Norges miljø- og biovitenskapelige universitet

Institutt for datavitenskap

Digitalisering på Ås

DAT121
Introduction to data science

3 Regression basics

3.4 Influence diagrams
3.5 Residual quantities
3.6 Time series

What is our uncertainty in regression?

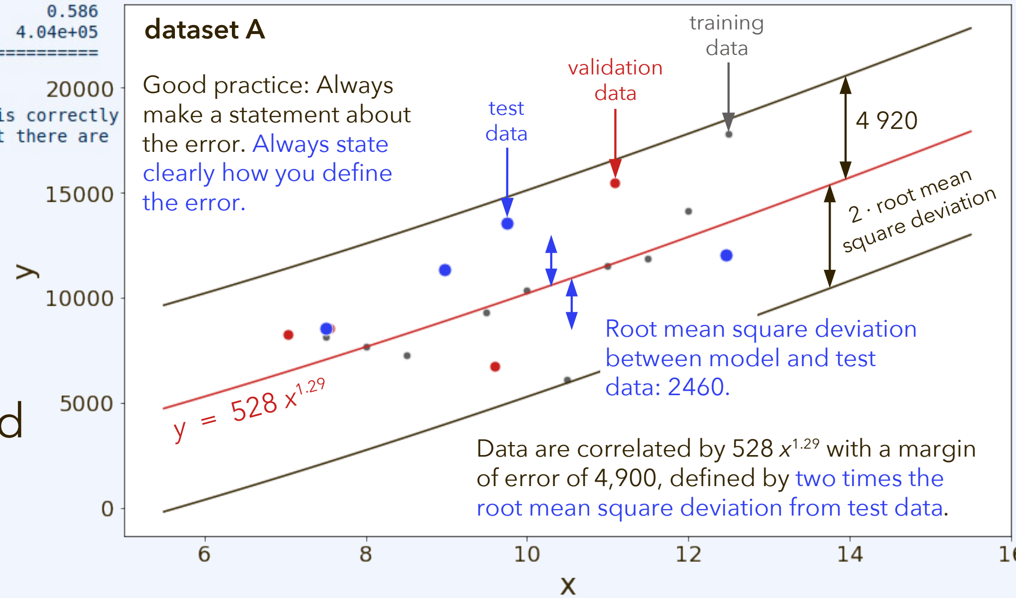
Norwegian University of Life Sciences

```
Fit a):
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.799
Model:                            OLS   Adj. R-squared:                  0.713
Method:                 Least Squares   F-statistic:                     9.265
Date:                Mon, 06 Dec 2021   Prob (F-statistic):            0.00781
Time:                        10:55:39   Log-Likelihood:                -95.687
No. Observations:                  11   AIC:                             199.4
Df Residuals:                       7   BIC:                             201.0
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -2.342e+05   1.79e+05     -1.311      0.231   -6.56e+05    1.88e+05
x1          7.821e+04   5.49e+04      1.424      0.198   -5.17e+04    2.08e+05
x2         -8367.3711   5558.872     -1.505      0.176   -2.15e+04    4777.273
x3           297.8664    185.111      1.609      0.152    -139.851     735.584
==============================================================================
Omnibus:                        2.762   Durbin-Watson:                   2.877
Prob(Omnibus):                  0.251   Jarque-Bera (JB):                1.069
Skew:                          -0.761   Prob(JB):                        0.586
Kurtosis:                       3.123   Cond. No.                     4.04e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
[2] The condition number is large, 4.04e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

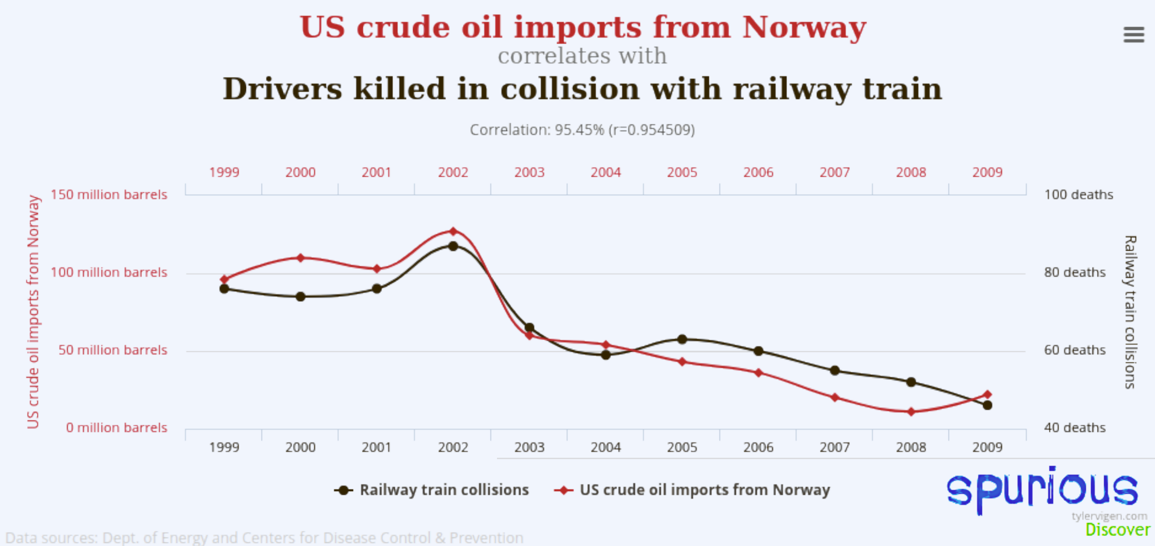
There was the uncertainty with which we can give the **coefficients of the regression**.

There was the deviation that new **single data points** would be expected to have from the regression.



Why does this feel wrong to us?

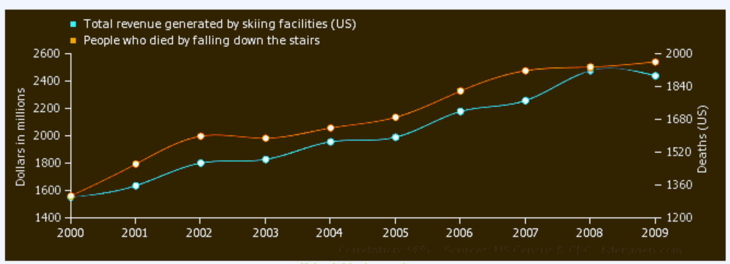
Norwegian University
of Life Sciences



spurious correlations
tylervigen.com
Discover a new correlation

Is there anything numerically wrong with the correlations?

