$ can't solve exact, 3 equation, 2 unknown

In matrix-vector form

vector of outputs

vector of parameters

$$y = U\beta$$

matrix of inputs

$y = \begin{bmatrix} 1 \\ 1.2 \\ 1.3 \end{bmatrix}$ $U = \begin{bmatrix} 1 & 1/2 \\ 1 & 3/4 \\ 1 & 1 \end{bmatrix}$ $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

For a particular choice of β construct the *residual* or *error vector*

residual, error vector; actual output; output of the model

$$e = y - U\beta$$

Minimising error

What choice of β minimises the error?

First, what is meant by small since the error is a vector? Obvious answer is the Euclidean norm

$$\|e\| = \left(\sum_i e_i^2\right)^{\frac{1}{2}}$$

which gives us least-squares.

But it's not the only answer!

- Least absolute deviation (more robust in some circumstances)

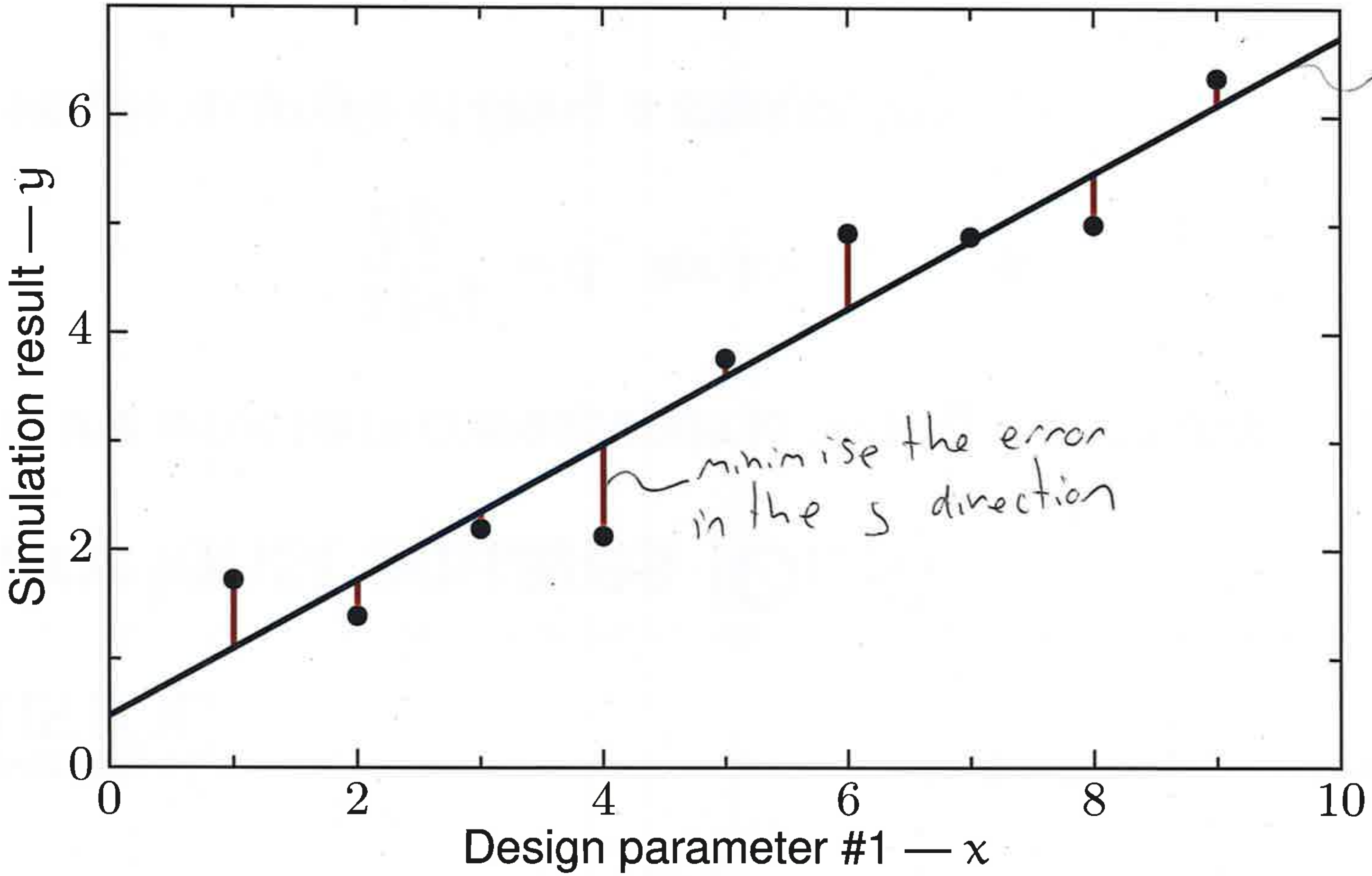
$$\|e\| = \sum_{i=1}^{n} |e_i|$$

often better if you have outliers

- For inverse problems: Tikhonov regularisation or the Lasso method

Ordinary least squares (OLS)

Ordinary least squares corresponds to minimising the error in the y direction — errors in the x direction are ignored!



result of the model

$y = \beta_1 + \beta_2 X$

Ordinary least squares (OLS)

Minimising the error term corresponds to finding β such that

$$\frac{d\,\|e\|}{d\,\beta_i} = 0 \quad \text{for } i = 1, \ldots, p$$

Take the simple example of fitting a straight line

$$y = \beta_1 + \beta_2 x$$

Here we have

output model output

$$\|e\|^2 = \sum_{i=1}^{n} (y_i - \overline{\beta_1 - \beta_2 x_i})^2$$

Note that minimising $\|e\|^2$ gives the same results as minimising $\|e\|$ in this context.

