Predictive confidence distributions for time series

Gudmund Hermansen (with Nils Lid Hjort)

University of Oslo

May 13, 2015

Motivation

- In several applications the main goal of fitting a statistical model is to predict future outcomes.
- The construction of confidence intervals, or credibility intervals, are therefore of great interest, i.e. confidence distributions.
- Here we will discuss a framework for confidence distributions for a future unobserved variable Y_0 that we wish to predict.
- In addition, we discuss how to make confidence distributions for
  - the probability of exceeding a threshold and
  - the time point t corresponding to the event that $\mathrm{E}\, Y_t$ crosses y_0.
- The focus here is on various time series models.
- It is preliminary work and a extended discussion of ideas and methods presented in Schweder & Hjort (2015, Chapter 12)

Confidence distributions for future observations

- Following the framework of Schweder & Hjort (2015) the CD

$$C_{\text{pred}}(y_0, Y),$$

for variable Y_0 that we wish to predict based on data Y, has to satisfy two conditions:

(i) $C_{\text{pred}}(y_0, \boldsymbol{y}_{\text{obs}})$ is a distribution function in y_0 for each $\boldsymbol{y}_{\text{obs}}$ and

(ii) $C_{\text{pred}}(Y_0, Y) \sim U[0,1]$, when (Y_0, Y) has the joint distribution of the true model.

- Note that this is different from estimating $\mathrm{E}\, Y_0$ or $\mathrm{E}\, Y_0 \,|\, \boldsymbol{y}_{\text{obs}}$.

- We will (try to) motivate an approximate pivot for linear predictors and apply the general CD framework in some applications.

Prototype illustration with independent realisations

- Before we introduce dependency, consider the prototype case of predicting Y_{n+1} given a sequence of i.i.d. realisations.
- Suppose we wish to predict Y_{n+1} given a series of n i.i.d. realisations, where $Y_1 \sim \mathrm{N}(\mu, \sigma^2)$ (with μ and σ unknown).
- Given $Y = Y_1, \ldots, Y_n$ we know that that

$$R = \frac{Y_{n+1} - \bar{Y}_n}{\widehat{\sigma}} \sim H_n$$

  is a pivot (its distribution is independent of μ and σ).
- And hence that

$$C_{\text{pred}}(y_{n+1}, \boldsymbol{y}_{\text{obs}}) = H_n\left(\frac{y_{n+1} - \bar{y}_{\text{obs}}}{\widehat{\sigma}_{\text{obs}}}\right)$$

  satisfies the two criteria above and is a valid CD construction.

Prototype illustration with dependent realisations

- Let
$$\begin{pmatrix} Y_0 \\ Y_1 \end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$
- Then $Y_1 \,|\, Y_0 = y_0 \sim \mathrm{N}(\rho y_0, 1 - \rho^2)$ and if ρ and σ are known
$$\frac{Y_1 - \rho y_0}{\sqrt{1-\rho^2}} \,|\, y_0 \sim \mathrm{N}(0, 1),$$
but this is true for all y_0, meaning that it is also true unconditional.
- This motivates
$$C_{\text{pred}}(y_1) = \Phi\left(\frac{y_1 - \rho y_0}{\sqrt{1-\rho^2}}\right),$$
where Φ is the cdf. of a standard normal, but is it more 'natural' to consider
$$C_{\text{pred}}(y_1) = \Phi\left(\frac{y_1 - \rho y_0}{\sqrt{1-\rho^2}}\right) \,|\, y_0?$$
- It is less clear what happens in more general models, and in cases where parameters need to be estimated.

A family of time series modells

- Let
$$Y_t = m(x_t, \beta) + \epsilon_t, \qquad \text{for all } t \geq n \text{ and } n \geq 1,$$
where:
  - m is known, but depending on a unknown parameter β
  - x_t is a vector-values sequence of known covariates
  - ϵ_t is a stationary time series prosess with mean zero and dependency that is known up to a unknown parameter θ.
- Here m is typically a polynomial of low order, e.g.
$$Y_t = \beta_0 + \beta_1 x_{t,1} + \cdots + \beta_p x_{t,p} + \epsilon_t,$$
and we will also often assume that ϵ_t is a autoregressive process, i.e.
$$\epsilon_t = \rho_1 \epsilon_{t-1} + \cdots + \rho_q \epsilon_{t-q} + \sigma \delta_t,$$
where δ_t are independent $\mathrm{N}(0, 1)$.
- The framework is easily extended to other more general types of model constructions.

How to make good predictions?

- If ϵ_t above is Gaussian, the 'answer' is conditional expectation, i.e.

$$\widehat{y}_0 = \mathrm{E}\left[Y_0 \,|\, \boldsymbol{y}_{\mathrm{obs}}, X, \widehat{\beta}, \widehat{\theta}\,\right],$$

  where the unknown parameters are fitted by maximum likelihood.

- If ϵ_t is an autoregressive prosess, it is quite common to use so-called 'conditional quasi maximum likelihood' (which simplifies the estimation to ordinary least squares (OLS)).

Example: If $\epsilon_t \sim \mathrm{AR}(q)$ with Gaussian errors, then the likelihood $\ell_n(\beta, \theta \,|\, \boldsymbol{y}_{obs}) \,|\, y_1, \ldots y_q$ is structurally equivalent to that of a linear regression model.

- There is a rich theory that aims at linear predictors, say

$$\widehat{y}_0 = \widehat{y}_{n+h} = \alpha^t \boldsymbol{x}_{n+h} + \gamma^t \boldsymbol{y}_{\mathrm{obs}}, \qquad \text{for } h \geq 1,$$

  which minimises the mean squared error $\mathrm{E}\,(Y_0 - \widehat{Y}_0)^2$, see e.g. Grenander & Rosenblatt (1957).

- We will argue that linear predictors fitted by minimising the sum of squared prediction errors have some desired qualities.

Predicting with the true model

- As simple starting point suppose Y_t is an autoregressive model of order one, then

$$Y_t = \epsilon_t = \rho\epsilon_{t-1} + \sigma\delta_t = \rho Y_{t-1} + \sigma\delta_t, \qquad \text{for } t \leq n,$$

  where δ_t are independent $\mathrm{N}(0,1)$.

- Given the past $Y_1, \ldots, Y_n$ our best guess for $Y_0 = Y_{n+1}$ with ρ known is $\widehat{Y}_{n+1} = \rho Y_n$.

- Moreover, if σ is also known

$$R_n = \frac{Y_{n+1} - \widehat{Y}_{n+1}}{\sigma} = \frac{\rho Y_n + \sigma\epsilon_t - \rho Y_n}{\sigma} = \epsilon_t \sim \mathrm{N}(0,1),$$

  which is a pivot.

- Resulting in the simple confidence distribution construction:

$$C(y_0, \boldsymbol{y}_{\text{obs}}) = \Phi\left(\frac{y_0 - \rho y_n}{\sigma}\right).$$

What if some parameters are unknown?