Total revenue generated by skiing facilities (US)
correlates with
People who died by falling down the stairs

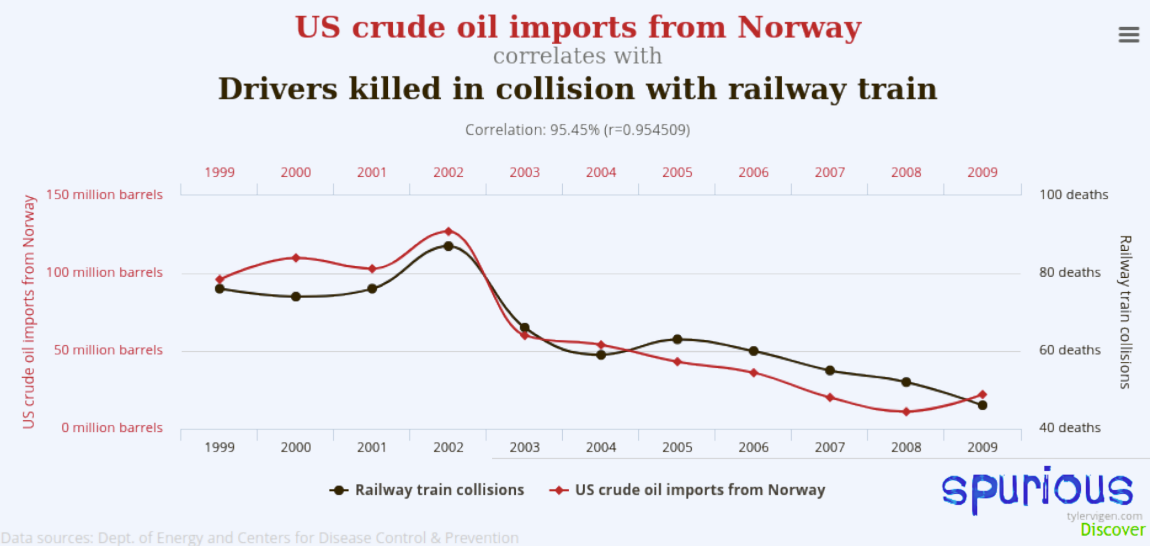


Upload this chart to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Total revenue generated by skiing facilities (US) *Dollars in millions (US Census)*	1,551	1,635	1,801	1,827	1,956	1,989	2,178	2,257	2,476	2,438

Why does this feel wrong to us?

Norwegian University of Life Sciences



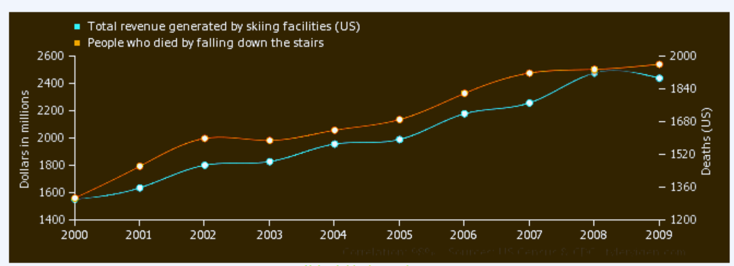
spurious correlations
tylervigen.com
Discover a new correlation

Or is it that "correlation does not imply causation"?

But the figures don't claim anything about causation.

Is there anything numerically wrong with the correlations?

Total revenue generated by skiing facilities (US)
correlates with
People who died by falling down the stairs



Upload this chart to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Total revenue generated by skiing facilities (US) *Dollars in millions (US Census)*	1,551	1,635	1,801	1,827	1,956	1,989	2,178	2,257	2,476	2,438

Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap

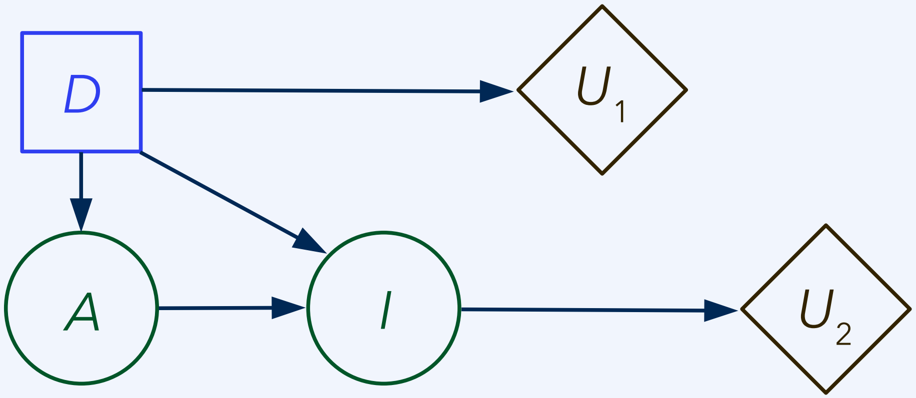
Digitalisering på Ås

3 Python basics

3.4 Influence diagrams

Influence diagrams

Influence diagrams (also: decision networks) visualize how different quantities are connected to each other in a decision-making process.



(Example based on Barber,[1] Fig. 7.6)

D: Should I work on a doctorate?

A: Academic recognition measure
I: Life income

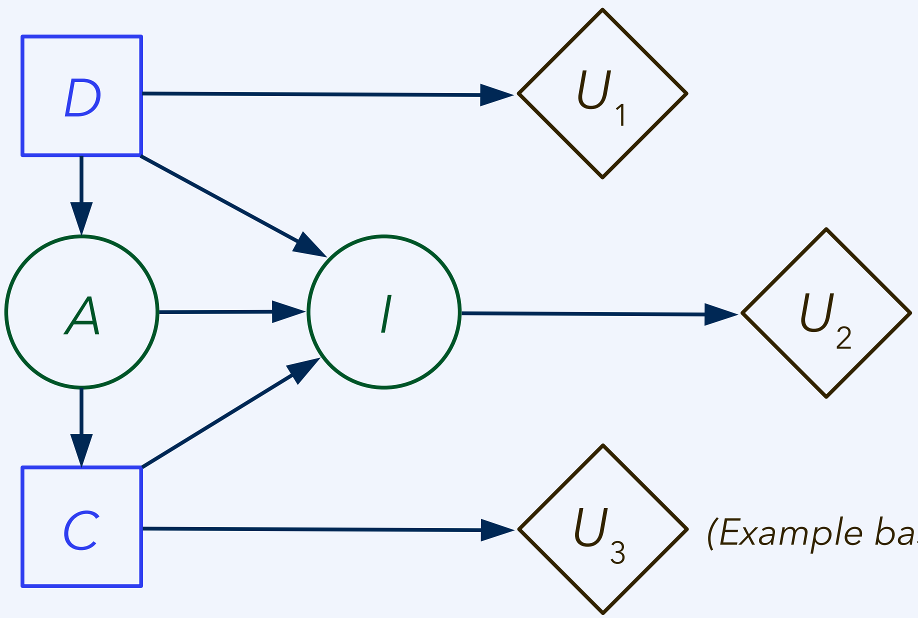
U_1, U_2: Contributions to utility.

[1] D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge Univ. Press, **2012**.

influence diagram

Influence diagrams

Influence diagrams (also: decision networks) visualize how different quantities are connected to each other in a decision-making process.



D: Should I work on a doctorate?
C: Should I found a consultancy?

A: Academic recognition measure
I: Life income

U_1, U_2, U_3: Contributions to utility.

(Example based on Barber,[1] Fig. 7.6)

Observation: An **influence diagram visualizes a process** by which quantities are evaluated. For it to represent a valid process, it **must not contain cycles**.