Ordinary least squares (OLS)

Differentiating $\|e\|^2$ w.r.t. β_1 and β_2 and equating to zero gives

$$\frac{d\,\|e\|^2}{d\,\beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i) = 0$$

$$\frac{d\,\|e\|^2}{d\,\beta_2} = -2\sum_{i=1}^{n}x_i(y_i - \beta_1 - \beta_2 x_i) = 0$$

These linear algebraic equations are called the *normal equations*

$$\beta_1 \boxed{n} + \beta_2 \boxed{\sum_{i=1}^{n} x_i} = \boxed{\sum_{i=1}^{n} y_i}$$

$$\beta_1 \boxed{\sum_{i=1}^{n} x_i} + \beta_2 \boxed{\sum_{i=1}^{n} x_i^2} = \boxed{\sum_{i=1}^{n} x_i y_i}$$

known values

Example calculation

$y = B_1 + B_2 x + B_3 x^2$ quadratic model

$\|e\|^2 = \sum_i (y_i - B_1 - B_2 x - B_3 x^2)^2$

$\frac{d\|e\|^2}{B_1} = -2 \sum_i (y_i - B_1 - B_2 x_i - B_3 x_i^2) = 0$

$\frac{d\|e\|^2}{B_2} = -2 \sum_i x_i (y_i - B_1 - B_2 x_i - B_3 x_i^2) = 0$

$\frac{d\|e\|^2}{B_3} = -2 \sum_i x_i^2 (y_i - B_1 - B_2 x_i - B_3 x_i^2) = 0$

$$\left.\begin{aligned} B_1 n + B_2 \sum x_i + B_3 \sum x_i^2 &= \sum y_i \\ B_1 \sum x_i + B_2 \sum x_i^2 + B_3 \sum x_i^3 &= \sum x_i y_i \\ B_1 \sum x_i^2 + B_2 \sum x_i^3 + B_3 \sum x_i^4 &= \sum x_i^2 y_i \end{aligned}\right\}$$

solve this to find B

→ use matlab polyfit

Maximum likelihood

The errors in the y direction are often assumed to be normally distributed i.i.d.; in this case least-squares is a maximum likelihood estimator (MLE)

Given a set of points y_i (ignore x for simplicity), what are the parameters of the normal distribution that are most likely to have generated that set of points?

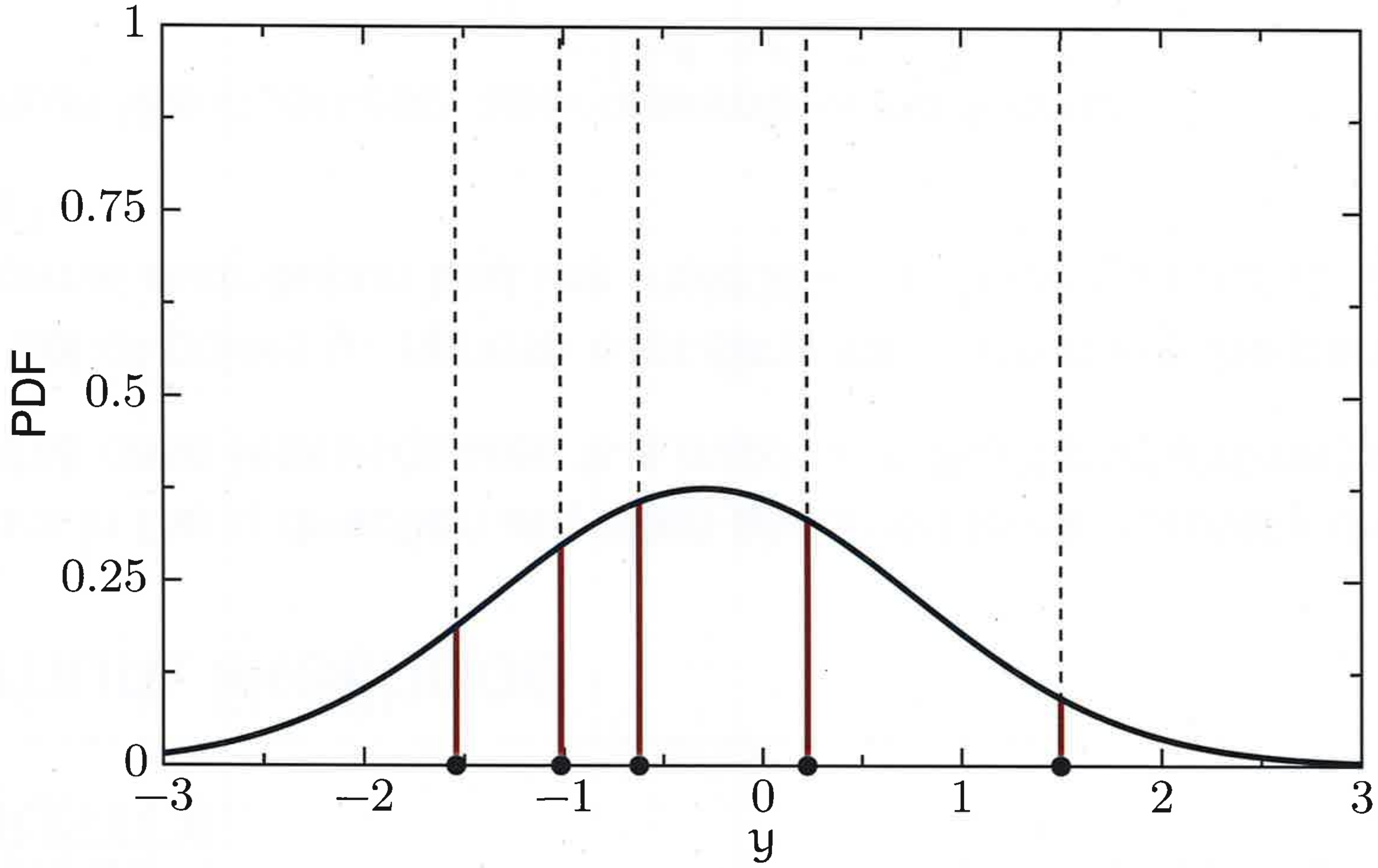
A *maximum likelihood estimator* maximises the function