- If ρ is known and σ has be estimated, then

$$R_n = \frac{Y_{n+1} - \widehat{Y}_{n+1}}{\widehat{\sigma}} \sim t_{n-2},$$

since

$$\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{t=2}^{n}(Y_t - \widehat{Y}_{t-1})^2 \sim \chi^2_{n-2}$$

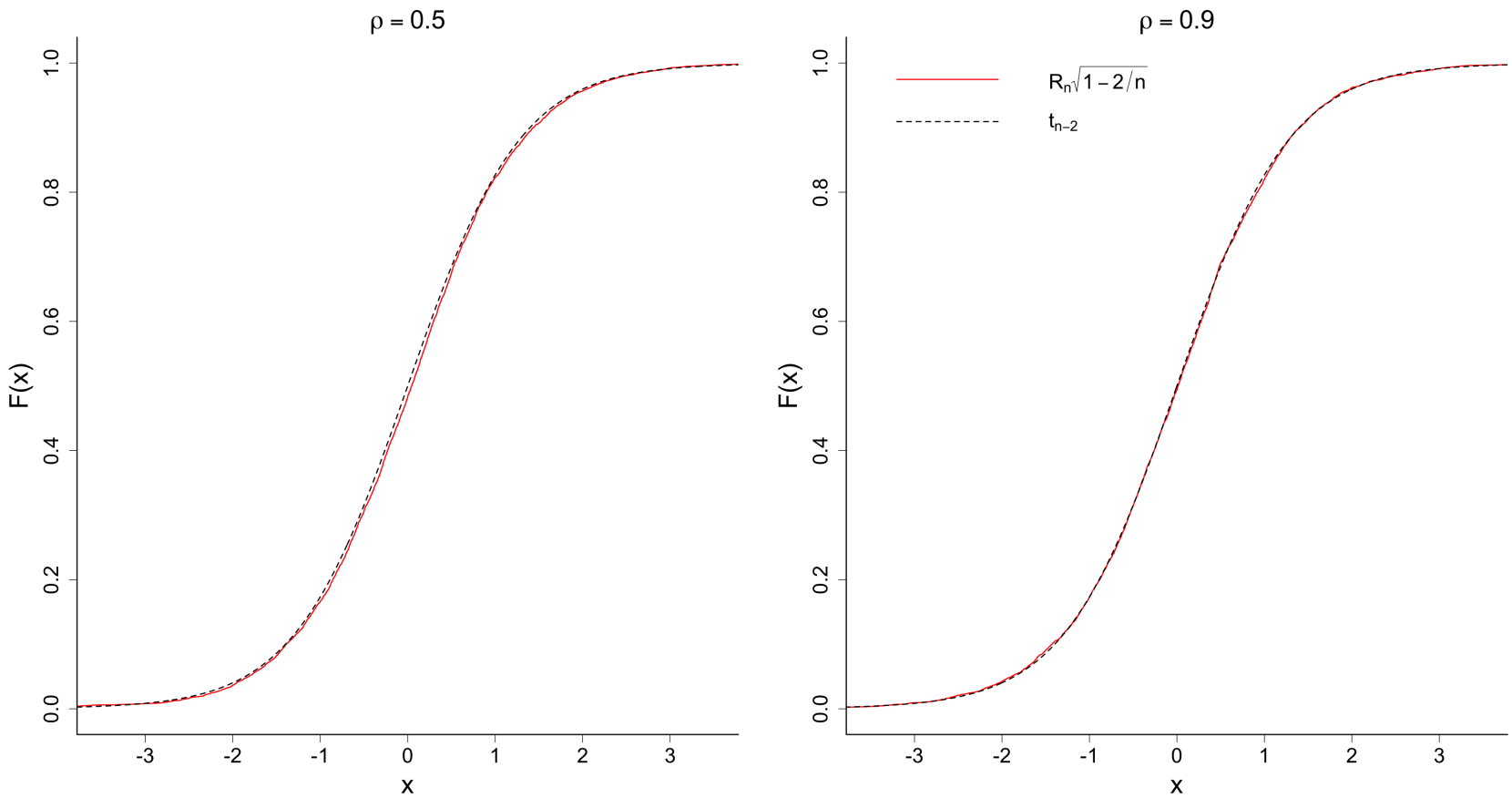
- If ρ also has to be estimated, the hope is that if $\widehat{\rho}$ is close to ρ

$$R_n = \frac{Y_{n+1} - \widehat{\rho}\, Y_n}{\widehat{\sigma}}$$

is an approximate pivot; at least for large n.

- In simulations studies, we observe that $R_n\sqrt{1-2/n}$ is close to a t_{n-2}-distribution, for all ρ and σ.

‘Proof of concept’ with $n = 10$ samples



A more general modelling framework

- In practice, we (currently) obtain the best simulation results (fit and coverage) with

$$R_n = \frac{Y_{n+1} - \widehat{Y}_{n+1}}{s},$$

where

$$s^2 = \frac{1}{n-2}\sum_{t=2}^{n}(r_t - \bar{r}_n)^2$$

and where $r_t = Y_t - \widehat{Y}_t$ are the residuals.

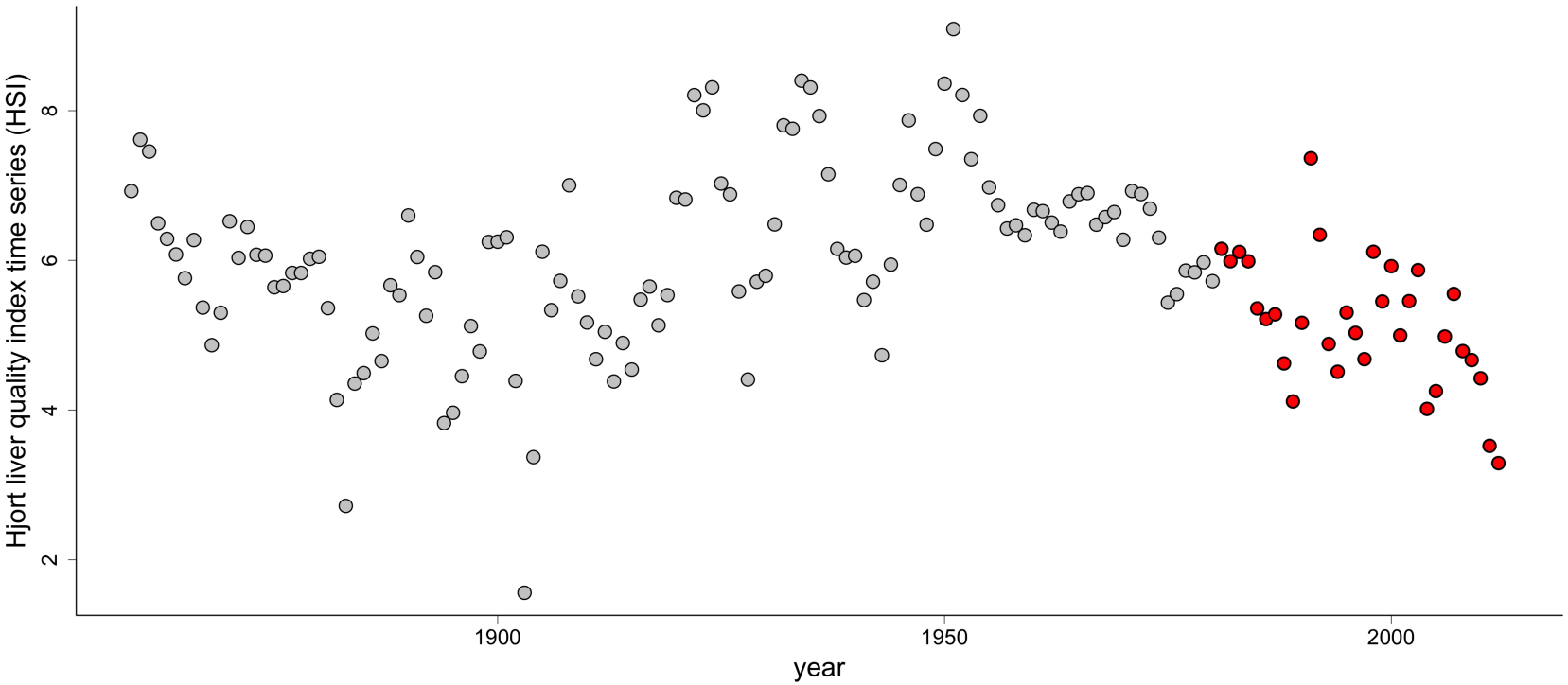
- A similar results seems to be true (in simulations) for general linear predictors of the type

