02417: Time Series Analysis

Week 2 - Regression based methods, 1st part

Peder Bacher
DTU Compute

Based on material previous material from the course

February 13, 2026

General form of the regression model

$$Y_t = f(\boldsymbol{X}_t, t; \boldsymbol{\theta}) + \varepsilon_t$$

General form of the regression model

$$Y_t = f(\boldsymbol{X}_t, t; \boldsymbol{\theta}) + \varepsilon_t$$

Where:

- Y_t is the output we aim to model

General form of the regression model

$$Y_t = f(\boldsymbol{X}_t, t; \boldsymbol{\theta}) + \varepsilon_t$$

Where:

- Y_t is the output we aim to model
- $\boldsymbol{X}_t$ indicates the p independent variables $\boldsymbol{X}_t = (X_{1t}, \cdots, X_{pt})^T$

General form of the regression model

$$Y_t = f(\boldsymbol{X}_t, t; \boldsymbol{\theta}) + \varepsilon_t$$

Where:

- Y_t is the output we aim to model
- $\boldsymbol{X}_t$ indicates the p independent variables $\boldsymbol{X}_t = (X_{1t}, \cdots, X_{pt})^T$
- t is the time index

General form of the regression model

$$Y_t = f(\boldsymbol{X}_t, t; \boldsymbol{\theta}) + \varepsilon_t$$

Where:

- Y_t is the output we aim to model
- $\boldsymbol{X}_t$ indicates the p independent variables $\boldsymbol{X}_t = (X_{1t}, \cdots, X_{pt})^T$
- t is the time index
- $\boldsymbol{\theta}$ indicates m unknown parameters $(\theta_1, \cdots, \theta_m)^T$

General form of the regression model

$$Y_t = f(\boldsymbol{X}_t, t; \boldsymbol{\theta}) + \varepsilon_t$$

Where:

- Y_t is the output we aim to model
- $\boldsymbol{X}_t$ indicates the p independent variables $\boldsymbol{X}_t = (X_{1t}, \cdots, X_{pt})^T$
- t is the time index
- $\boldsymbol{\theta}$ indicates m unknown parameters $(\theta_1, \cdots, \theta_m)^T$
- ε_t is a sequence of random variables with mean zero, variance σ_t^2, and $\mathrm{Cov}[\varepsilon_{t_i}, \varepsilon_{t_j}] = \sigma_t^2 \Sigma_{ij}$

General form of the regression model

$$Y_t = f(\boldsymbol{X}_t, t; \boldsymbol{\theta}) + \varepsilon_t$$

Where:

- Y_t is the output we aim to model
- $\boldsymbol{X}_t$ indicates the p independent variables $\boldsymbol{X}_t = (X_{1t}, \cdots, X_{pt})^T$
- t is the time index
- $\boldsymbol{\theta}$ indicates m unknown parameters $(\theta_1, \cdots, \theta_m)^T$
- ε_t is a sequence of random variables with mean zero, variance σ_t^2, and $\mathrm{Cov}[\varepsilon_{t_i}, \varepsilon_{t_j}] = \sigma_t^2 \Sigma_{ij}$

For now we restrict the discussion to the case where $\boldsymbol{X}_t$ is non-random and thus we write $\boldsymbol{x}_t$ instead of $\boldsymbol{X}_t$.

Ordinary least squares (OLS) estimates

Observations:

$$(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \cdots, (y_n, \boldsymbol{x}_n)$$

Ordinary least squares (OLS) estimates

Observations:

$$(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \cdots, (y_n, \boldsymbol{x}_n)$$

Ordinary Least Square (unweighted) estimates are found from

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} S(\boldsymbol{\theta})$$

Ordinary least squares (OLS) estimates

Observations:

$$(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \cdots, (y_n, \boldsymbol{x}_n)$$

Ordinary Least Square (unweighted) estimates are found from

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} S(\boldsymbol{\theta})$$

where

$$S(\boldsymbol{\theta}) = \sum_{t=1}^{n} [y_t - f(\boldsymbol{x}_t; \boldsymbol{\theta})]^2 = \sum_{t=1}^{n} \varepsilon_t^2(\boldsymbol{\theta})$$

For the unweighted method to result in reliable estimates, the errors must be assumed to all have the same variance and be mutually uncorrelated.

OLS – Variance of error and estimates

If the model errors ε_t are i.i.d.

- The variance of the model errors is estimated as:

$$\widehat{\sigma}^2 = \frac{S(\widehat{\boldsymbol{\theta}})}{n-p}$$

  where p is the number of estimated parameters.

OLS – Variance of error and estimates

If the model errors ε_t are i.i.d.

- The variance of the model errors is estimated as:

$$\widehat{\sigma}^2 = \frac{S(\widehat{\boldsymbol{\theta}})}{n-p}$$

  where p is the number of estimated parameters.