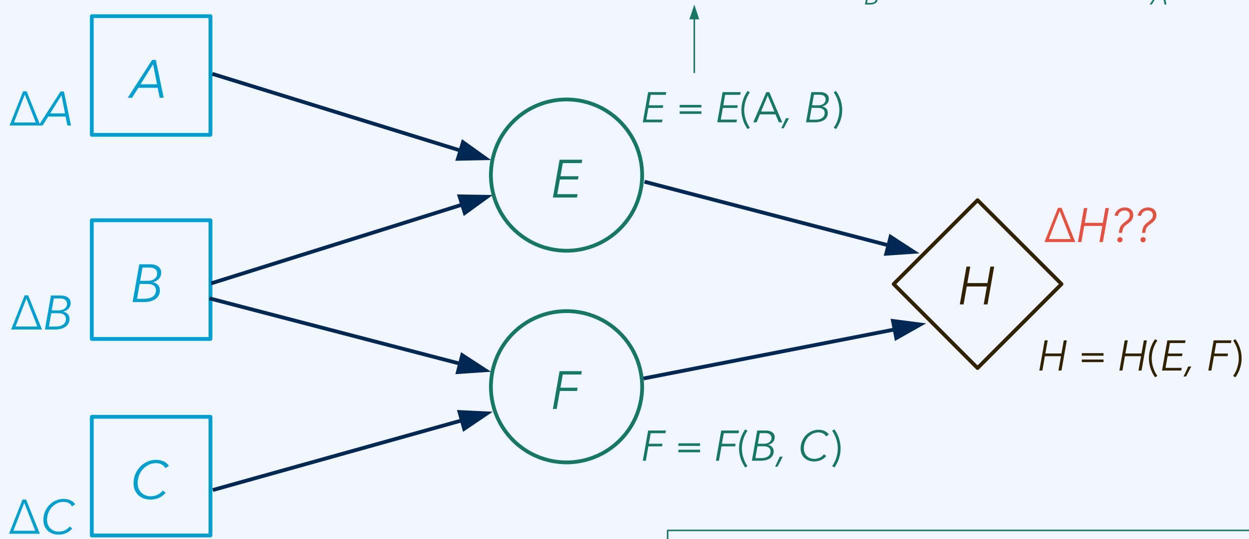
[1] D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge Univ. Press, **2012**.

Linear uncertainty propagation

Norwegian University of Life Sciences

Idea: Treat the **propagated uncertainty** as analogous to a **total differential**.

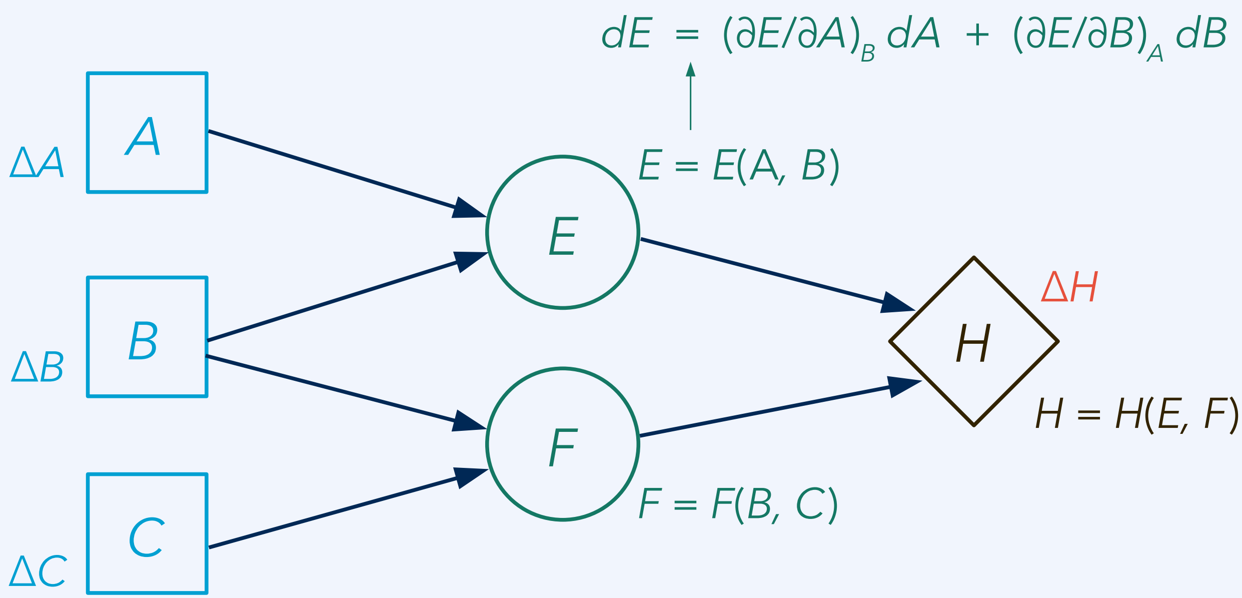
$$dE = (\partial E/\partial A)_B \, dA + (\partial E/\partial B)_A \, dB$$



$$\boxed{\Delta E = |(\partial E/\partial A)_B| \, \Delta A + |(\partial E/\partial B)_A| \, \Delta B}$$

Linear uncertainty propagation

Idea: Treat the propagated uncertainty as analogous to a total differential.



$$\Delta E \;=\; |(\partial E/\partial A)_B|\,\Delta A \;+\; |(\partial E/\partial B)_A|\,\Delta B$$

$$\Delta F \;=\; |(\partial F/\partial B)_C|\,\Delta B \;+\; |(\partial F/\partial C)_B|\,\Delta C$$

$$\Delta H \;=\; |(\partial H/\partial E)_F|\,\Delta E \;+\; |(\partial H/\partial F)_E|\,\Delta F$$

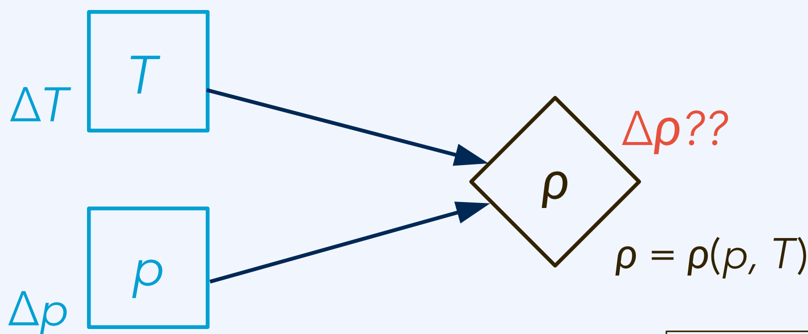
Advantages:

- Simple calculus that can be applied to complicated influence diagrams.
- No need to resolve backward, *e.g.*, we don't need to find $H(A, B, C)$.

Linear uncertainty propagation: “Nice” example

We are given information on the thermodynamic state of pure liquid water:

- Temperature given as $T = 280\text{ K} \pm 3\text{ K}$
- Pressure given as $p = 1\text{ MPa} \pm 0.1\text{ MPa}$



$$\boxed{\Delta\rho = |(\partial\rho/\partial p)_T|\,\Delta p + |(\partial\rho/\partial T)_p|\,\Delta T}$$

What is the uncertainty in the density of the water?

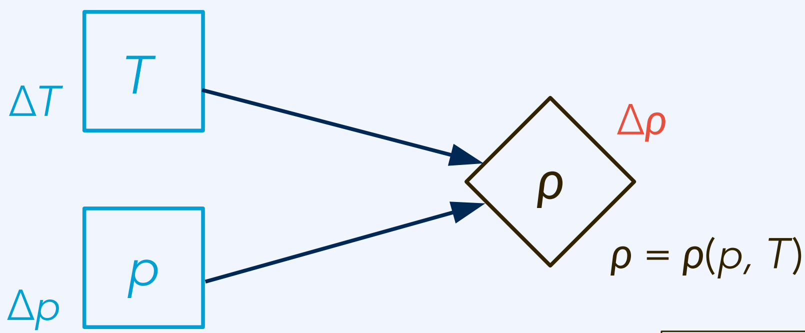
$\rho = \rho(p, T) \pm \Delta\rho(p, \Delta p, T, \Delta T)$

https://webbook.nist.gov/chemistry/fluid/

Linear uncertainty propagation: “Nice” example

We are given information on the thermodynamic state of pure liquid water:

- Temperature given as $T = 280\ \mathrm{K} \pm 3\ \mathrm{K}$
- Pressure given as $p = 1\ \mathrm{MPa} \pm 0.1\ \mathrm{MPa}$



$$\boxed{\Delta\rho = |(\partial\rho/\partial p)_T|\,\Delta p + |(\partial\rho/\partial T)_p|\,\Delta T}$$

What is the uncertainty in the density of the water?