$$\widehat{y}_0 = \widehat{y}_{n+h} = \widehat{\alpha}^t \boldsymbol{x}_{n+h} + \widehat{\gamma}^t \boldsymbol{y}_{\text{obs}}, \qquad \text{for } h \geq 1,$$

where $\widehat{\alpha}$ and $\widehat{\gamma}$ are fitted by minimising the empirical sum of squared prediction errors.

- The approximation seems to be quite robust, also in very small samples, and to work well under model misspecification.

Illustration: The liver quality index

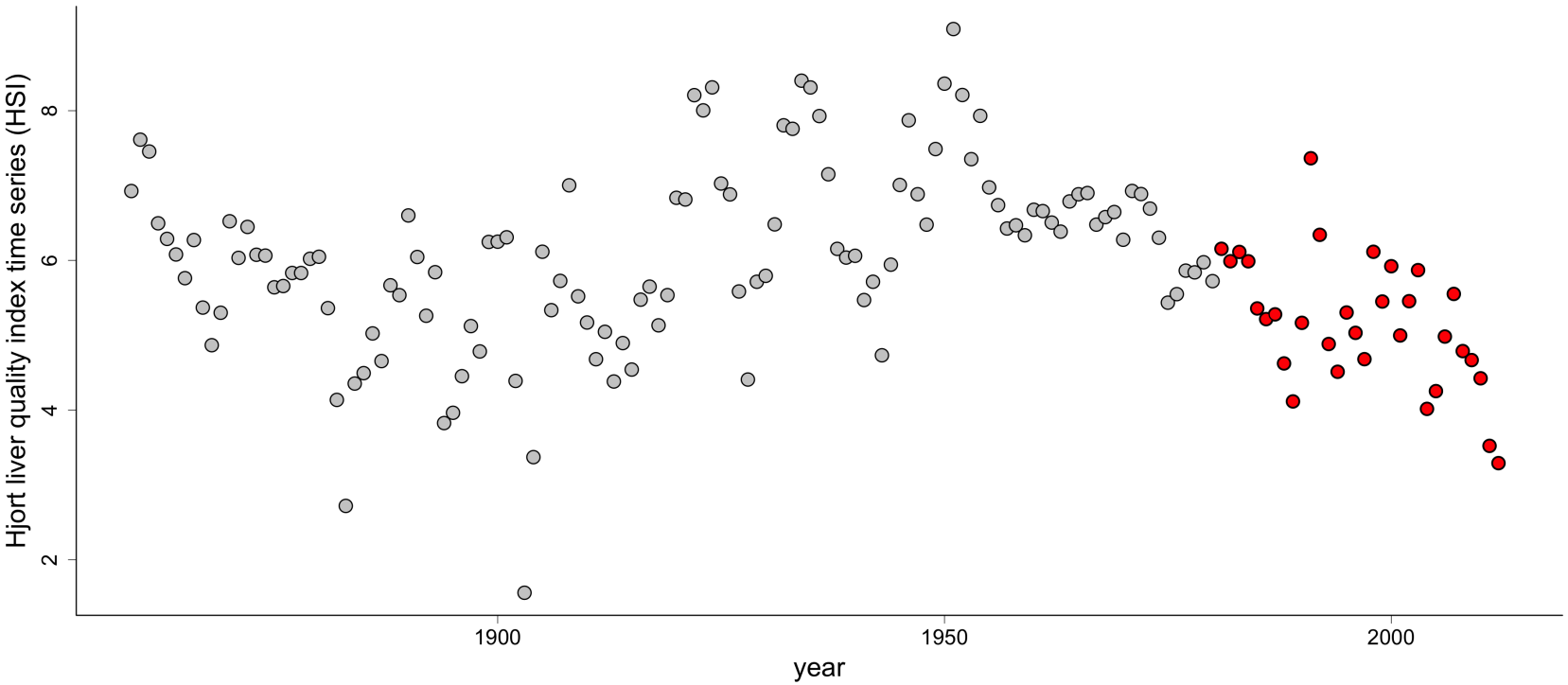


- The liver quality index (HSI) for a individual fish is defined as

$$\text{HSI}_{\text{fish}} = 100 \times \frac{\text{weight of liver}}{\text{weight of fish}}.$$

- The HSI index is viewed as a measure of the quality of life and is e.g. related to reproduction.