find a normal distribution that best fits y and

$$\prod_{i=1}^{n} p_Y(y_i \mid \mu_Y, \sigma_Y^2)$$

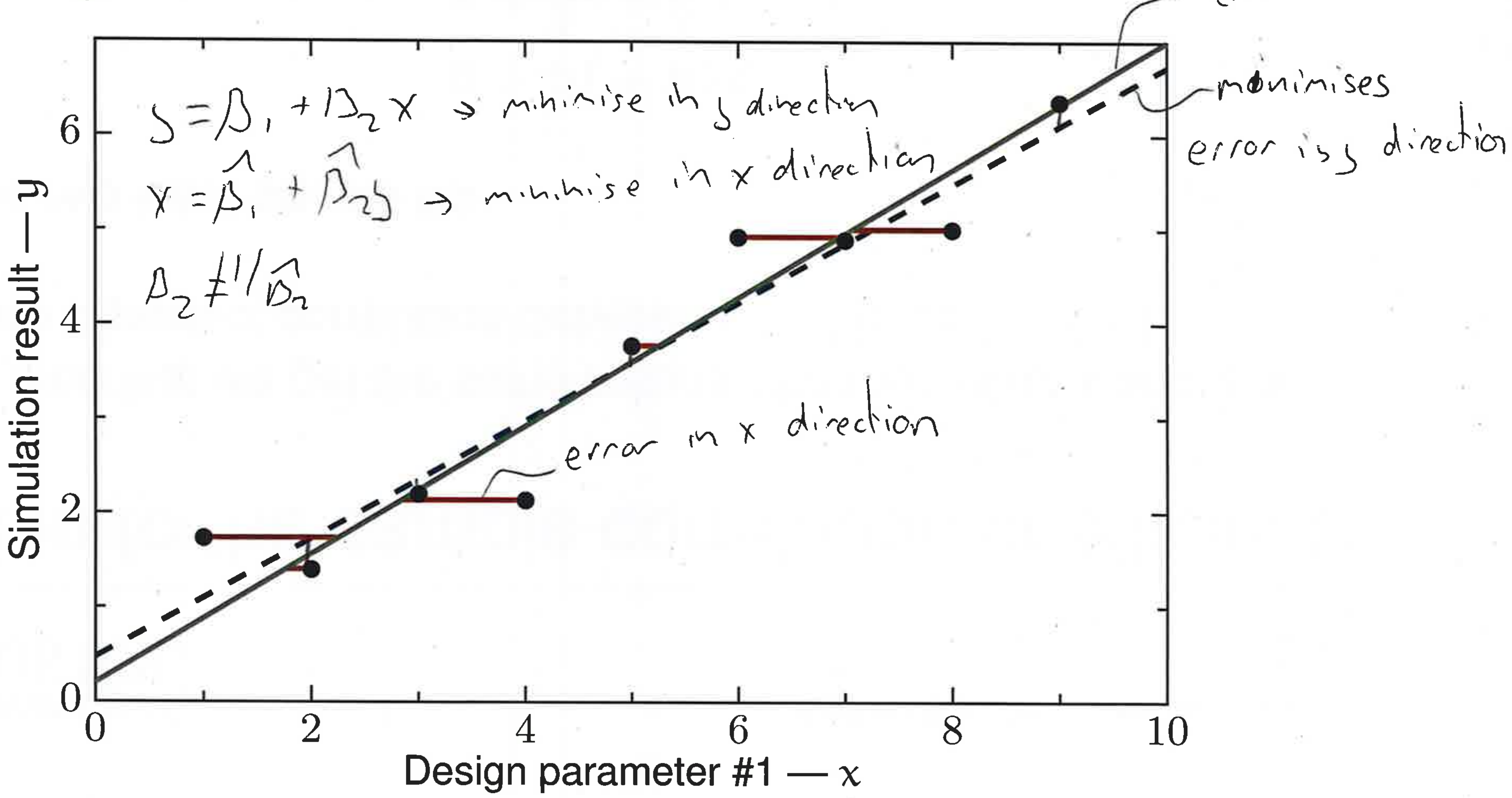
where y_i are i.i.d. samples and p_Y is the corresponding probability density function. (Often maximise the log of this function.)

Maximum likelihood



Errors in the x direction

We've assumed no errors in the x direction — swapping the x and y coordinates gives different answers!



Relation to the sample correlation coefficient

The only time that we get the same results when we swap x and y is when there is perfect correlation between x and y; i.e., $r = \pm 1$.

If we have two least-square fits

$$y = \beta_1 + \beta_2 x$$
$$x = \hat{\beta}_1 + \hat{\beta}_2 y$$

then it's possible to show that

$$\beta_2 \cdot \hat{\beta}_2 = r^2$$

Hence $\beta_2 = 1/\hat{\beta}_2$ only when $r = \pm 1$.