$\rho = 55.521\ \mathrm{mol\ l^{-1}} \pm \Delta\rho$ $\quad (\partial\rho/\partial p)_T = 0.0275\ \mathrm{mol\ l^{-1}\ MPa^{-1}}$ $\quad (\partial\rho/\partial T)_p = -0.0025\ \mathrm{mol\ l^{-1}\ K^{-1}}$

https://webbook.nist.gov/chemistry/fluid/

Limitations of the method

Norwegian University of Life Sciences

Influence diagrams supply the inductive reasoning with **helpful domain knowledge**. But they are limited to saying what quantities depend on what other quantities – they do not directly describe *how*, even if we know it.

There are also **limitations** to the method of **linear uncertainty propagation**. In practice these can cause a major risk of making potentially serious mistakes:

- The approximation as *linear* may not be warranted, which can make the *uncertainty appear smaller* than it actually is.
  Characteristic example: Potential energy of a harmonic oscillator.

- When there are diamonds in the diagram, relying on intermediate variables can make the *uncertainty appear greater* than it actually is.
  Typical example: Subtraction of a large number from a similar one.

- Non-trivial interaction between variables. (Not rare in decision making.)

Limitations of the method

Influence diagrams supply the inductive reasoning with helpful domain knowledge. But they are limited to saying what quantities depend on what other quantities – they do not directly describe *how*, even if we know it.

There are also limitations to the method of linear uncertainty propagation.
In practice these can cause a major risk of making potentially serious mistakes:

- The approximation as *linear* may not be warranted, which can make the *uncertainty appear smaller* than it actually is.
  Characteristic example: Potential energy of a harmonic oscillator.
  Remedy: Keep higher-order terms; only neglect them if you know it is OK!
- When there are diamonds in the diagram, relying on intermediate variables can make the *uncertainty appear greater* than it actually is.
  Typical example: Subtraction of a large number from a similar one.
  Remedies: a) Accept; b) avoid diamonds; c) expand in original variables.
- Non-trivial interaction between variables. (Not rare in decision making.)
  Remedies: Multicriteria and sampling methods (*e.g.*, Monte Carlo).

Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap

Digitalisering på Ås

3 Python basics

3.4 Influence diagrams

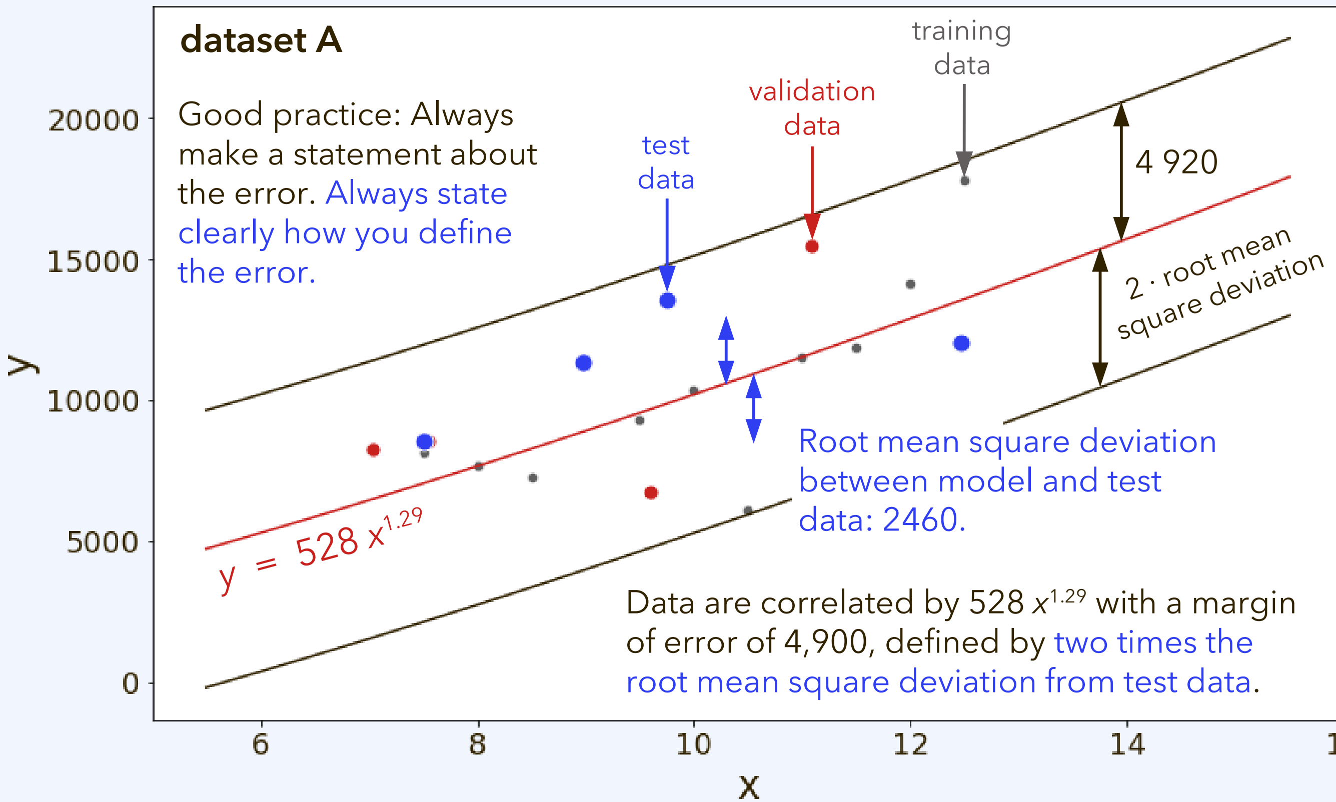
3.5 Residual quantities

Residual with respect to a model

uncertainty

Other ways of using a residual w.r.t. to a model, *e.g.*, a model from regression:

- We can express the **level of confidence or uncertainty** for a **model** from the **magnitude of the residual**, as we implicitly did in the last lecture:

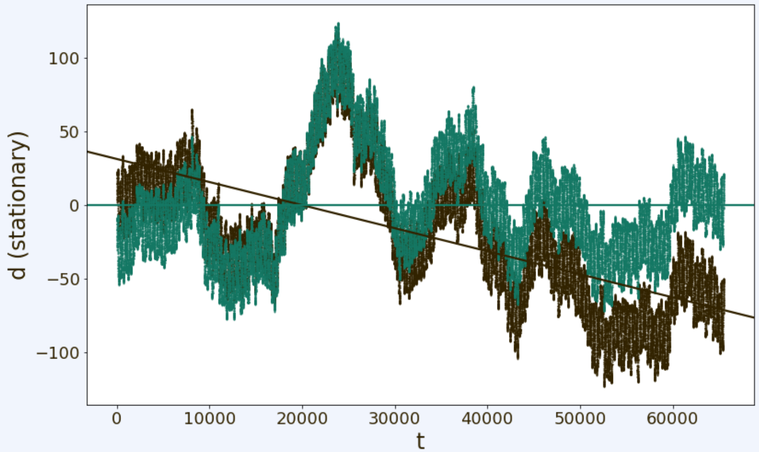


Residual with respect to a model

residual quantity

Other ways of using a residual w.r.t. to a model, *e.g.*, a model from regression:

- We can express the **level of confidence or uncertainty** for a **model** from the **magnitude of the residual**, as we implicitly did in the last lecture.