Illustration: The liver quality index

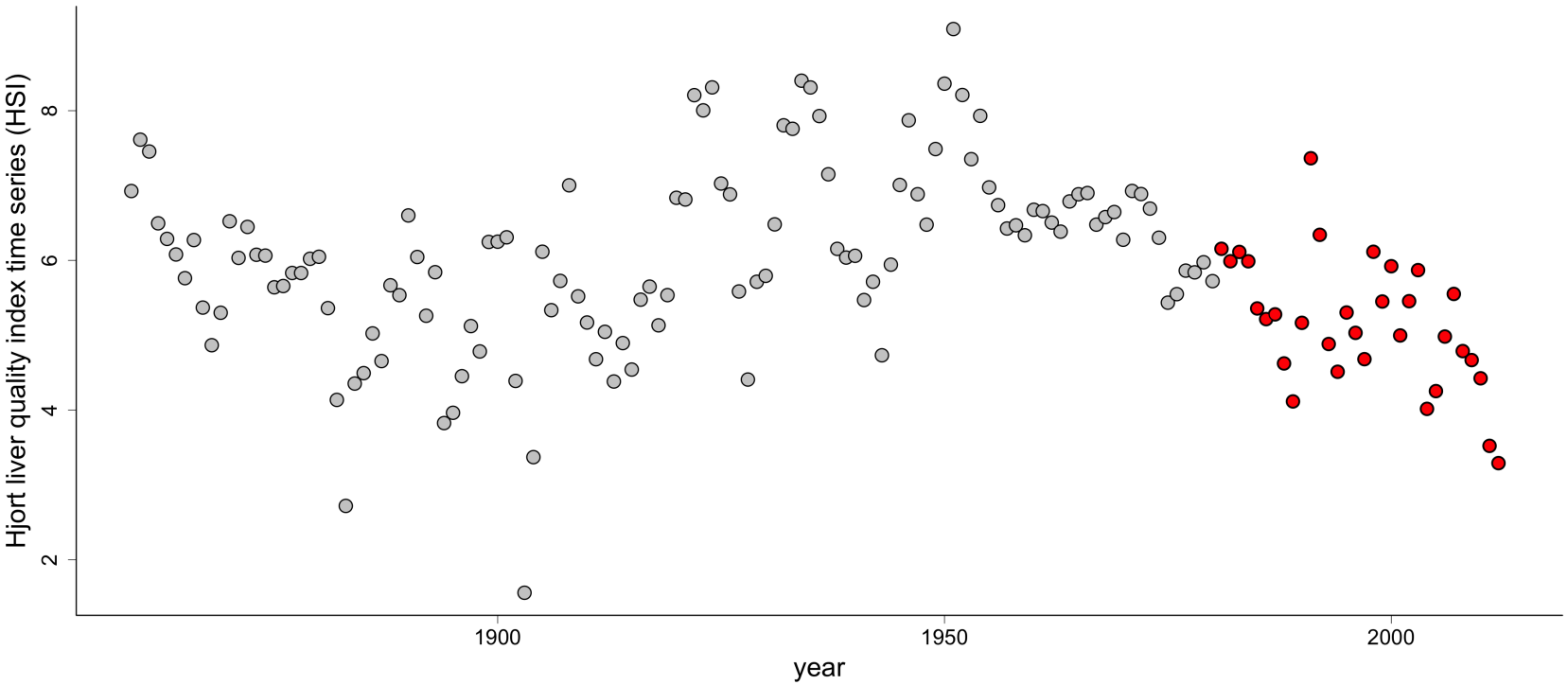


- For a population of fish in a given year i, the liver quality index is commonly estimated by

$$\widehat{\text{HSI}}_{\text{bulk},i} = 100 \times \frac{\text{total amount of liver}}{\text{total amount of fish}},$$

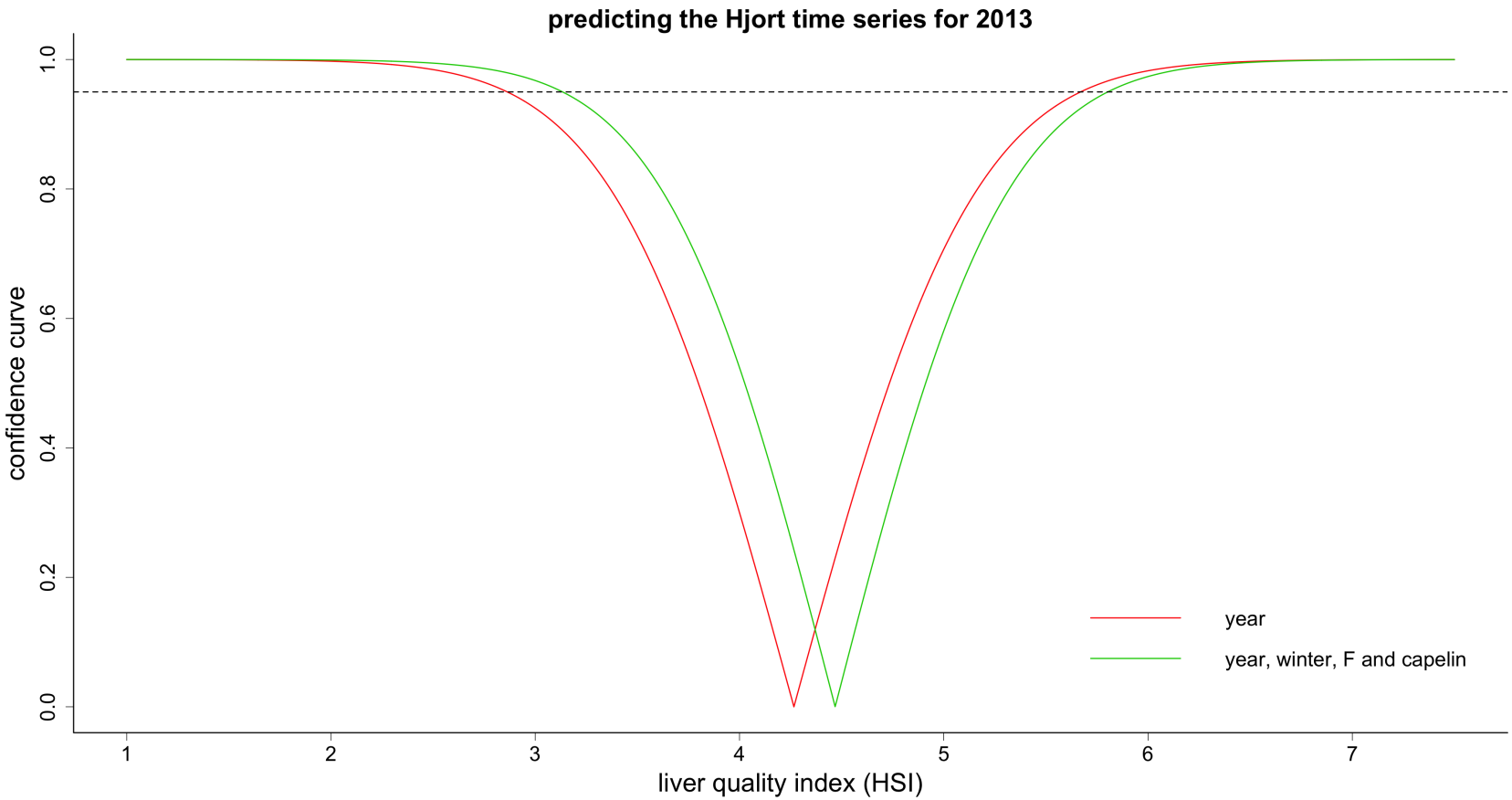
for fish caught in year i.

Illustration: The liver quality index



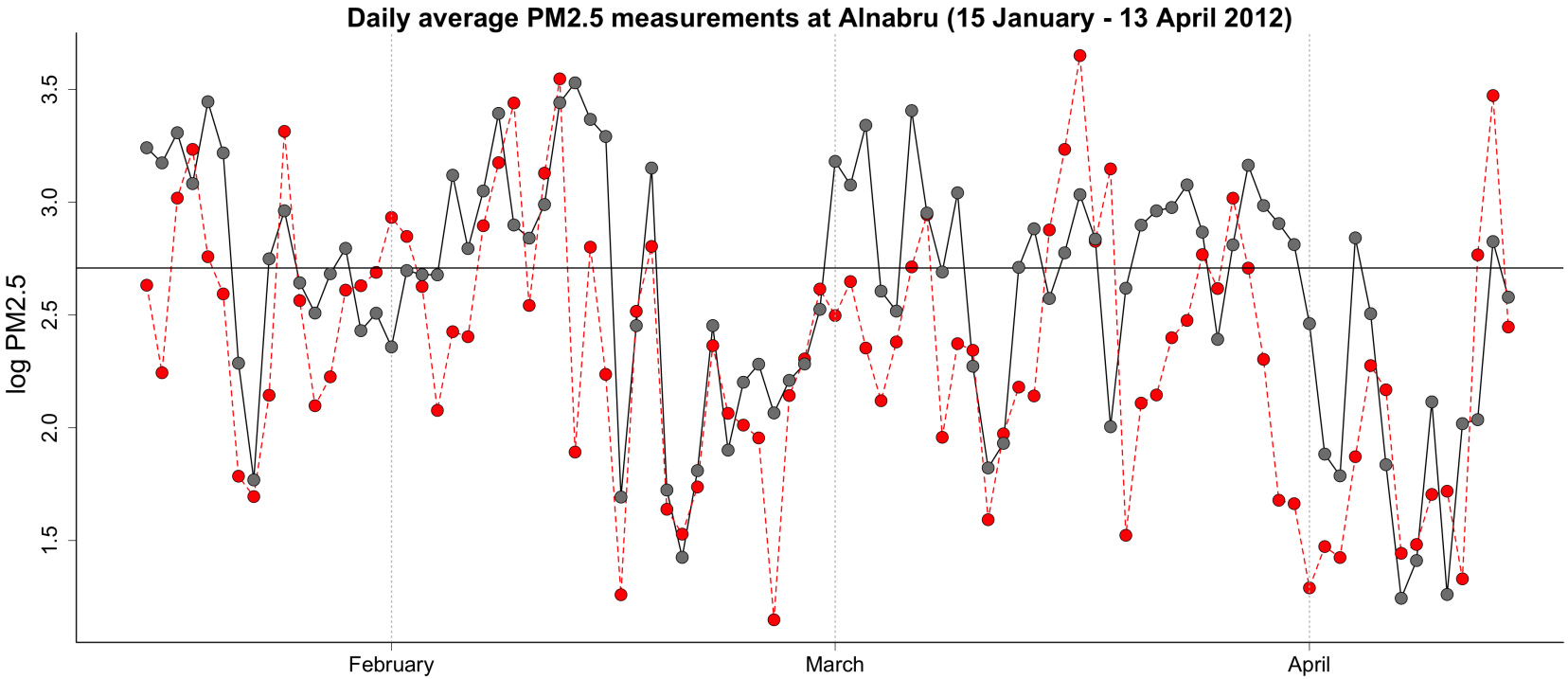
- It is important to understand the dynamic of the HSI index, e.g. how external factors, like changes in sea temperature, death rates and food supply affects the series.
- As an illustration we will compare two confidence curves for the 'future' HSI index of 2013 (i.e. the next observation).