- The variance-covariance matrix of the estimates is approximately

$$V[\widehat{\boldsymbol{\theta}}] = 2\widehat{\sigma}^2 \left[\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} S(\boldsymbol{\theta})\right]^{-1}\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

The General Linear Model (GLM)

$$Y_t = \boldsymbol{x}_t^T \boldsymbol{\theta} + \varepsilon_t$$

The General Linear Model (GLM)

$$Y_t = \boldsymbol{x}_t^T \boldsymbol{\theta} + \varepsilon_t$$

Is this model a GLM?

$$Y_t = \theta_0 + \theta_1 z_t + \varepsilon_t$$

The General Linear Model (GLM)

$$Y_t = \boldsymbol{x}_t^T \boldsymbol{\theta} + \varepsilon_t$$

Is this model a GLM?

$$Y_t = \theta_0 + \theta_1 z_t + \varepsilon_t$$

Yes, what about this one?

$$Y_t = \theta_0 + \theta_1 z_t + \theta_2 z_t^2 + \varepsilon_t$$

The General Linear Model (GLM)

$$Y_t = \boldsymbol{x}_t^T \boldsymbol{\theta} + \varepsilon_t$$

Is this model a GLM?

$$Y_t = \theta_0 + \theta_1 z_t + \varepsilon_t$$

Yes, what about this one?

$$Y_t = \theta_0 + \theta_1 z_t + \theta_2 z_t^2 + \varepsilon_t$$

Also yes, since it can be written as

$$y_t = \begin{pmatrix} 1 & z_t & z_t^2 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix} + \varepsilon_t$$

The General Linear Model (GLM)

$$Y_t = \boldsymbol{x}_t^T \boldsymbol{\theta} + \varepsilon_t$$

Is this model a GLM?

$$Y_t = \theta_0 + \theta_1 z_t + \varepsilon_t$$

Yes, what about this one?

$$Y_t = \theta_0 + \theta_1 z_t + \theta_2 z_t^2 + \varepsilon_t$$

Also yes, since it can be written as

$$y_t = \begin{pmatrix} 1 & z_t & z_t^2 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix} + \varepsilon_t$$

It is linearity in the parameters that matters!

OLS-estimates

- Non-linear regression: Numerical optimization is required; see the book for a simple example (Newton-Raphson)

OLS-estimates

- Non-linear regression: Numerical optimization is required; see the book for a simple example (Newton-Raphson)
- For the general linear model a closed-form solution exists.

OLS-estimates

- Non-linear regression: Numerical optimization is required; see the book for a simple example (Newton-Raphson)
- For the general linear model a closed-form solution exists.
  For all observations the model equations are written as:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad or \quad \boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

OLS-estimates

- Non-linear regression: Numerical optimization is required; see the book for a simple example (Newton-Raphson)
- For the general linear model a closed-form solution exists.
  For all observations the model equations are written as:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad or \quad \boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

  We minimise $S(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})^T(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})$, by solving $\frac{\partial}{\partial\widehat{\boldsymbol{\theta}}}S(\widehat{\boldsymbol{\theta}}) = 0$.
- The solution is $\boxed{\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}}$ (if $\boldsymbol{x}$ has full rank).

OLS-estimates

- Non-linear regression: Numerical optimization is required; see the book for a simple example (Newton-Raphson)
- For the general linear model a closed-form solution exists.
  For all observations the model equations are written as:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad or \quad \boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

  We minimise $S(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})^T(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})$, by solving $\frac{\partial}{\partial\widehat{\boldsymbol{\theta}}}S(\widehat{\boldsymbol{\theta}}) = 0$.
- The solution is $\boxed{\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}}$ (if $\boldsymbol{x}$ has full rank). What if $\boldsymbol{x}$ does not have full rank?

OLS-estimates

- Non-linear regression: Numerical optimization is required; see the book for a simple example (Newton-Raphson)
- For the general linear model a closed-form solution exists.
  For all observations the model equations are written as:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad or \quad \boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

  We minimise $S(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})^T(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})$, by solving $\frac{\partial}{\partial\widehat{\boldsymbol{\theta}}}S(\widehat{\boldsymbol{\theta}}) = 0$.
- The solution is $\boxed{\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}}$ (if $\boldsymbol{x}$ has full rank). What if $\boldsymbol{x}$ does not have full rank?
- $\widehat{\sigma}^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}/(n-p)$ and $V[\widehat{\boldsymbol{\theta}}] = \widehat{\sigma}^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$.

OLS-estimates

- Non-linear regression: Numerical optimization is required; see the book for a simple example (Newton-Raphson)
- For the general linear model a closed-form solution exists.
  For all observations the model equations are written as:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad or \quad \boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

  We minimise $S(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})^T(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})$, by solving $\frac{\partial}{\partial\widehat{\boldsymbol{\theta}}}S(\widehat{\boldsymbol{\theta}}) = 0$.
- The solution is $\boxed{\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}}$ (if $\boldsymbol{x}$ has full rank). What if $\boldsymbol{x}$ does not have full rank?
- $\widehat{\sigma}^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}/(n-p)$ and $V[\widehat{\boldsymbol{\theta}}] = \widehat{\sigma}^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$.
- Important result: The minimiser of $E[(a - \boldsymbol{Y})^2]$ is $E[\boldsymbol{Y}]$!