Exercise

Calculate line of best fit

$x =$ death by entanglement

$y =$ cheese consumption

$$y = \beta_1 + \beta_2 x$$

normal equations

$$n\beta_1 + \sum x_i \beta_2 = \sum y_i$$

$$\sum x_i \beta_1 + \sum x_i^2 \beta_2 = \sum x_i y_i$$

$$10\beta_1 + 5886\beta_2 = 315.2$$

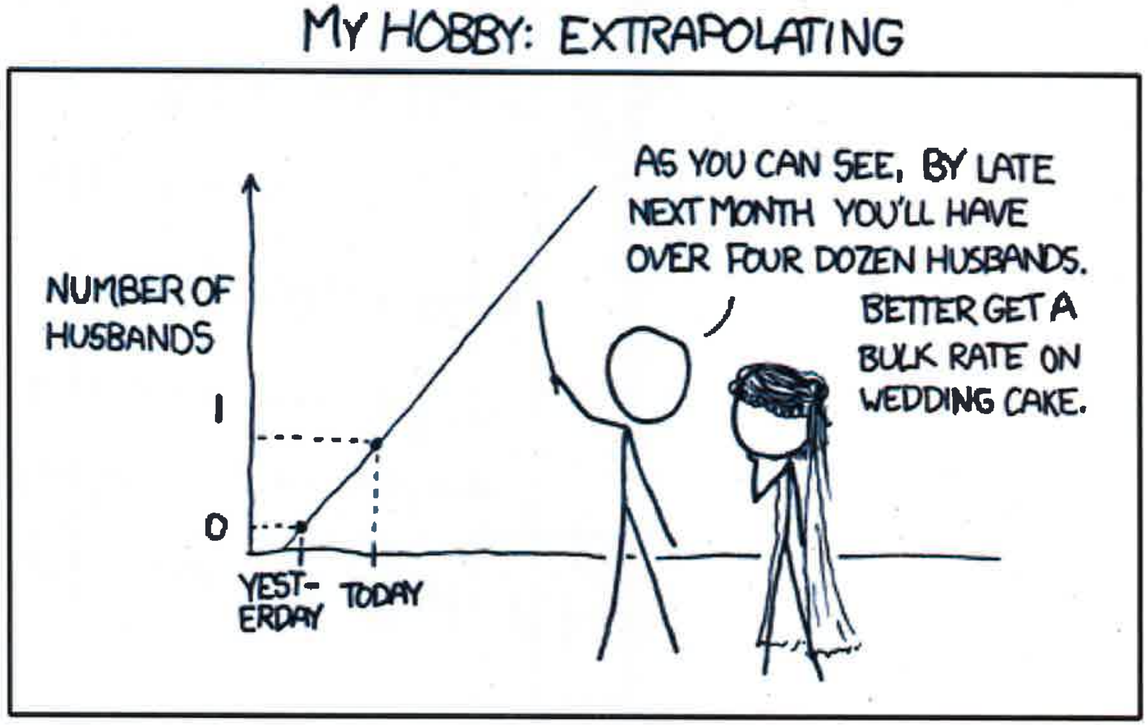
$$5896\beta_1 + 3659072\beta_2 = 187069.4$$

solve

$$\boxed{y = \underset{\beta_1}{26.85} + \underset{\beta_2}{0.008}x}$$

model

Be careful with data fitting. . .

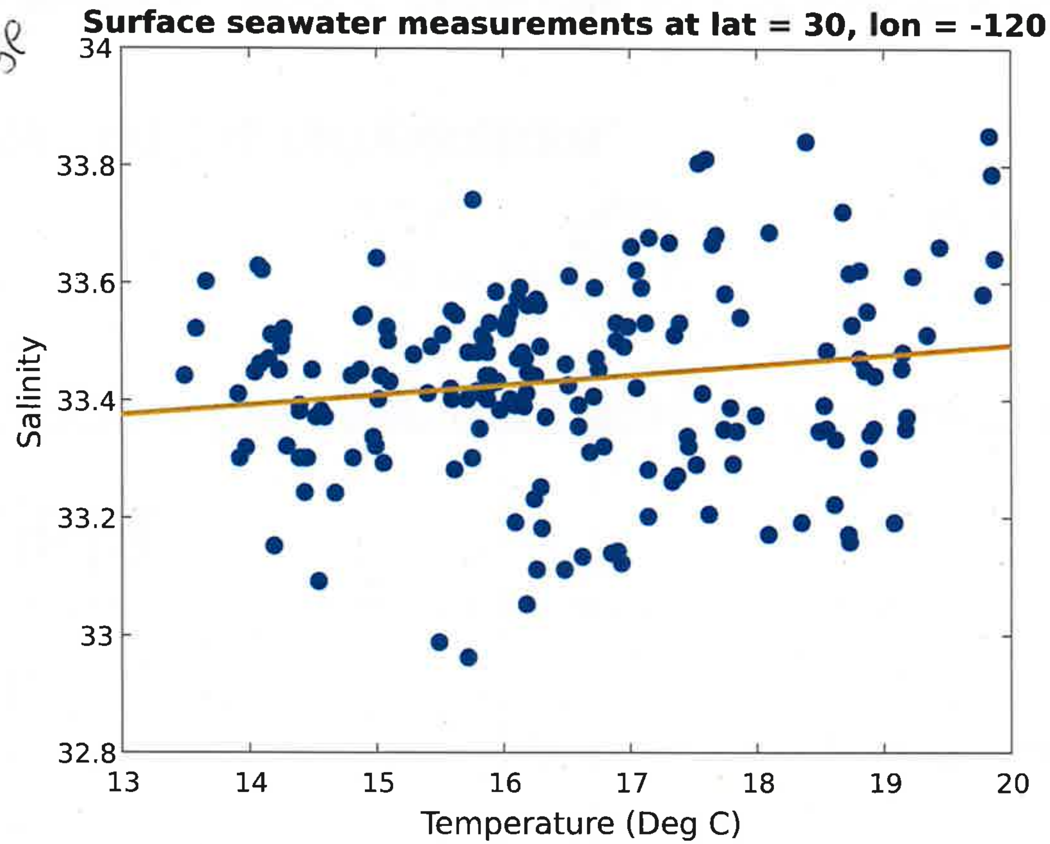


[XKCD #605]

CalCOFI data

California Cooperative Oceanic Fisheries Investigations (CalCOFI) dataset is a database of oceanic measurements taken regularly since 1949. Amongst other aims, it collects data related to climate change.

affects circulation of water, important for climate change



CalCOFI data

Can fit a linear function between temperature and salinity such that

$$\underbrace{s}_{\text{salinity}} = m\underset{\uparrow \text{temp}}{t} + c \qquad e_i = \underbrace{s_i}_{\text{actual value}} \overbrace{- mt_i - c}^{\text{prediction}}$$

where s is salinity and t is temperature.

minimise $\sum_i \|e\|^2$ wrt m & c

Finding the constants m and c amounts to solving the normal equations

$$\begin{bmatrix} n & \sum_i t_i \\ \sum_i t_i & \sum_i t_i^2 \end{bmatrix} \begin{bmatrix} c \\ m \end{bmatrix} = \begin{bmatrix} \sum_i s_i \\ \sum_i s_i t_i \end{bmatrix}$$

$$\frac{d\|e\|^2}{dc} = \sum_i -2(s_i - mt_i - c) = 0$$

$$\frac{d\|e\|^2}{dm} = \sum_i -2t(s_i - mt_i - c) = 0$$

Normal equations

$$\underset{\#\text{samples}}{nc} + m\sum_i t_i = \sum_i s_i$$

$$c\sum_i t_i + m\sum_i t_i^2 = \sum_i s_i t_i$$

In this case

$$\begin{bmatrix} 185 & 3048.1 \\ 3048.1 & 50676.9 \end{bmatrix} \begin{bmatrix} c \\ m \end{bmatrix} = \begin{bmatrix} 6184.4 \\ 101899.6 \end{bmatrix}$$

giving $c = 33.205$ and $m = 0.0136$ $\quad s = 0.0136t + 33.205$

Making inferences from the data

Might want to make inferences, e.g., what is the probability that the salinity will be over 33.5 given a temperature of 16 °C?