Confidence distribution for the next observation



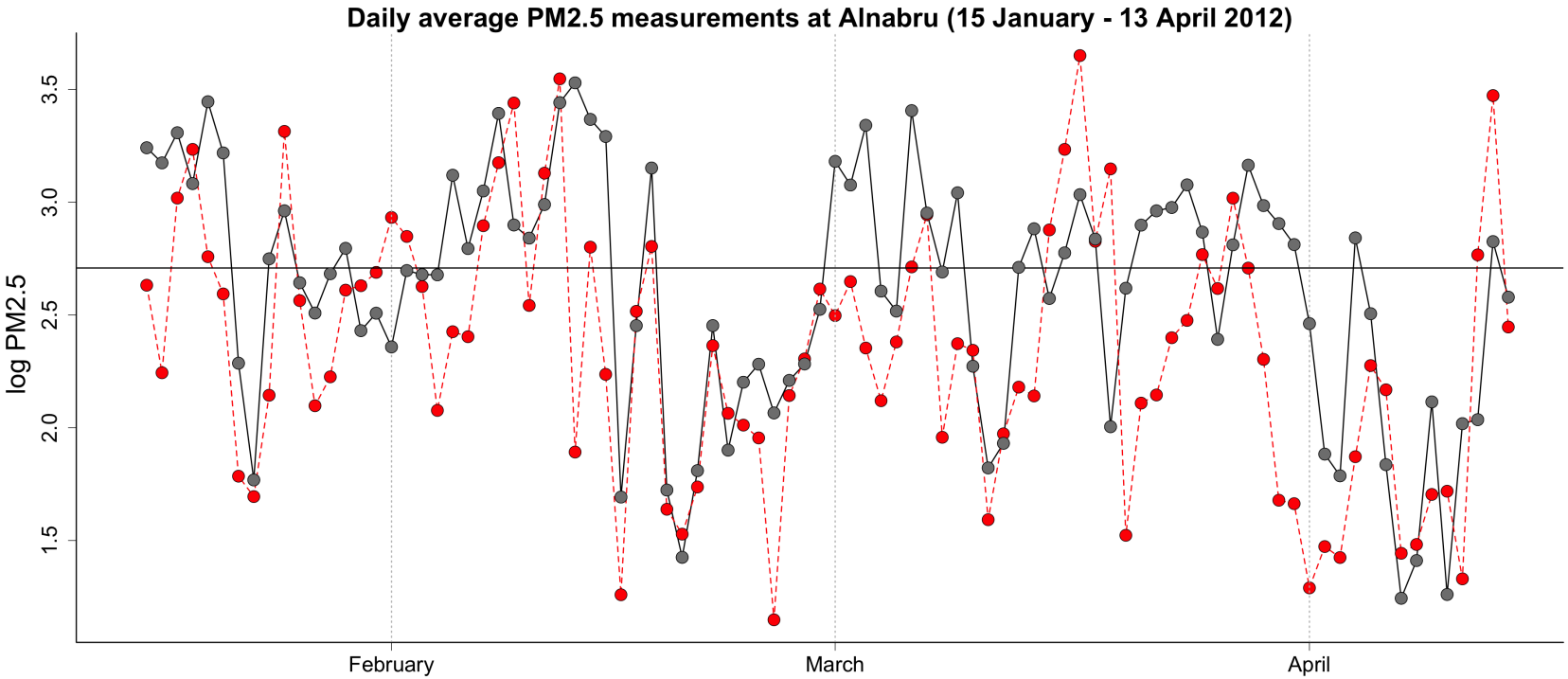
- The confidence curve are constructed using the approximatins introduced above.
- The idea is to see how adding additional covariates affects the prediction with respect to the confidence curve.

Illustration: Particle pollution ($PM_{2.5}$)