OLS-estimates

- Non-linear regression: Numerical optimization is required; see the book for a simple example (Newton-Raphson)
- For the general linear model a closed-form solution exists.
  For all observations the model equations are written as:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad or \quad \boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

  We minimise $S(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})^T(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})$, by solving $\frac{\partial}{\partial\widehat{\boldsymbol{\theta}}}S(\widehat{\boldsymbol{\theta}}) = 0$.
- The solution is $\boxed{\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}}$ (if $\boldsymbol{x}$ has full rank). What if $\boldsymbol{x}$ does not have full rank?
- $\widehat{\sigma}^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}/(n-p)$ and $V[\widehat{\boldsymbol{\theta}}] = \widehat{\sigma}^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$.
- Important result: The minimiser of $E[(a - \boldsymbol{Y})^2]$ is $E[\boldsymbol{Y}]$! What does this say about our predictions in the linear model $\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$? Predicions of the mean!

OLS-estimates

- Non-linear regression: Numerical optimization is required; see the book for a simple example (Newton-Raphson)
- For the general linear model a closed-form solution exists.
  For all observations the model equations are written as:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad or \quad \boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

  We minimise $S(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})^T(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\theta})$, by solving $\frac{\partial}{\partial\widehat{\boldsymbol{\theta}}}S(\widehat{\boldsymbol{\theta}}) = 0$.
- The solution is $\boxed{\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}}$ (if $\boldsymbol{x}$ has full rank). What if $\boldsymbol{x}$ does not have full rank?
- $\widehat{\sigma}^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}/(n-p)$ and $V[\widehat{\boldsymbol{\theta}}] = \widehat{\sigma}^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$.
- Important result: The minimiser of $E[(a - \boldsymbol{Y})^2]$ is $E[\boldsymbol{Y}]$! What does this say about our predictions in the linear model $\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$? Predicions of the mean!
- Implies: Least squares methods always estimate the mean.

Properties of the OLS-estimator of a GLM

- Notice that $\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{Y}$ is a linear combination of the observations.

Properties of the OLS-estimator of a GLM

- Notice that $\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{Y}$ is a linear combination of the observations. When does this imply that $\widehat{\boldsymbol{\theta}}$ is normal distributed?

Properties of the OLS-estimator of a GLM

- Notice that $\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}$ is a linear combination of the observations. When does this imply that $\widehat{\boldsymbol{\theta}}$ is normal distributed?
- Similarly $\widehat{\boldsymbol{Y}}$ is a linear combination of $\widehat{\boldsymbol{\theta}}$.

Properties of the OLS-estimator of a GLM

- Notice that $\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}$ is a linear combination of the observations. When does this imply that $\widehat{\boldsymbol{\theta}}$ is normal distributed?
- Similarly $\widehat{\boldsymbol{Y}}$ is a linear combination of $\widehat{\boldsymbol{\theta}}$.
- It is unbiased, i.e. $E[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$

Properties of the OLS-estimator of a GLM

- Notice that $\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}$ is a linear combination of the observations. When does this imply that $\widehat{\boldsymbol{\theta}}$ is normal distributed?
- Similarly $\widehat{\boldsymbol{Y}}$ is a linear combination of $\widehat{\boldsymbol{\theta}}$.
- It is unbiased, i.e. $E[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$
- $V[\widehat{\boldsymbol{\theta}}] = E\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right] = \sigma^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$

Properties of the OLS-estimator of a GLM

- Notice that $\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}$ is a linear combination of the observations. When does this imply that $\widehat{\boldsymbol{\theta}}$ is normal distributed?
- Similarly $\widehat{\boldsymbol{Y}}$ is a linear combination of $\widehat{\boldsymbol{\theta}}$.
- It is unbiased, i.e. $E[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$
- $V[\widehat{\boldsymbol{\theta}}] = E\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right] = \sigma^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$
- $\hat{\theta}$ is BLUE (Best Linear Unbiased Estimator), which means that it has the smallest variance among all estimators which are a linear function of the observations.

WLS-estimates

In OLS all observations receive the same weight. When does it make sense to put different weight on different observations? In general and for time series in particular.

WLS-estimates

In OLS all observations receive the same weight. When does it make sense to put different weight on different observations? In general and for time series in particular.

- Equation for all observations: $\boldsymbol{Y} = \boldsymbol{x\theta} + \boldsymbol{\varepsilon}$

WLS-estimates

In OLS all observations receive the same weight. When does it make sense to put different weight on different observations? In general and for time series in particular.

- Equation for all observations: $\boldsymbol{Y} = \boldsymbol{x\theta} + \boldsymbol{\varepsilon}$
- $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $V[\boldsymbol{\varepsilon}] = E[\boldsymbol{\varepsilon\varepsilon}^T] = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is known

WLS-estimates

In OLS all observations receive the same weight. When does it make sense to put different weight on different observations? In general and for time series in particular.

- Equation for all observations: $\boldsymbol{Y} = \boldsymbol{x\theta} + \boldsymbol{\varepsilon}$
- $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $V[\boldsymbol{\varepsilon}] = E[\boldsymbol{\varepsilon\varepsilon}^T] = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is known
- We want to minimize $(\boldsymbol{Y} - \boldsymbol{x\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{x\theta})$

WLS-estimates

In OLS all observations receive the same weight. When does it make sense to put different weight on different observations? In general and for time series in particular.