- We can establish a **hierarchy of models**, with each new model approximating that what remains as a residual after the previous ones.
- Remove some **undesired behaviour** or bias **before further analysing** data.
  - Always document such steps very clearly, or you are engaging in fraud!



Black: Data with a tendency downward.

Green: Residual w.r.t. the linear regression.

(Example will be used again later. See the autocorrelation-statsmodels.ipynb notebook.)

Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap

Digitalisering på Ås

3 Python basics

Time series in Python

Methods for time series from pandas are summarized in the Python for Data Analysis book, Chapter 11 (https://wesmckinney.com/book/time-series):

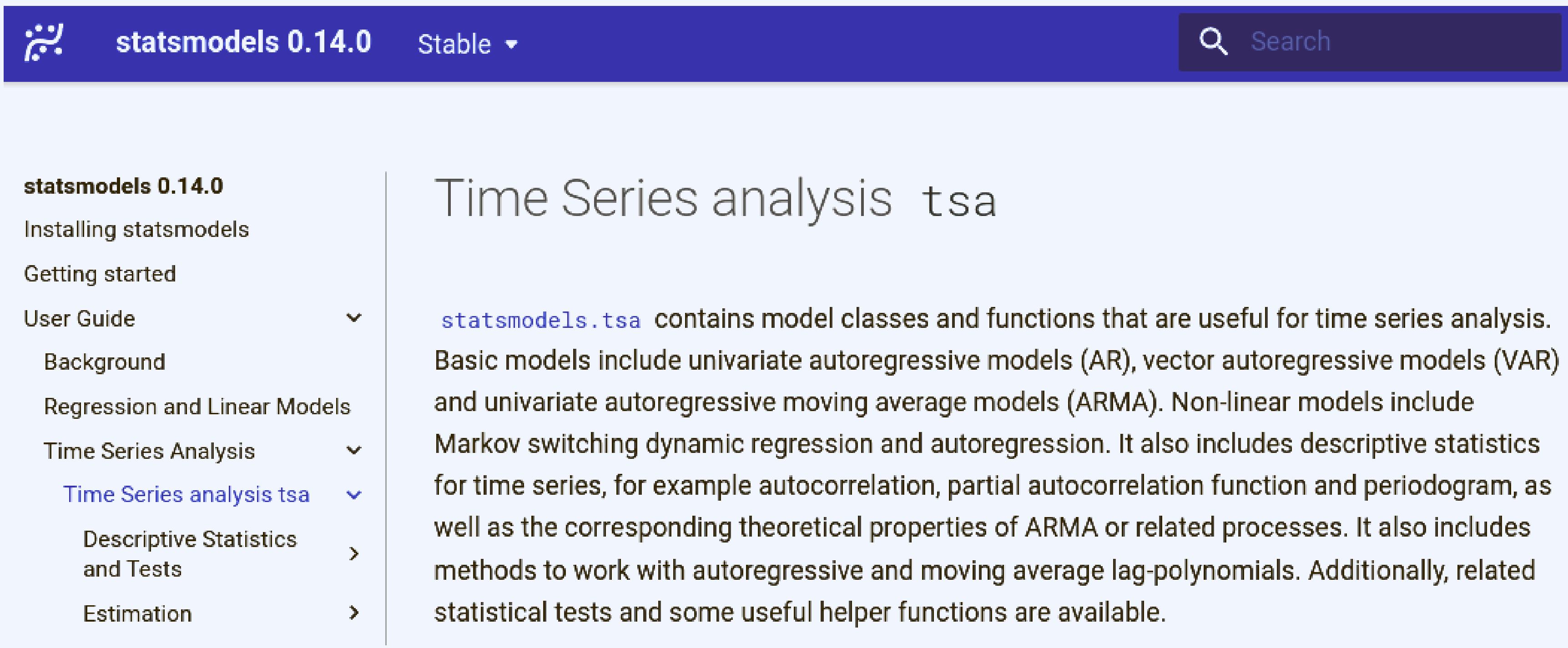
```
In [254]: close_px["AAPL"].rolling(250).mean().plot()
```

Figure 11.4: Apple price with 250-day moving average

Time series in Python

https://www.statsmodels.org/stable/tsa.html

https://www.statsmodels.org/stable/examples/index.html#time-series-analysis



statsmodels 0.14.0 Stable ▾ Search

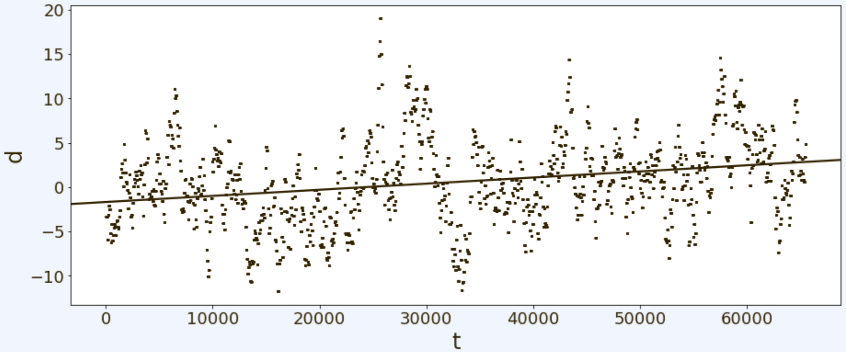
statsmodels 0.14.0
Installing statsmodels
Getting started
User Guide
Background
Regression and Linear Models
Time Series Analysis
Time Series analysis tsa
Descriptive Statistics and Tests
Estimation

Time Series analysis `tsa`

`statsmodels.tsa` contains model classes and functions that are useful for time series analysis. Basic models include univariate autoregressive models (AR), vector autoregressive models (VAR) and univariate autoregressive moving average models (ARMA). Non-linear models include Markov switching dynamic regression and autoregression. It also includes descriptive statistics for time series, for example autocorrelation, partial autocorrelation function and periodogram, as well as the corresponding theoretical properties of ARMA or related processes. It also includes methods to work with autoregressive and moving average lag-polynomials. Additionally, related statistical tests and some useful helper functions are available.

Autocorrelation of time series data

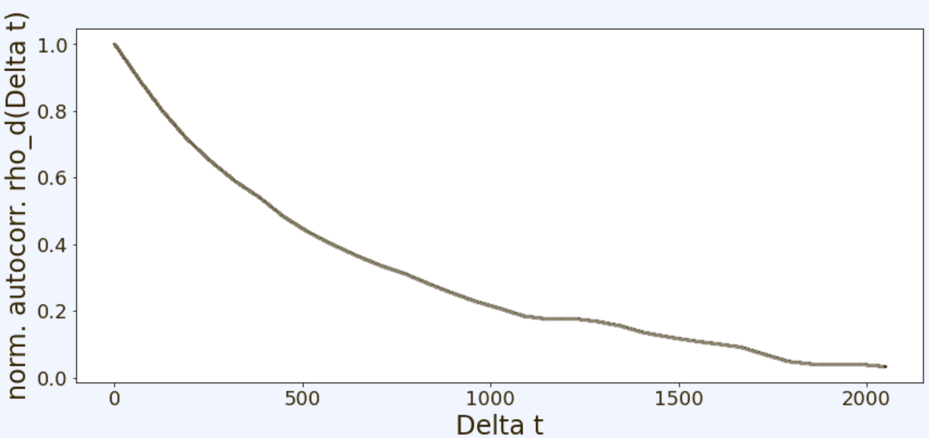
Time series data are **autocorrelated**. This means that *data points taken at times close to each other cannot be regarded as independent* items of information.