Under what conditions can we use linear regression to answer questions like this?

need this to hold to do

Three conditions to check —

- Normal residuals — errors or normally distributed
- Constant variability (homoscedasticity) — variability in the errors doesn't depend on the independent variable
  e.g. error in variables doesn't ~~depende~~ depend on the temperature
- Independence of observations
  (data points are independent)

Normal residuals

The residuals are

actual, prediction

$$e_i = y_i - \sum_j \beta_j x_{i,j}, \quad \text{e.g.,} \quad e_i = y_i - \beta_1 - \beta_2 x_i$$

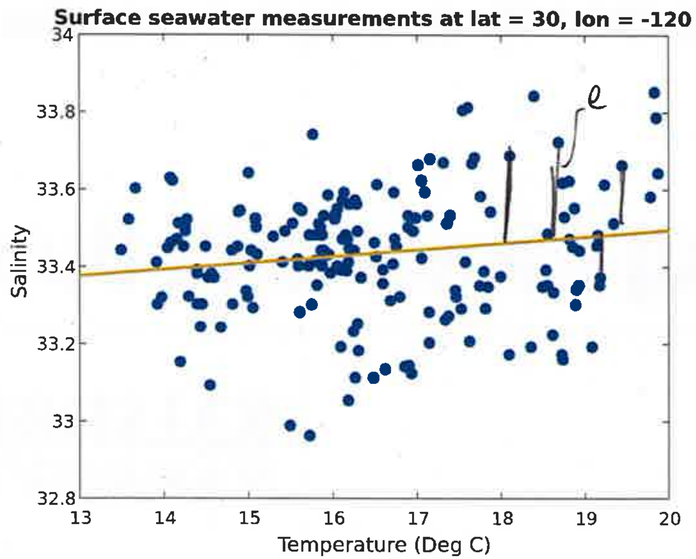
$s = \beta_1 + \beta_2 x$

$e_i = s_i - m t_i - c$

→ residual for every observation

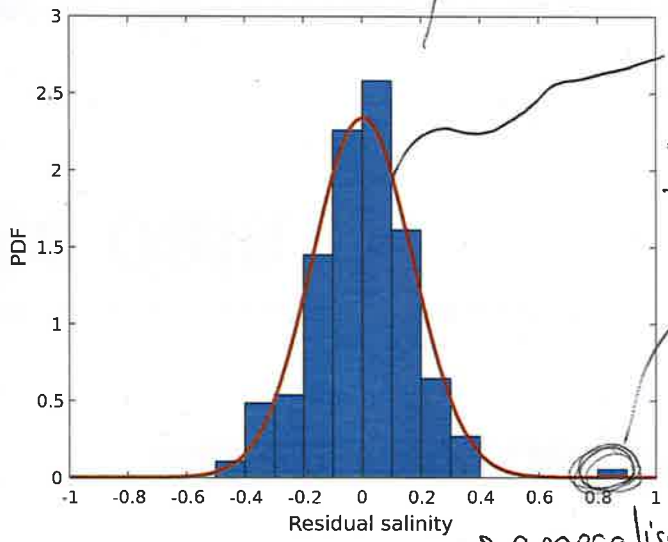
i.e., the difference between the line of best fit and the actual observation.

Plot a histogram of the residuals.



distribution of residuals

residual plot



pretty good (close to normal distribution

normal distribution with the same mean and std as the observed residuals

outlier

impact on least squares

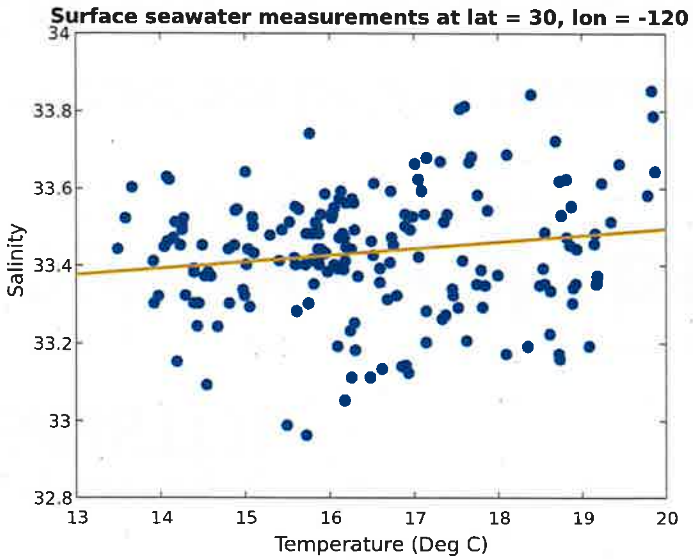
→ filter your data

→ generalised linear model GLM

Constant variability - homoscedasticity

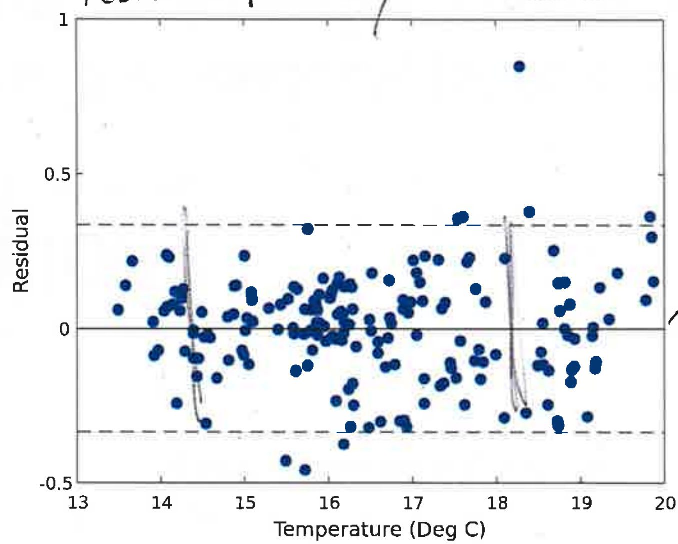
There should not be any obvious trends in the residuals (should be normal random variables) and variability (e.g., standard deviation) should not change as the independent variable does.