- Equation for all observations: $\boldsymbol{Y} = \boldsymbol{x\theta} + \boldsymbol{\varepsilon}$
- $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $V[\boldsymbol{\varepsilon}] = E[\boldsymbol{\varepsilon\varepsilon}^T] = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is known
- We want to minimize $(\boldsymbol{Y} - \boldsymbol{x\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{x\theta})$ - For OLS:

WLS-estimates

In OLS all observations receive the same weight. When does it make sense to put different weight on different observations? In general and for time series in particular.

- Equation for all observations: $\boldsymbol{Y} = \boldsymbol{x\theta} + \boldsymbol{\varepsilon}$
- $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $V[\boldsymbol{\varepsilon}] = E[\boldsymbol{\varepsilon\varepsilon}^T] = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is known
- We want to minimize $(\boldsymbol{Y} - \boldsymbol{x\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{x\theta})$ - For OLS: $(\boldsymbol{Y} - \boldsymbol{x\theta})^T(\boldsymbol{Y} - \boldsymbol{x\theta})$

WLS-estimates

In OLS all observations receive the same weight. When does it make sense to put different weight on different observations? In general and for time series in particular.

- Equation for all observations: $\boldsymbol{Y} = \boldsymbol{x\theta} + \boldsymbol{\varepsilon}$
- $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $V[\boldsymbol{\varepsilon}] = E[\boldsymbol{\varepsilon\varepsilon}^T] = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is known
- We want to minimize $(\boldsymbol{Y} - \boldsymbol{x\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{x\theta})$ - For OLS: $(\boldsymbol{Y} - \boldsymbol{x\theta})^T(\boldsymbol{Y} - \boldsymbol{x\theta})$
- The solution is

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

(if $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ is invertible)

WLS-estimates

In OLS all observations receive the same weight. When does it make sense to put different weight on different observations? In general and for time series in particular.

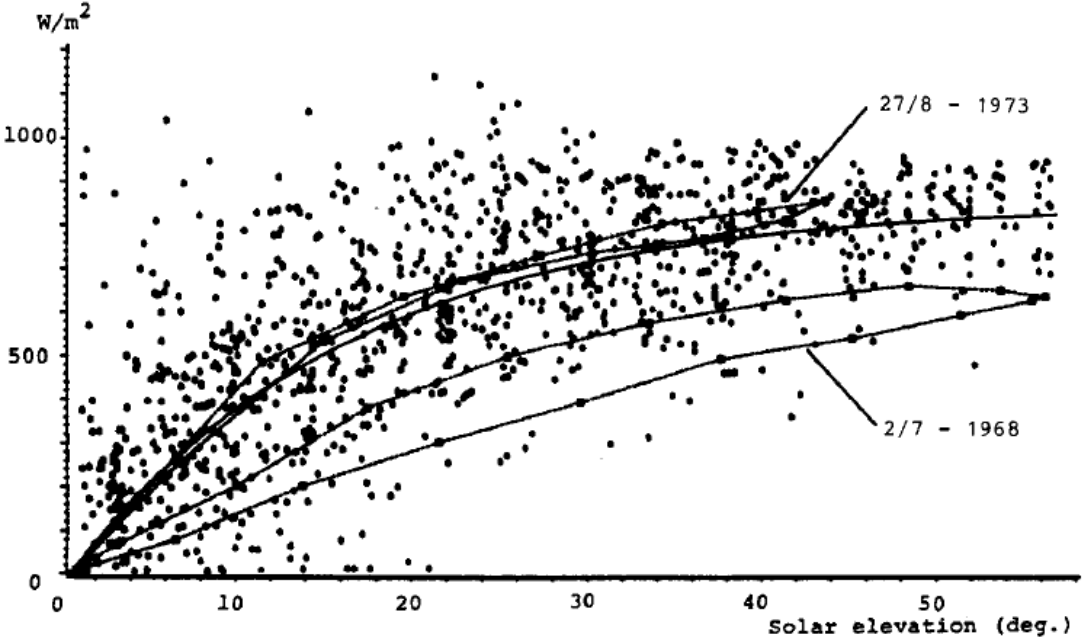
- Equation for all observations: $\boldsymbol{Y} = \boldsymbol{x\theta} + \boldsymbol{\varepsilon}$
- $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $V[\boldsymbol{\varepsilon}] = E[\boldsymbol{\varepsilon\varepsilon}^T] = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is known
- We want to minimize $(\boldsymbol{Y} - \boldsymbol{x\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{x\theta})$ - For OLS: $(\boldsymbol{Y} - \boldsymbol{x\theta})^T(\boldsymbol{Y} - \boldsymbol{x\theta})$
- The solution is

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

  (if $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ is invertible)
- An estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-p}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\theta}})$$

Example WLS/OLS: Clear sky radiation



Example WLS/OLS: Clear sky radiation

$$Y_t^{\text{dir}} = a(1 - \exp(-bh_t)) + \varepsilon_t \tag{1}$$

$$\varepsilon_t \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma}) \tag{2}$$