Assume we are given time series data $d(t)$:

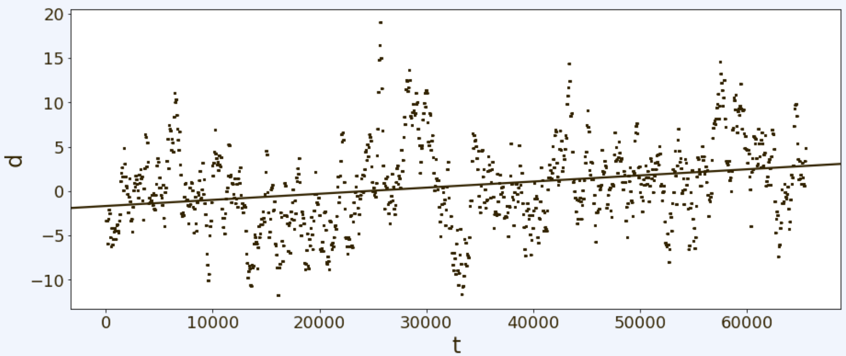
autocorrelation $R(\Delta t) = \langle d(t)\, d(t + \Delta t)\rangle$
autocovariance $\langle [d(t) - \langle d\rangle]\, [d(t + \Delta t) - \langle d\rangle]\rangle$

Often **normalized** by Var(d) to yield $\rho(\Delta t)$.



Autocorrelation of time series data

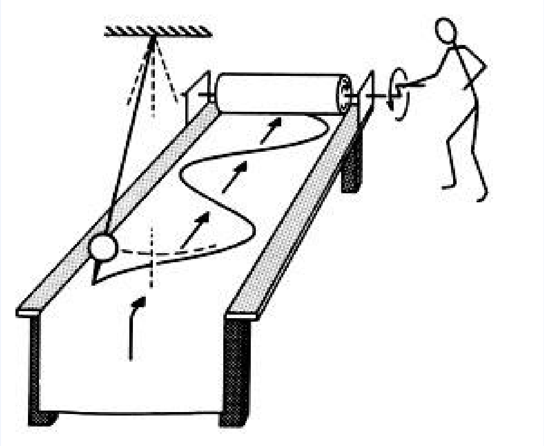
Time series data are **autocorrelated**. This means that *data points taken at times close to each other cannot be regarded as independent* items of information.

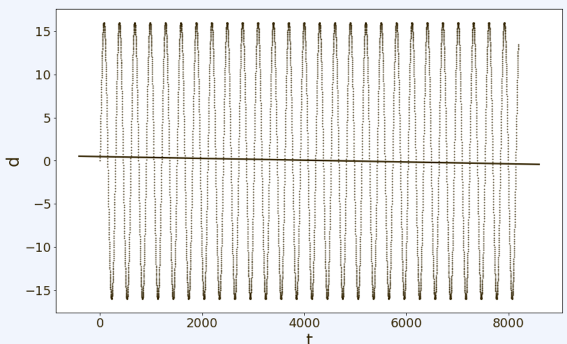


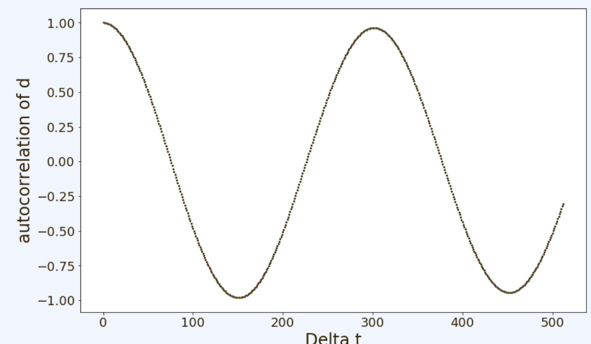
Assume we are given time series data $d(t)$:

autocorrelation $R(\Delta t) = \langle d(t)\, d(t + \Delta t)\rangle$
autocovariance $\langle [d(t) - \langle d\rangle]\, [d(t + \Delta t) - \langle d\rangle]\rangle$

Often **normalized** by Var(d) to yield $\rho(\Delta t)$.







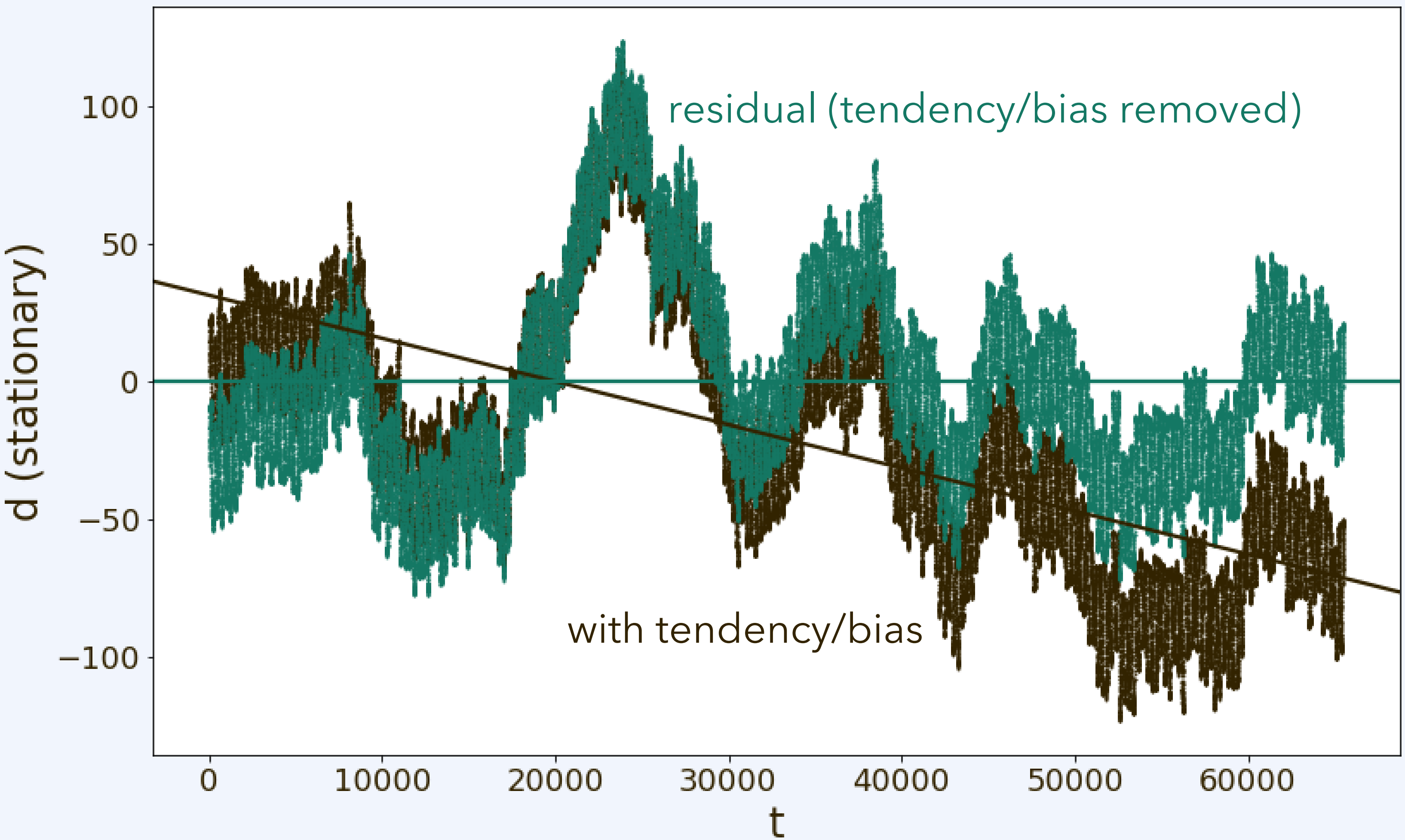
harmonic.ipynb

Attention: Autocorrelations may persist for a long time!