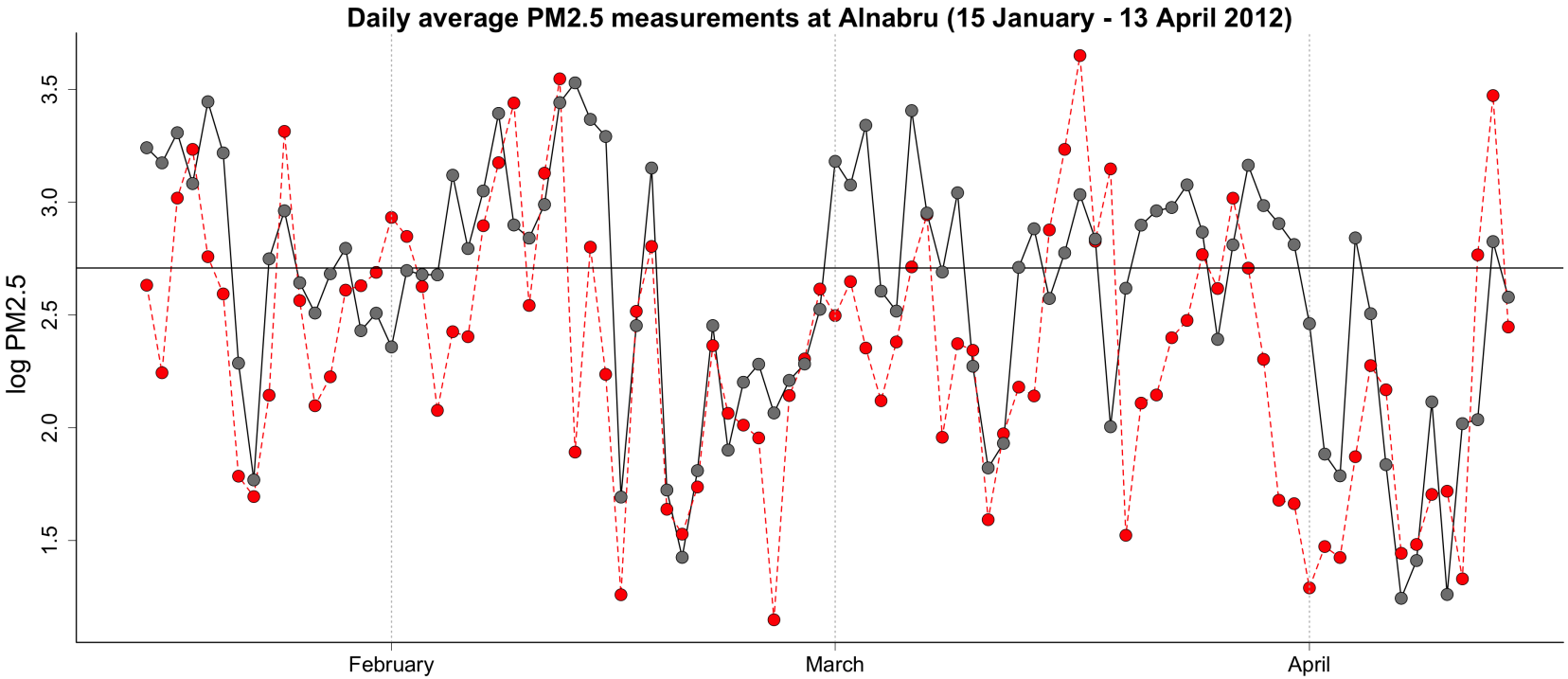
- Daily average measurements of fine particulate matter ($PM_{2.5}$), a type of air pollutant of tiny particles, or droplets, in the air.
- Exposure can cause short-term effects, like eye, nose, throat and lung irritation, long-term exposure can affect lung function and has also been shown to be related to asthma and heart disease.

Illustration: Particle pollution ($PM_{2.5}$)



- For health reasons, government regulations typically restrict the daily average emission to 25-35 $\mu g/\mathrm{m}^3$.
- New directives in Norway suggest that the level may be set as low as 15 $\mu g/\mathrm{m}^3$ with the possibility of 8 exceedances per year.

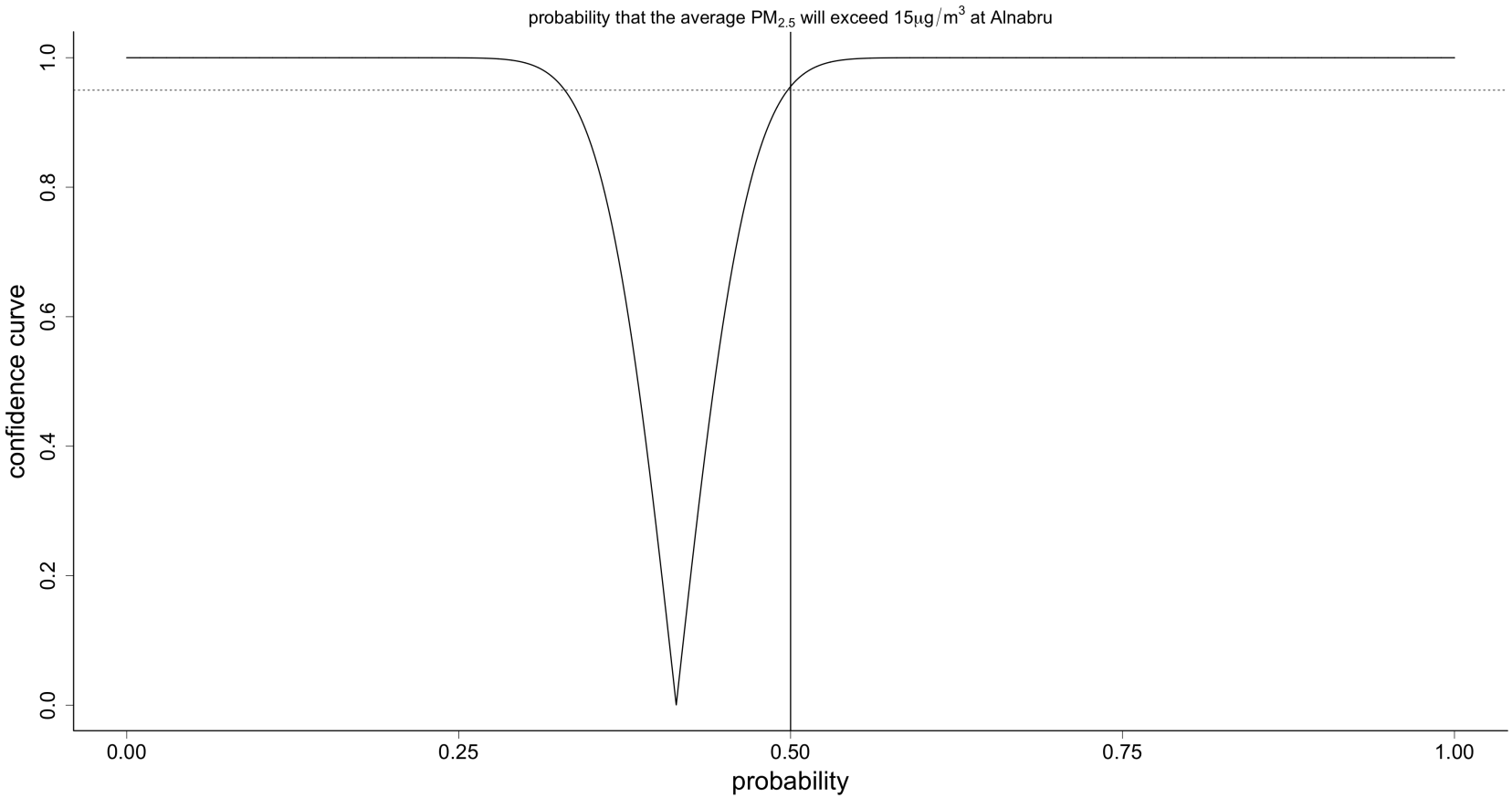
Illustration: Particle pollution ($PM_{2.5}$)



- One of the main objective is to use forecast to detect periods, or days with potential high emissions, enabling the legislature to take appropriate actions in advanced.
- As an application we will make the confidence curve for