Example WLS/OLS: Clear sky radiation

$$Y_t^{\text{dir}} = a(1 - \exp(-bh_t)) + \varepsilon_t \tag{1}$$

$$\varepsilon_t \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma}) \tag{2}$$

Actually what we are measuring is Y_t^{glob} and Y_t^{dif} where

$$Y_t^{\text{glob}} = \sin(h_t)\, Y_t^{\text{dir}} + Y_t^{\text{dif}},$$

Example WLS/OLS: Clear sky radiation

$$Y_t^{\text{dir}} = a(1 - \exp(-bh_t)) + \varepsilon_t \tag{1}$$

$$\varepsilon_t \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma}) \tag{2}$$

Actually what we are measuring is Y_t^{glob} and Y_t^{dif} where

$$Y_t^{\text{glob}} = \sin(h_t)\, Y_t^{\text{dir}} + Y_t^{\text{dif}},$$

so

$$Y_t^{\text{dir}} = \frac{Y_t^{\text{glob}} - Y_t^{\text{dif}}}{\sin(h_t)}.$$

Example WLS/OLS: Clear sky radiation

$$Y_t^{\text{dir}} = a(1 - \exp(-bh_t)) + \varepsilon_t \tag{1}$$

$$\varepsilon_t \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma}) \tag{2}$$

Actually what we are measuring is Y_t^{glob} and Y_t^{dif} where

$$Y_t^{\text{glob}} = \sin(h_t)\, Y_t^{\text{dir}} + Y_t^{\text{dif}},$$

so

$$Y_t^{\text{dir}} = \frac{Y_t^{\text{glob}} - Y_t^{\text{dif}}}{\sin(h_t)}.$$

How does this affect the variance of the observations?

Example WLS/OLS: Clear sky radiation

$$Y_t^{\text{dir}} = a(1 - \exp(-bh_t)) + \varepsilon_t \tag{1}$$

$$\varepsilon_t \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma}) \tag{2}$$

Actually what we are measuring is Y_t^{glob} and Y_t^{dif} where

$$Y_t^{\text{glob}} = \sin(h_t)\, Y_t^{\text{dir}} + Y_t^{\text{dif}},$$

so

$$Y_t^{\text{dir}} = \frac{Y_t^{\text{glob}} - Y_t^{\text{dif}}}{\sin(h_t)}.$$

How does this affect the variance of the observations?
Potential Variance structures:

- i.i.d: $\mathbf{\Sigma} = \boldsymbol{I}$

Example WLS/OLS: Clear sky radiation

$$Y_t^{\text{dir}} = a(1 - \exp(-bh_t)) + \varepsilon_t \tag{1}$$

$$\varepsilon_t \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma}) \tag{2}$$

Actually what we are measuring is Y_t^{glob} and Y_t^{dif} where

$$Y_t^{\text{glob}} = \sin(h_t)\, Y_t^{\text{dir}} + Y_t^{\text{dif}},$$

so

$$Y_t^{\text{dir}} = \frac{Y_t^{\text{glob}} - Y_t^{\text{dif}}}{\sin(h_t)}.$$

How does this affect the variance of the observations?
Potential Variance structures:

- i.i.d: $\mathbf{\Sigma} = \boldsymbol{I}$
- Only variance: $\Sigma_{ii} = \dfrac{1}{\sin(h_{t_i})^2}$

Example WLS/OLS: Clear sky radiation

$$Y_t^{\text{dir}} = a(1 - \exp(-bh_t)) + \varepsilon_t \tag{1}$$

$$\varepsilon_t \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma}) \tag{2}$$

Actually what we are measuring is Y_t^{glob} and Y_t^{dif} where

$$Y_t^{\text{glob}} = \sin(h_t)\, Y_t^{\text{dir}} + Y_t^{\text{dif}},$$

so

$$Y_t^{\text{dir}} = \frac{Y_t^{\text{glob}} - Y_t^{\text{dif}}}{\sin(h_t)}.$$

How does this affect the variance of the observations?
Potential Variance structures:

- i.i.d: $\mathbf{\Sigma} = \boldsymbol{I}$
- Only variance: $\Sigma_{ii} = \dfrac{1}{\sin(h_{t_i})^2}$
- Only correlation: $\Sigma_{ij} = \rho^{|t_i - t_j|}$

Example WLS/OLS: Clear sky radiation

$$Y_t^{\text{dir}} = a(1 - \exp(-bh_t)) + \varepsilon_t \tag{1}$$

$$\varepsilon_t \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma}) \tag{2}$$

Actually what we are measuring is Y_t^{glob} and Y_t^{dif} where

$$Y_t^{\text{glob}} = \sin(h_t)\, Y_t^{\text{dir}} + Y_t^{\text{dif}},$$

so

$$Y_t^{\text{dir}} = \frac{Y_t^{\text{glob}} - Y_t^{\text{dif}}}{\sin(h_t)}.$$

How does this affect the variance of the observations?
Potential Variance structures:

- i.i.d: $\mathbf{\Sigma} = \boldsymbol{I}$
- Only variance: $\Sigma_{ii} = \dfrac{1}{\sin(h_{t_i})^2}$
- Only correlation: $\Sigma_{ij} = \rho^{|t_i - t_j|}$
- Both correlation and variance: $\Sigma_{ij} = \dfrac{\rho^{|t_i - t_j|}}{\sin(h_{t_i})\sin(h_{t_j})}$

Example WLS/OLS: Solar radiation

Structure	$\hat{a}_N$	$\hat{b}_N$	$\hat{\sigma}_N^2$

Numbers in parenthesis are standard deviations.