Figure: IOP, https://spark.iop.org/broomstick-pendulum-sinusoidal-motion.

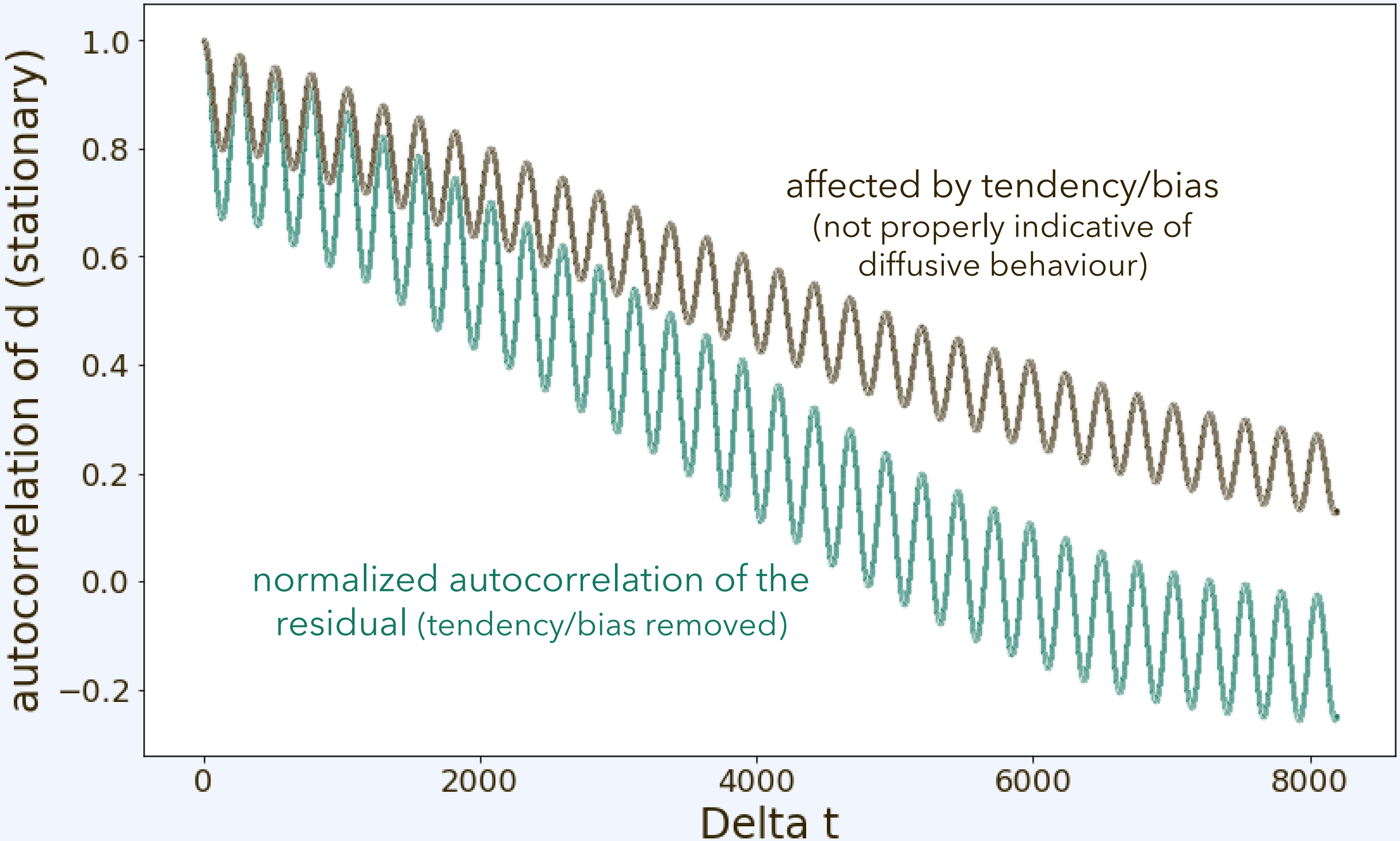
Autocorrelation function (normalized autocovariance)

autocorrelation-statsmodels.ipynb



Autocorrelation function (normalized autocovariance)

autocorrelation-statsmodels.ipynb



Decorrelation time

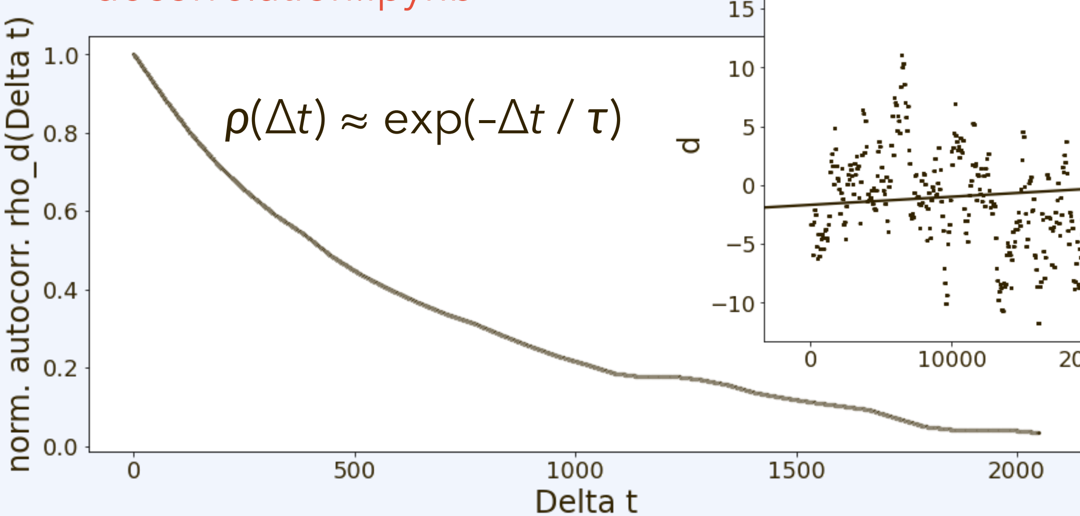
decorrelation time

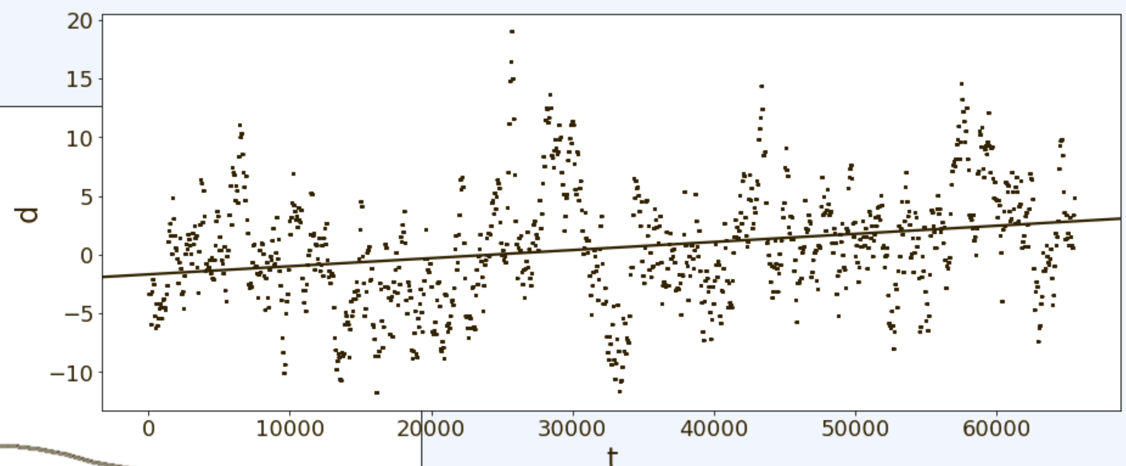
Our previously introduced approach to uncertainty estimation was based on:

- Some set of data that are representative of the phenomenon;
- Separation of data points into training, validation, and test data.