Plot the residuals against the independent variable — this is the most common sort of residual plot.



residual plot

as temperature increases variability is ~~reducing~~ should stay ~~the~~ the same



not good

constant variability

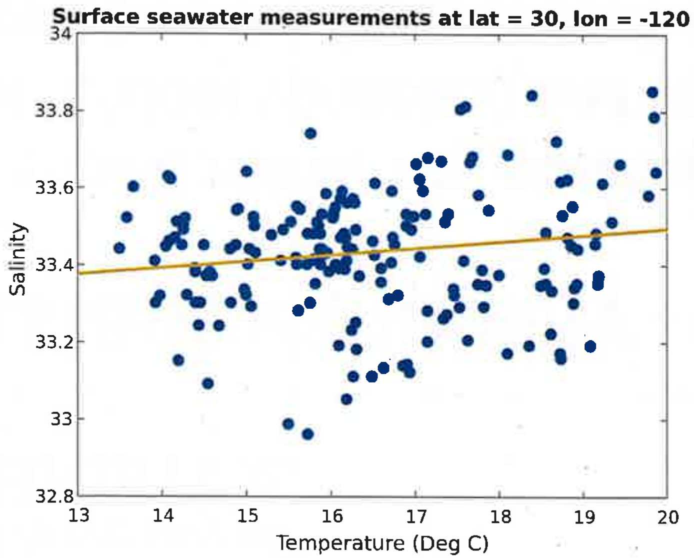
generalised least square

Independence of observations

L challenging when you have time series

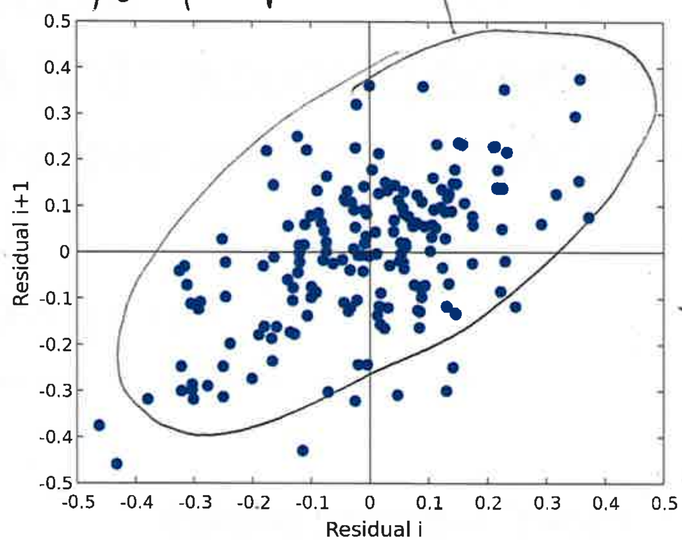
There should not be any obvious trends in the residuals (should be normal random variables) and the value of each residual should not depend on another.

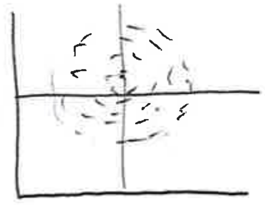
Plot the i-th residual against the i + 1-residual.