Example WLS/OLS: Solar radiation

Structure	$\hat{a}_N$	$\hat{b}_N$	$\hat{\sigma}^2_N$
Identity	798.6 (13.1)	0.0798 (0.0043)	34631.8

Numbers in parenthesis are standard deviations.

Example WLS/OLS: Solar radiation

Structure	$\hat{a}_N$	$\hat{b}_N$	$\hat{\sigma}^2_N$
Identity	798.6 (13.1)	0.0798 (0.0043)	34631.8
Variance	822.6 (5.2)	0.0706 (0.0012)	3947.6

Numbers in parenthesis are standard deviations.

Example WLS/OLS: Solar radiation

Structure	$\hat{a}_N$	$\hat{b}_N$	$\hat{\sigma}^2_N$
Identity	798.6 (13.1)	0.0798 (0.0043)	34631.8
Variance	822.6 (5.2)	0.0706 (0.0012)	3947.6
Correlation	827.8 (7.3)	0.0551 (0.0022)	25180.1

Numbers in parenthesis are standard deviations.

Example WLS/OLS: Solar radiation

Structure	$\hat{a}_N$	$\hat{b}_N$	$\hat{\sigma}^2_N$
Identity	798.6 (13.1)	0.0798 (0.0043)	34631.8
Variance	822.6 (5.2)	0.0706 (0.0012)	3947.6
Correlation	827.8 (7.3)	0.0551 (0.0022)	25180.1
Var. and Cor.	842.3 (7.4)	0.0614 (0.0013)	2302.1

Numbers in parenthesis are standard deviations.

Maximum Likelihood (ML)

The likelihood is defined by the joint probability function of the data

$$L(\mu, \sigma) \equiv p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N | \mu, \sigma)$$

Hence, it's a function of the two parameters (the sample is observed, so it is not varying).

Maximum Likelihood (ML)

The likelihood is defined by the joint probability function of the data

$$L(\mu, \sigma) \equiv p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N | \mu, \sigma)$$

Hence, it's a function of the two parameters (the sample is observed, so it is not varying).

Due to independence of observations:

$$= \prod_{i=1}^{N} p(\boldsymbol{y}_i | \mu, \sigma)$$

Maximum Likelihood (ML)

The likelihood is defined by the joint probability function of the data

$$L(\mu, \sigma) \equiv p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N | \mu, \sigma)$$

Hence, it's a function of the two parameters (the sample is observed, so it is not varying).

Due to independence of observations:

$$= \prod_{i=1}^{N} p(\boldsymbol{y}_i | \mu, \sigma)$$

If observations are correlated in time?

$$\begin{aligned} L(\boldsymbol{\theta}) &= p(\mathcal{Y}_N | \boldsymbol{\theta}) = p(\boldsymbol{y}_N, \boldsymbol{y}_{N-1}, \ldots, \boldsymbol{y}_0 | \boldsymbol{\theta}) \\ &= \left(\prod_{k=1}^{N} p(\boldsymbol{y}_k | \mathcal{Y}_{k-1}, \boldsymbol{\theta}) \right) p(\boldsymbol{y}_0 | \boldsymbol{\theta}) \end{aligned}$$

where $\mathcal{Y}_k = \{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_k\}$.

Maximum Likelihood (ML) - estimates

- Assume that the observations are Gaussian;

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$$

Maximum Likelihood (ML) - estimates

- Assume that the observations are Gaussian;

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$$

- and that $\boldsymbol{\Sigma}$ is known.

Maximum Likelihood (ML) - estimates

- Assume that the observations are Gaussian;

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$$

- and that $\boldsymbol{\Sigma}$ is known.
- The ML-estimator is (here) the same as the WLS-estimator:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

Maximum Likelihood (ML) - estimates

- Assume that the observations are Gaussian;

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$$

- and that $\boldsymbol{\Sigma}$ is known.
- The ML-estimator is (here) the same as the WLS-estimator:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

- The ML-estimator for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\theta}})$$

Maximum Likelihood (ML) - estimates

- Assume that the observations are Gaussian;

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$$

- and that $\boldsymbol{\Sigma}$ is known.
- The ML-estimator is (here) the same as the WLS-estimator:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

- The ML-estimator for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\theta}})$$

- OLS and WLS estimates can be interpreted as?

Maximum Likelihood (ML) - estimates

- Assume that the observations are Gaussian;

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$$

- and that $\boldsymbol{\Sigma}$ is known.
- The ML-estimator is (here) the same as the WLS-estimator:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

- The ML-estimator for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\theta}})$$

- OLS and WLS estimates can be interpreted as? assumptions of Gaussianity.