    - *What would happen if we split the points into these three at random?*

Different data points on a time series are not independent tests, they are correlated – this is exactly what is expressed by the autocorrelation function.

decorrelation.ipynb





Decorrelation time

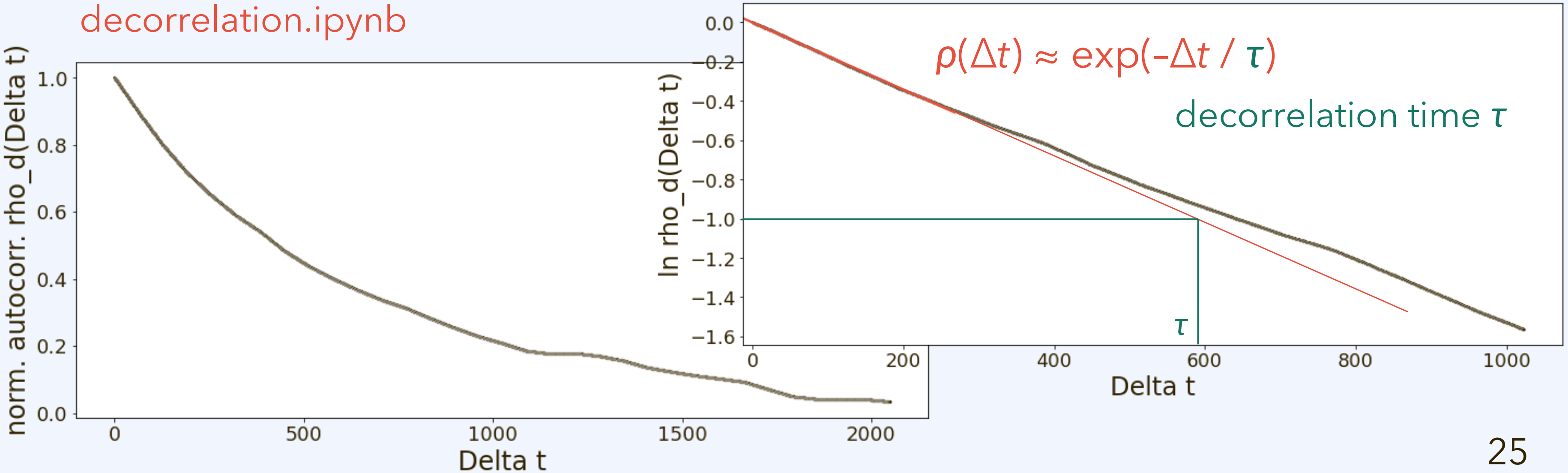
decorrelation time

Our previously introduced approach to uncertainty estimation was based on:

- Some set of data that are representative of the phenomenon;
- Separation of data points into training, validation, and test data.
    - *What would happen if we split the points into these three at random?*

Different data points on a time series are not independent tests, they are correlated – this is exactly what is expressed by the autocorrelation function.

decorrelation.ipynb



Uncertainty in time series

Our previously introduced approach to uncertainty estimation was based on:

- Some set of data that are representative of the phenomenon;
- Separation of data points such that they are *decorrelated*.

Once Δt exceeds 3τ, the normalized autocorrelation is small, $\rho(\Delta t) < 0.05$. We can *average over blocks* with size 3τ (or more) and then treat each of these **block averages** as independent data points.

Warning: This only works if the autocorrelation *actually is* decaying.

It will lead to mistakes where there is a strong correlation over long timespans, such as when analysing a periodic signal. If your data points are not really decorrelated, you can never treat them as independent items of information.

Uncertainty in time series

block averaging

Our previously introduced approach to uncertainty estimation was based on:

- Some set of data that are representative of the phenomenon;
- Separation of data points such that they are *decorrelated*.

Once Δt exceeds 3τ, the normalized autocorrelation is small, $\rho(\Delta t) < 0.05$. We can *average over blocks* with size 3τ (or more) and then treat each of these **block averages** as independent data points. This is called **block averaging**.

N_b such blocks correspond to $N_b - 1$ independent deviations from the mean.

Variance of the block averages: $\sigma_b^2 = (N_b - 1)^{-1} \Sigma (B_i - \langle B \rangle)^2$

Uncertainty based on σ, where $\sigma = N_b^{-1/2} \sigma_b$ from central limit theorem

A rigorous theory of block averaging was developed by Flyvbjerg and Petersen[1] (which is therefore also called *Flyvbjerg-Petersen block averaging*).

[1] H. Flyvbjerg, H. G. Petersen, *J. Chem. Phys.* **91**: 461–466, doi:10.1063/1.457480, **1989**.

Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap

Digitalisering på Ås

Conclusion

Glossary terms

Proposed glossary[1] terms:

- How do we best define them? Is the definition controversial?
- What is the best translation into Norwegian bokmål/nynorsk?
- Are there more key concepts that would require an agreed definition?

decorrelation time

(also: autocorrelation time)

residual quantity

influence diagram

suggested for discussion

uncertainty

block averaging

[1] https://home.bawue.de/~horsch/teaching/dat121/glossary-en.html

Norges miljø- og biovitenskapelige universitet

Institutt for datavitenskap

Digitalisering på Ås

DAT121
Introduction to data science

3 Regression basics

3.4 Influence diagrams

3.5 Residual quantities

3.6 Time series

Schedule for DAT121 part 3

Monday, 21st August 2023

9.15 – 10.00	Q&A session
10.15 – 11.00	first lecture on regression
11.15 – 12.00	discussion and problem solving
13.15 – 15.00	project work and tutorial

Tuesday, 22nd August 2023

10.15 – 12.00	scheduling of group sessions and of the final presentations
13.15 – 15.00	project work and tutorial

Wednesday, 23rd August 2023

10.15 – 11.00	second lecture on regression
11.15 – 12.00	**interest group sessions**
13.15 – 15.00	project work and tutorial

Thursday, 24th August 2023

10.15 – 11.00	third lecture on regression
11.15 – 12.00	discussion and problem solving

Schedule for DAT121 parts 4 and 5

Friday, 25th August 2023

10.15 – 11.00 lecture on good practice

11.15 – 12.00 interest group sessions

13.15 – 15.00 project work and tutorial

Monday, 28th August 2023

9.15 – 10.00 first multidimensionality lecture

10.15 – 10.?? Pangasia presentation in TF1-115

11.15 – 12.00 discussion and problem solving

13.15 – 15.00 project work and tutorial

Tuesday, 29th August 2023

9.15 – 10.00 Q&A session and discussion

10.15 – 11.00 second multidimensionality lecture

11.15 – 12.00 interest group sessions

13.15 – 15.00 project work and tutorial