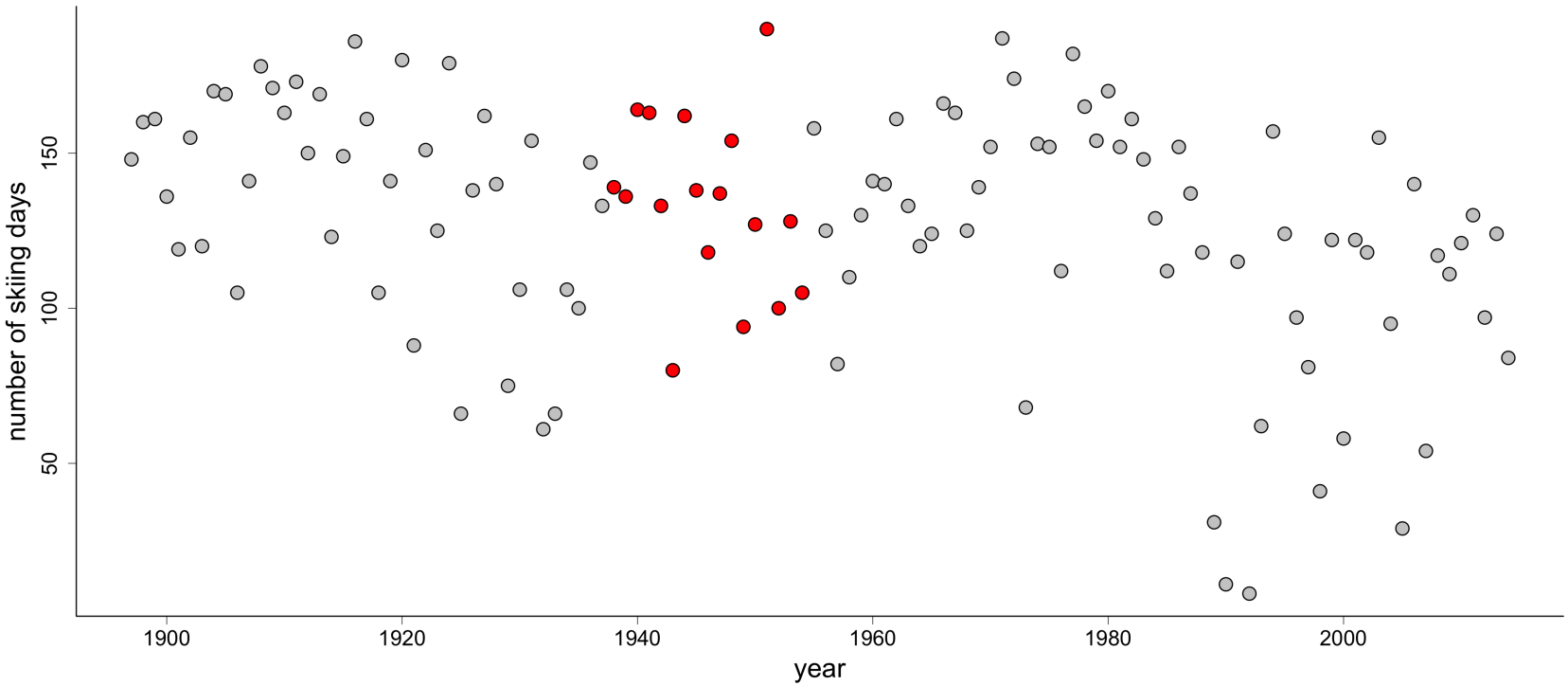
$$\psi = \psi(\theta) = \Pr\{Y_{n+1} \geq \log(15) \,|\, \boldsymbol{x}_{\text{forecast}}, \boldsymbol{y}_{\text{obs}}, \theta\}.$$

Illustration: Particle pollution ($PM_{2.5}$)



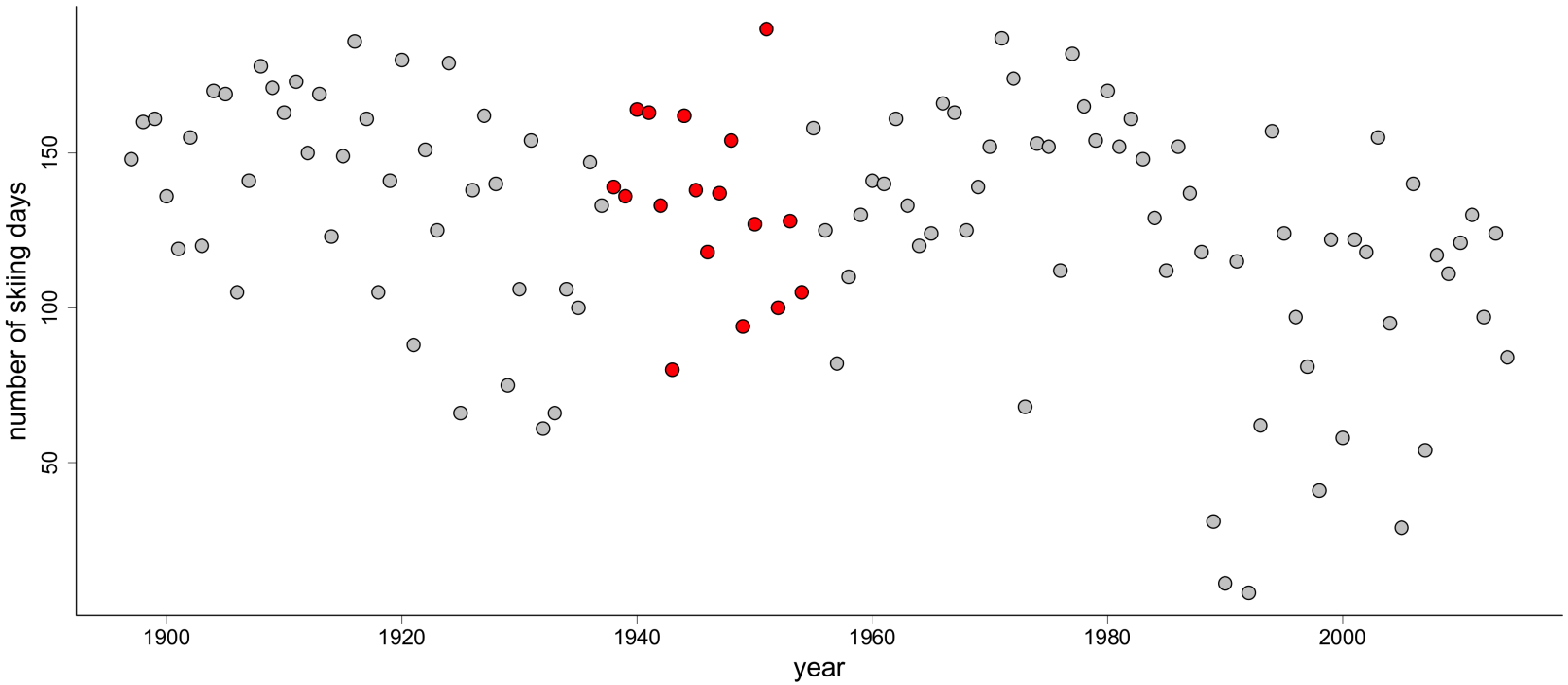
- The confidence curve is constructed using the delta method and the asymptotic normality of the estimated parameters.
- In simulations the large-sample approximate seems robust and the profile log-likelihood approach is often unstable.