Maximum Likelihood (ML) - estimates

- Assume that the observations are Gaussian;

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$$

- and that $\boldsymbol{\Sigma}$ is known.
- The ML-estimator is (here) the same as the WLS-estimator:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

- The ML-estimator for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\theta}})$$

- OLS and WLS estimates can be interpreted as? assumptions of Gaussianity.

Let's take a look at the `example_likelihood.R` script

Properties of the ML-estimator

- It is a linear function of the observations which implies that it is normally distributed.
- It is unbiased, i.e. $E[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$.
- The variance $V[\widehat{\boldsymbol{\theta}}] = E[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] = (\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x})^{-1} \sigma^2$.
- It is an efficient estimator (minimum variance of unbiased estimators).

Prediction in the general linear model

Theorem 3.8 and 3.9

- If the expected value of the squared prediction error is to be minimized, then
- the expected mean $E[Y|\boldsymbol{X} = \boldsymbol{x}]$ is the optimal predictor.

Prediction in the general linear model

Theorem 3.8 and 3.9

- If the expected value of the squared prediction error is to be minimized, then
- the expected mean $E[Y|\boldsymbol{X} = \boldsymbol{x}]$ is the optimal predictor.

- Known parameters:

$$\widehat{Y}_t = E_\theta[Y_t|\boldsymbol{X}_t = \boldsymbol{x}_t] = \boldsymbol{x}_t^T\boldsymbol{\theta}$$
$$V_\theta[Y_t - \widehat{Y}_t] = V_\theta[\varepsilon_t] = \sigma^2$$

Prediction in the general linear model

Theorem 3.8 and 3.9

- If the expected value of the squared prediction error is to be minimized, then
- the expected mean $E[Y|\boldsymbol{X} = \boldsymbol{x}]$ is the optimal predictor.

- Known parameters:

$$\widehat{Y}_t = E_\theta[Y_t|\boldsymbol{X}_t = \boldsymbol{x}_t] = \boldsymbol{x}_t^T\boldsymbol{\theta}$$
$$V_\theta[Y_t - \widehat{Y}_t] = V_\theta[\varepsilon_t] = \sigma^2$$

- Estimated parameters:

Prediction in the general linear model

Theorem 3.8 and 3.9

- If the expected value of the squared prediction error is to be minimized, then
- the expected mean $E[Y|\boldsymbol{X} = \boldsymbol{x}]$ is the optimal predictor.

- Known parameters:

$$\widehat{Y}_t = E_\theta[Y_t|\boldsymbol{X}_t = \boldsymbol{x}_t] = \boldsymbol{x}_t^T\boldsymbol{\theta}$$
$$V_\theta[Y_t - \widehat{Y}_t] = V_\theta[\varepsilon_t] = \sigma^2$$

- Estimated parameters:

$$\widehat{Y}_t = E_{\widehat{\theta}}[Y_t|\boldsymbol{X}_t = \boldsymbol{x}_t] = \boldsymbol{x}_t^T\widehat{\boldsymbol{\theta}}$$

Prediction in the general linear model

Theorem 3.8 and 3.9

- If the expected value of the squared prediction error is to be minimized, then
- the expected mean $E[Y|\boldsymbol{X} = \boldsymbol{x}]$ is the optimal predictor.

- Known parameters:

$$\widehat{Y}_t = E_\theta[Y_t|\boldsymbol{X}_t = \boldsymbol{x}_t] = \boldsymbol{x}_t^T\boldsymbol{\theta}$$
$$V_\theta[Y_t - \widehat{Y}_t] = V_\theta[\varepsilon_t] = \sigma^2$$

- Estimated parameters:

$$\widehat{Y}_t = E_{\widehat{\theta}}[Y_t|\boldsymbol{X}_t = \boldsymbol{x}_t] = \boldsymbol{x}_t^T\widehat{\boldsymbol{\theta}}$$

$$V_{\widehat{\theta}}[Y_t - \widehat{Y}_t] = V_{\widehat{\theta}}[\varepsilon_t + \boldsymbol{x}_t^T(\theta - \widehat{\theta})] = \widehat{\sigma}^2[1 + \boldsymbol{x}_t^T(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}_t]$$

Model selection and the bias-variance tradeoff

We want to find a *suitable model* – the model which is neither too simple (under-fitted) nor too complex (over-fitted):

Model selection and the bias-variance tradeoff

We want to find a *suitable model* – the model which is neither too simple (under-fitted) nor too complex (over-fitted):

- Divide the data into a training set and a test set

Model selection and the bias-variance tradeoff

We want to find a *suitable model* – the model which is neither too simple (under-fitted) nor too complex (over-fitted):