residual plot

not independent



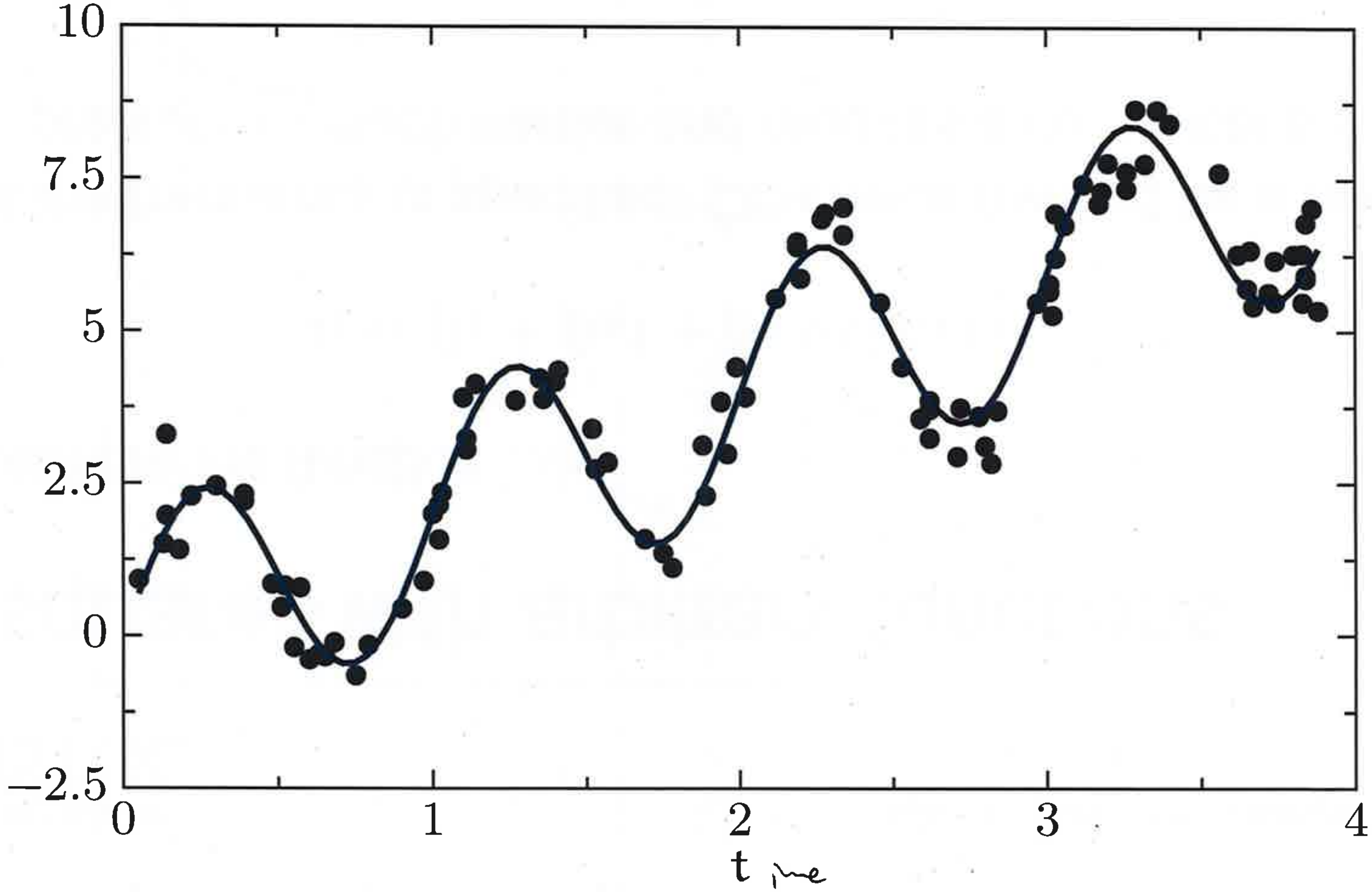


linear regression

→ AR
autoregressive model

Least squares with arbitrary functions

Often data contains non-polynomial trends that we want to investigate



Least squares with arbitrary functions

Take for example the function

$$y = \overbrace{\beta_1 + \beta_2 t}^{\text{linear trend}} + \overbrace{\beta_3 \sin(2\pi t)}^{\text{seasonal trend}}$$

Notice that the frequency is specified! Otherwise it would be a nonlinear regression problem — much harder and requires a numerical solution.

residual

Define the error as

$$\|e\|^2 = \sum_{i=1}^{n} (\overbrace{y_i}^{\text{actual}} - \overbrace{\beta_1 - \beta_2 t_i - \beta_3 \sin(2\pi t_i)}^{\text{prediction}})^2$$

and derive the normal equations in the usual way. (Find derivatives, set them equal to zero...)

Least squares with arbitrary functions

The normal equations in this case are

$$\beta_1 n + \beta_2 \sum_{i=1}^{n} t_i + \beta_3 \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} y_i$$ derivative of $\|e\|^2$ wrt β_1 $=0$

$$\beta_1 \sum_{i=1}^{n} t_i + \beta_2 \sum_{i=1}^{n} t_i^2 + \beta_3 \sum_{i=1}^{n} t_i S_i = \sum_{i=1}^{n} t_i y_i$$ derivative of $\|e\|^2$ wrt β_2

$$\beta_1 \sum_{i=1}^{n} S_i + \beta_2 \sum_{i=1}^{n} t_i S_i + \beta_3 \sum_{i=1}^{n} S_i^2 = \sum_{i=1}^{n} S_i y_i$$ 〃 β_3

where $S_i = \sin(2\pi t_i)$. (Notice the symmetry again!)

Again in the form $A\beta = b$.

$$A = \begin{bmatrix} n & \sum_i t_i & \sum_i s_i \\ \sum_i t_i & \sum_i t_i^2 & \sum_i t_i s_i \\ \sum_i s_i & \sum_i t_i s_i & \sum_i s_i^2 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \quad b = \begin{bmatrix} \sum_i y_i \\ \sum_i t_i y_i \\ \sum_i s_i y_i \end{bmatrix}$$

General approach

Deriving the normal equations each time is straightforward but tedious. Is there a general equation? Yes!

General linear regression uses

$$y_i = \beta_1 x_{i,1}, + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}$$

the $x_{i,j}$ terms can be whatever we want! In the last example $p = 3$ and

$$\begin{aligned} x_{i,1} &= 1 \\ x_{i,2} &= t_i \\ x_{i,3} &= \sin(2\pi t_i) \end{aligned}$$

to give

$$y_i = \beta_1 + \beta_2 t_i + \beta_3 \sin(2\pi t_i)$$

General approach

Hence writing this in matrix-vector form gives

$$y = X\beta$$

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & t_1 & \sin(2\pi t_1) \\ \vdots & \vdots & \vdots \\ 1 & t_n & \sin(2\pi t_n) \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

The corresponding error vector is given by

prediction　　vector

$$\|e\|^2 = (y - X\beta)^T (y - X\beta)$$

vector
= scalar

actual values

General approach

Expanding out the error vector and differentiating gives the *general normal equations*

$$\overbrace{(X^T X)}^{A}\underline{\beta} = \overbrace{X^T y}^{b}$$

i.e., the same $A\beta = b$ form as before, and so

$$\beta = \overbrace{(X^T X)^{-1}}^{A^{-1}} X^T y$$

In Matlab this is achieved very simply with the command

```
beta = X \ y;
```

For example with $y = \beta_1 + \beta_2 t + \beta_3 \sin(2\pi t)$ with y and t as column vectors we have

```
beta = [ones(size(t)), t, sin(2*pi*t)] \ y;
```