Illustration: Number of skiing days



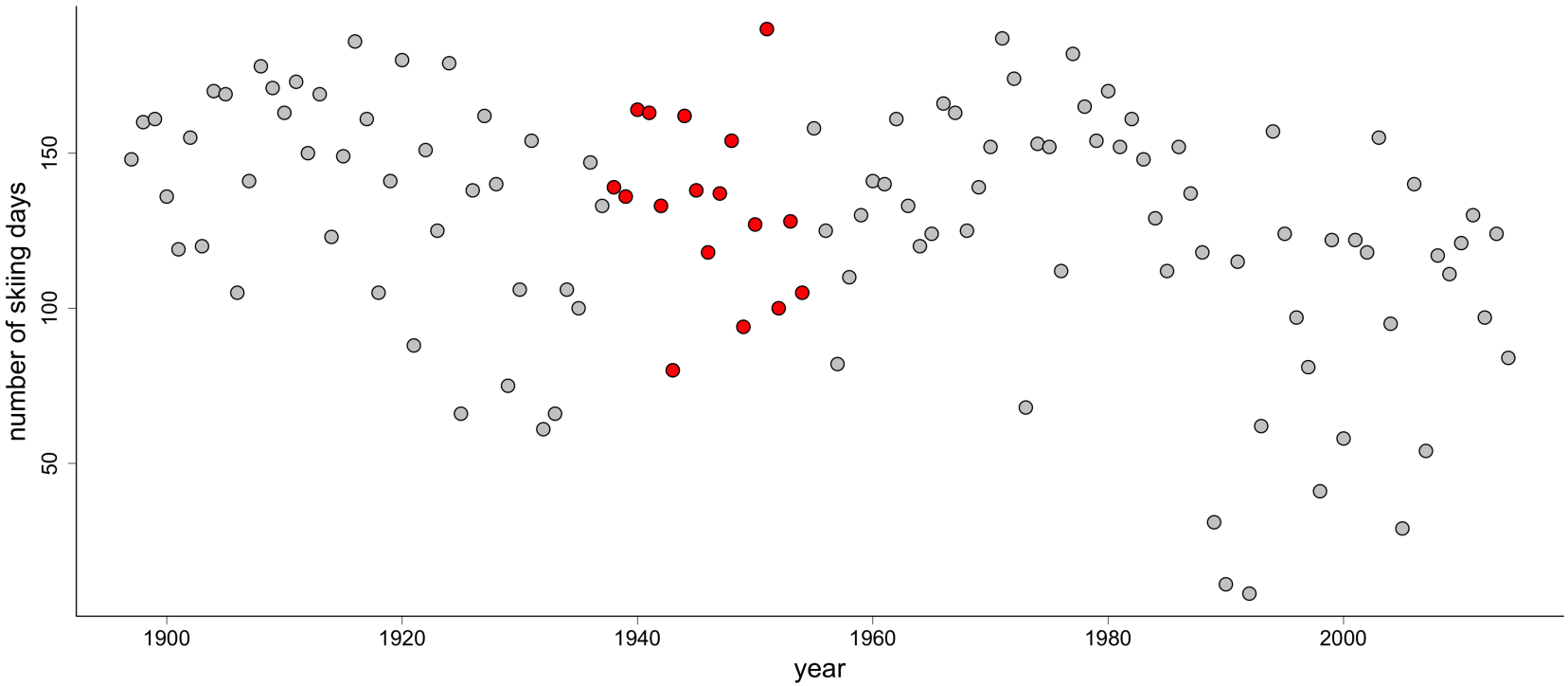
- The number of skiing days in a winter season is defined as the number of days with at least 25 cm snow.
- Meteorologist Gustav Bjørbæk introduced the definition as the least amount of snow needed to avoid serious injury (in case of a fall).

Illustration: Number of skiing days



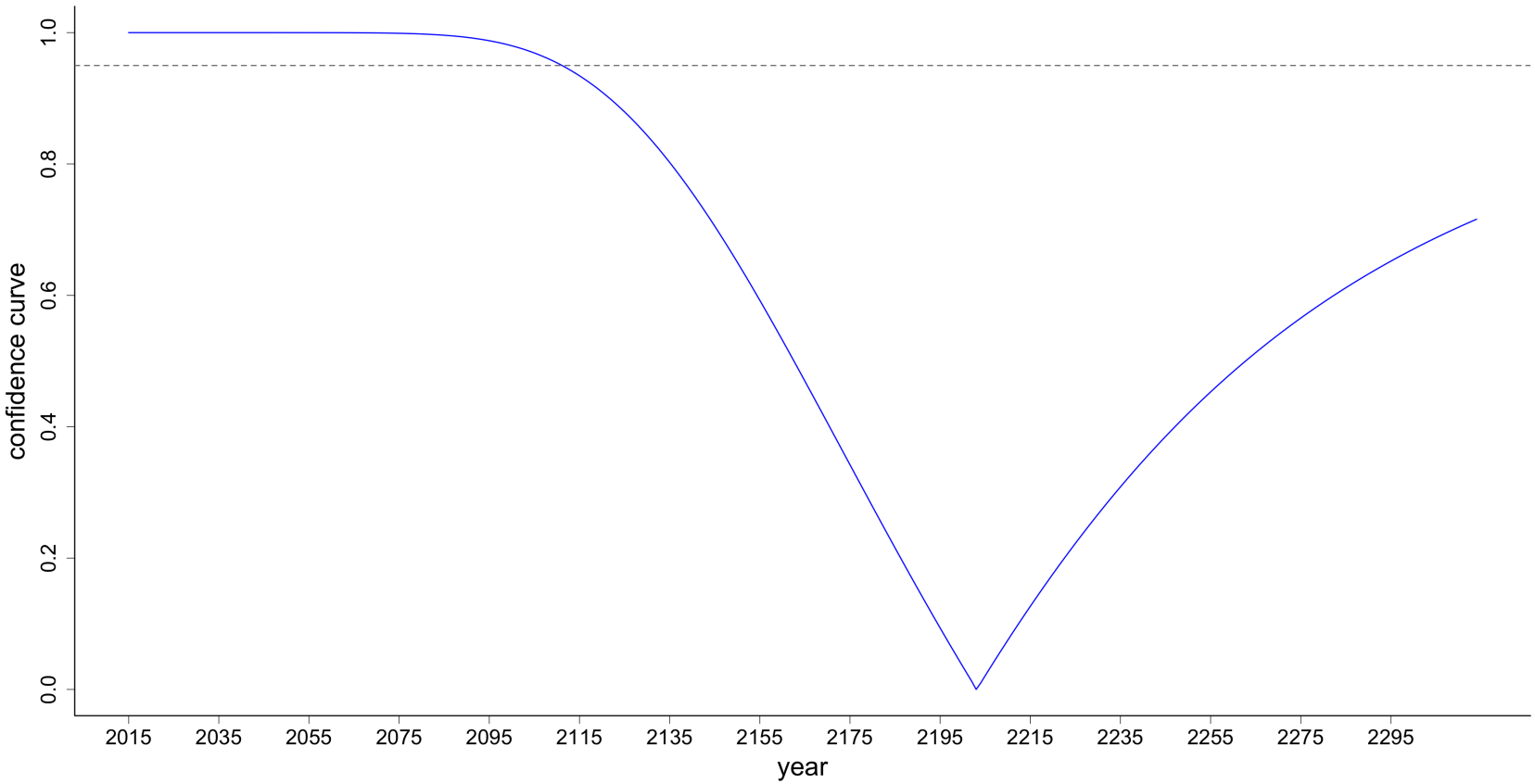
- Besides being of great interest to skiing enthusiasts, the number of skiing days can be seen as a indicator of how cold a winter is and is also an indication of the general temperature over a given period of time.
- Since 1896 the average decline is about -0.33 skiing days per year.

Illustration: Number of skiing days



- Since the mid 1950s the decline has a estimated slope of -0.9.
- The goal is to study, and make a confidence curve for, the time point t the expected value $\mathrm{E}\, Y_t$ crosses $y_0 = 30$.

Illustration: Number of skiing days



- The confidence curve is fitted using the profile log-likelihood approach and converting the deviance function.