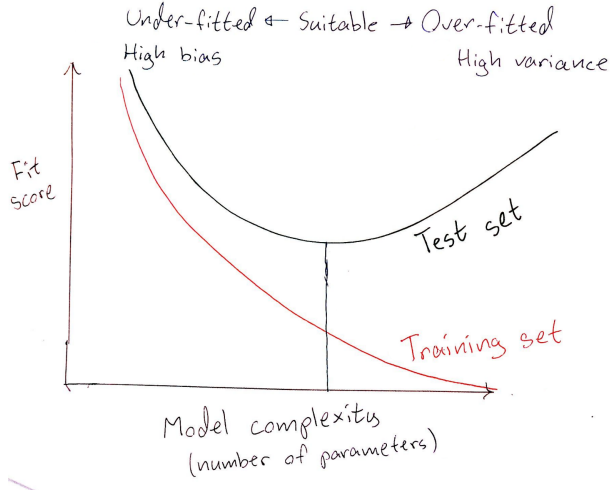
- Divide the data into a training set and a test set
- Define a fit score (smaller is better e.g. summed squared error)

Model selection and the bias-variance tradeoff

We want to find a *suitable model* – the model which is neither too simple (under-fitted) nor too complex (over-fitted):

- Divide the data into a training set and a test set
- Define a fit score (smaller is better e.g. summed squared error)

How do we find the balance?



Model selection

How do we find the balance?

Model selection

How do we find the balance?

- Information criteria (AIC, BIC)
- Goodness of fit test (likelihood-ratio test, F-test)
- Cross-validation technique (n-fold CV)

Model selection

How do we find the balance?

- Information criteria (AIC, BIC)
- Goodness of fit test (likelihood-ratio test, F-test)
- Cross-validation technique (n-fold CV)

Model selection procedure

- Forward selection (start with a small model and extend)

Model selection

How do we find the balance?

- Information criteria (AIC, BIC)
- Goodness of fit test (likelihood-ratio test, F-test)
- Cross-validation technique (n-fold CV)

Model selection procedure

- Forward selection (start with a small model and extend)
- Backward selection (start with a large model and remove)

Model validation (use also while finding a model)

After fitting the model then analyse the residuals

If no patterns are left, hence we have white noise residual, then we are done!

Model validation (use also while finding a model)

After fitting the model then analyse the residuals

If no patterns are left, hence we have white noise residual, then we are done!

Residuals analysis

Plot, plot, plot!

Model validation (use also while finding a model)

After fitting the model then analyse the residuals

If no patterns are left, hence we have white noise residual, then we are done!

Residuals analysis

Plot, plot, plot!

- Time series plots of residuals aligned with input series
- Scatter plots of residuals vs. inputs
- ACF and CCF (We will get back to them!)

Model validation (use also while finding a model)

After fitting the model then analyse the residuals

If no patterns are left, hence we have white noise residual, then we are done!

Residuals analysis

Plot, plot, plot!

- Time series plots of residuals aligned with input series
- Scatter plots of residuals vs. inputs
- ACF and CCF (We will get back to them!)

Forward selection

Fit a simple model, analyse the residuals: Can you see some patterns left related to some inputs? Improve the model and repeat...

Model validation (use also while finding a model)

After fitting the model then analyse the residuals

If no patterns are left, hence we have white noise residual, then we are done!

Residuals analysis

Plot, plot, plot!

- Time series plots of residuals aligned with input series
- Scatter plots of residuals vs. inputs
- ACF and CCF (We will get back to them!)

Forward selection

Fit a simple model, analyse the residuals: Can you see some patterns left related to some inputs? Improve the model and repeat...

- Good approach for modelling with new data
- Good approach for articles/reports (you get a story)

Training and Test Set Techniques for Time Series

- When data is time dependent, we can't just select some random points as test set and for cross-validation
- We have to take time into account

Training and Test Set Techniques for Time Series

- When data is time dependent, we can't just select some random points as test set and for cross-validation
- We have to take time into account
- **Holdout Method:**
  - Split data into training and test sets
  - **Illustration:**

Training and Test Set Techniques for Time Series

- When data is time dependent, we can't just select some random points as test set and for cross-validation
- We have to take time into account
- **Holdout Method:**
  - Split data into training and test sets
  - **Illustration:**

  - Problems often occur: if the underlying process changed from training to test set

Training and Test Set Techniques for Time Series

- When data is time dependent, we can't just select some random points as test set and for cross-validation
- We have to take time into account
- **Holdout Method:**
  - Split data into training and test sets
  - **Illustration:**

  - Problems often occur: if the underlying process changed from training to test set
- **Cross-Validation:**
  - K-fold, Leave-One-Out
  - **Illustration:**

  - This is often a good compromise (a burn-in period can be needed).

Rolling Horizon Test Sets

- **Concept:** Use a moving window to create multiple training and test sets

Rolling Horizon Test Sets

- **Concept:** Use a moving window to create multiple training and test sets
- **Illustration:**

Rolling Horizon Test Sets

- **Concept:** Use a moving window to create multiple training and test sets
- **Illustration:**

- **Steps:**
  1. Start with an initial training set
  2. Train the model
  3. Test on the next time point
  4. Expand the training set to include the test point
  5. Repeat the process

Rolling Horizon Test Sets

- **Concept:** Use a moving window to create multiple training and test sets
- **Illustration:**

- **Steps:**
  1. Start with an initial training set
  2. Train the model
  3. Test on the next time point
  4. Expand the training set to include the test point
  5. Repeat the process
- A burn-in period is needed
- Recursive estimation is really nice, it always does this!