Exercise

A particular process follows a daily cycle which suggests a least-squares fit using

$$y = \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t)$$

where t is measured in days, with the data

t	0.00	0.24	0.48	0.72	0.95
y	1.15	0.55	−1.12	−1.04	0.97

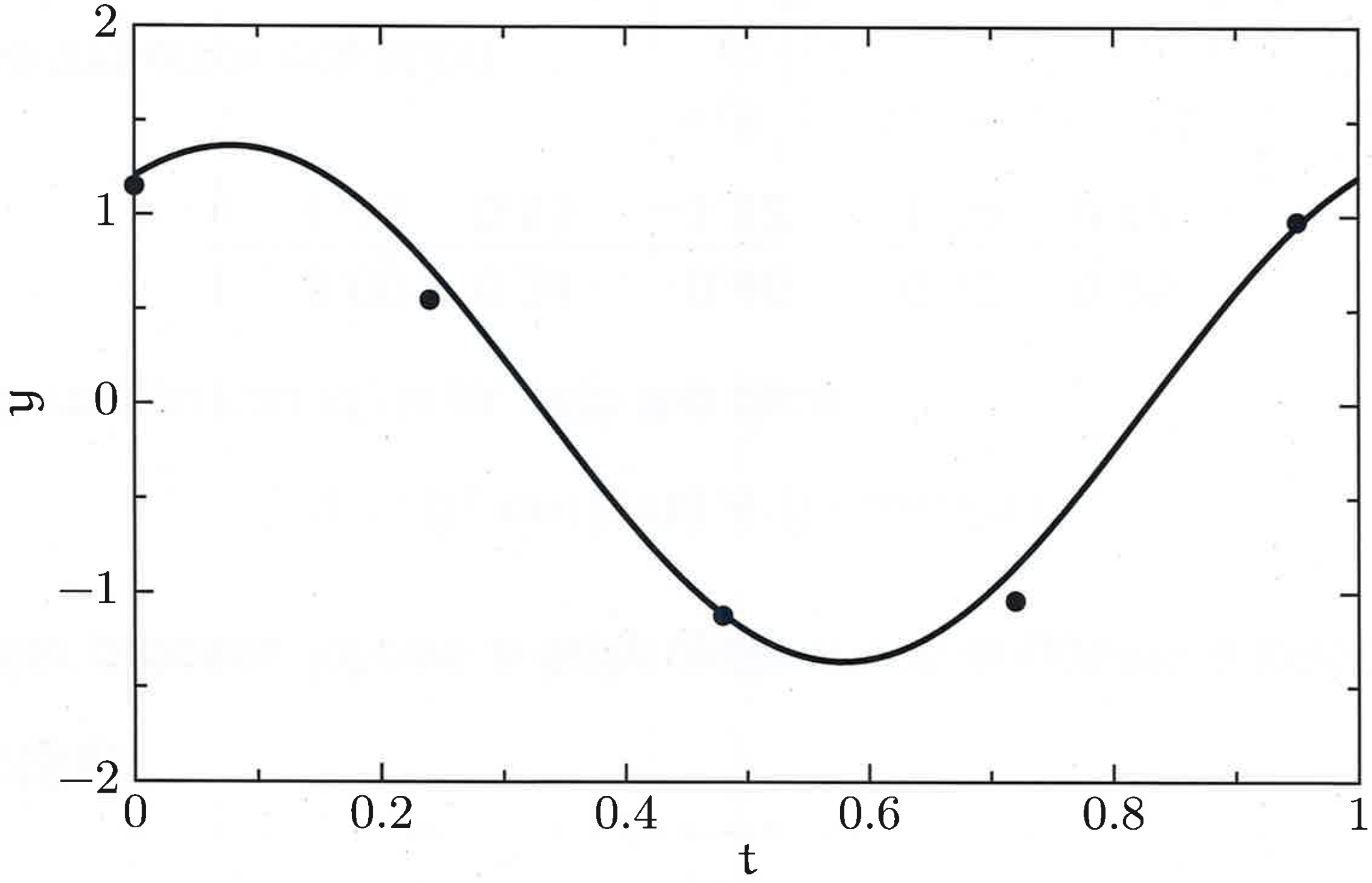
$\|e\|^2 = \sum (y_i - \beta_i \sin(2\pi t_i) - \beta_2 \cos(2\pi t_i))^2$

$X = \begin{bmatrix} \sin(2\pi t_1) & \cos(2\pi t_1) \\ \sin(2\pi t_2) & \cos(2\pi t_2) \end{bmatrix}$ $\quad \|e\|^2 = (y - X\beta)^T(y - X\beta)$

1. State the error equation
2. Derive the normal equations by differentiating w.r.t. β_1 and β_2
3. Calculate the required quantities
4. Solve the normal equations for β_1 and β_2

(Or use the general equation but that's probably harder by hand. . .)

Answers



In this case, least-squares is a good alternative to an FFT (small number of data points and sampling freq is incommensurate with the period).

† Uncertainty in least-squares estimates

When calculating least squares estimates for β of the form

$$y = X\beta$$

it is important to remember that β is a random variable with its own distribution.

The mean of the distribution is the value of β calculated, but what is the variance? Multiple linked variables means we have to compute the *covariance matrix*.

With this information, we can say how confident we are about any estimates.

† Uncertainty in least-squares estimates

The general solution for the least-squares estimator gives the estimated value for β

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

From this expression it's possible to show that

$$\text{covar}(\hat{\beta}) = (X^T X)^{-1} \sigma^2$$

where σ^2 is the variance of the noise.

Can estimate the variance of the noise from the data

$$\hat{\sigma}^2 = \frac{1}{n-p} \|y - X\beta\|^2 = \frac{1}{n-p} \|e\|^2$$

where p is the number of parameters to be estimated.

† Uncertainty in least-squares estimates

For the previous exercise with

$$y = \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t)$$

we have

$$X = \begin{bmatrix} 0 & 1.0000 \\ 0.9980 & 0.0628 \\ 0.1253 & -0.9921 \\ -0.9823 & -0.1874 \\ -0.3090 & 0.9511 \end{bmatrix}$$

$$\hat{\sigma}^2 = \frac{1}{n-p} \|y - X\beta\|^2 = 0.0216$$

Hence

$$\mathrm{covar}(\beta) = \begin{bmatrix} 0.0105 & 0.0006 \\ 0.0006 & 0.0074 \end{bmatrix}$$

† Exercise

Given the covariance matrix for β (from the previous exercise)

$$\mathrm{covar}(\beta) = \begin{bmatrix} 0.0105 & 0.0006 \\ 0.0006 & 0.0074 \end{bmatrix}$$

and that $\beta_1 = 0.6451$, what is the range of maximum value that β_1 could take and still pass a hypothesis test to 5% significance?

Ignore the covariance in this case — just use the first element of the matrix as the variance of β_1.

A two-tailed t test with 3 degrees of freedom at 5% significance gives a critical value of 3.182.