Multiple linear regression

model description and application

Markus Kunze

Institut für Meteorologie
Freie Universität Berlin

Ausgewählte Probleme der mittleren Atmosphäre,
WS 2009/10

Outline

Outline

Outline

Multiple linear regression model I

- Model: the response y_t may be related to $k + 1$ regressor variables.

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t$$

$$y_t = \beta_0 + \sum_{j=1}^{k} \beta_j x_{tj} + \varepsilon_t, \quad t = 1, 2, \ldots, n \tag{1}$$

Multiple linear regression model I

- Model: the response y_t may be related to $k+1$ regressor variables.

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t$$

$$y_t = \beta_0 + \sum_{j=1}^{k} \beta_j x_{tj} + \varepsilon_t, \quad t = 1, 2, \ldots, n \tag{1}$$

- Task: find the unknown regression coefficients $\beta_j, \quad j = 0, 1, \ldots, k$. The number of observations n must be greater than the number of unknown regressor variables: $n > k + 1$.

Multiple linear regression model I

- Model: the response y_t may be related to $k+1$ regressor variables.

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t$$

$$y_t = \beta_0 + \sum_{j=1}^{k} \beta_j x_{tj} + \varepsilon_t, \quad t = 1, 2, \ldots, n \tag{1}$$

- Task: find the unknown regression coefficients $\beta_j, \quad j = 0, 1, \ldots, k$. The number of observations n must be greater than the number of unknown regressor variables: $n > k + 1$.
- The model is a linear model because it is a linear function of the regression coefficients.

Multiple linear regression model II

- Matrix notation of the multiple linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon \tag{2}$$

Multiple linear regression model II

- Matrix notation of the multiple linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon \tag{2}$$

- with the following matrices (n: number of observations, $p = k + 1$: number of regressor variables (basis functions)):

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, n \times 1 \text{ Matrix}; \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, n \times p \text{ Matrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, p \times 1 \text{ Matrix}; \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, n \times 1 \text{ Matrix}$$

The least-squares function

- Like for simple linear regression, one looks for the parameters that minimises the least squares function $S(\beta_0, \beta_1, \ldots, \beta_k)$.

The least-squares function

- Like for simple linear regression, one looks for the parameters that minimises the least squares function $S(\beta_0, \beta_1, \ldots, \beta_k)$.
-

$$\begin{aligned} S(\beta_0, \beta_1, \ldots, \beta_k) &= \sum_{t=1}^{n} \varepsilon_t^2 \\ &= \sum_{t=1}^{n} \left(y_t - \beta_0 - \sum_{j=1}^{k} \beta_j x_{tj} \right)^2 \end{aligned} \tag{3}$$

The normal equations

- $k + 1$ partial derivations of the least-squares function lead to $k + 1$ normal equations:

The normal equations

- $k+1$ partial derivations of the least-squares function lead to $k+1$ normal equations:
-

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_t - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{tj}) = 0$$

$$\frac{\partial S}{\partial \beta_j} = -2\sum_{t=1}^{n}(y_t - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{tj})x_{tj} = 0, \quad j = 1, 2, \ldots, k$$

The normal equations II

- The least squares equation in matrix notation:

$$
\begin{aligned}
S(\beta_0, \beta_1, \ldots, \beta_k) &= \sum_{t=1}^{n} \varepsilon_t^2 \\
S(\boldsymbol{\beta}) &= \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} \\
&= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} \qquad (4)
\end{aligned}
$$

The normal equations II

- The least squares equation in matrix notation:

$$
\begin{aligned}
S(\beta_0, \beta_1, \ldots, \beta_k) &= \sum_{t=1}^{n} \varepsilon_t^2 \\
S(\boldsymbol{\beta}) &= \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} \\
&= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} \qquad (4)
\end{aligned}
$$

- The least-squares normal equations in matrix notation:

$$
\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{X}\hat{\boldsymbol{\beta}} = 0 \qquad (5)
$$

$$
\boldsymbol{X}^T \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{y} \qquad (6)
$$

The solutions for the normal equations

- The least squares estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{7}$$

The solutions for the normal equations

- The least squares estimator of β:

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{7}$$

- The solution can be found, when the covariance Matrix $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ exists.

The solutions for the normal equations

- The least squares estimator of β:

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{7}$$

- The solution can be found, when the covariance Matrix $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ exists.
- That is the case, if the regressor variables x_k are linear independent. That means no column of the Matrix $\boldsymbol{X}$ can be created as a linear combination of the other columns.

Outline

Autocorrelated errors I

- Fundamental assumptions in linear regression:

$$\begin{aligned} E(\varepsilon_t) &= 0, && \varepsilon_t \text{ have zero mean.} \\ Var(\varepsilon_t) &= \sigma^2, && \varepsilon_t \text{ have constant variance.} \\ E(\varepsilon_t \varepsilon_{t-1}) &= 0, && \varepsilon_t \text{ are uncorrelated.} \end{aligned}$$

Autocorrelated errors I

- Fundamental assumptions in linear regression:

$$\begin{aligned} E(\varepsilon_t) &= 0, && \varepsilon_t \text{ have zero mean.} \\ Var(\varepsilon_t) &= \sigma^2, && \varepsilon_t \text{ have constant variance.} \\ E(\varepsilon_t \varepsilon_{t-1}) &= 0, && \varepsilon_t \text{ are uncorrelated.} \end{aligned}$$

- The assumption of uncorrelated or independent errors for time series data is often not appropriate:

$$E(\varepsilon_t \varepsilon_{t-1}) \neq 0. \tag{8}$$

Autocorrelated errors I

- Fundamental assumptions in linear regression:

$$\begin{aligned} E(\varepsilon_t) &= 0, && \varepsilon_t \text{ have zero mean.} \\ Var(\varepsilon_t) &= \sigma^2, && \varepsilon_t \text{ have constant variance.} \\ E(\varepsilon_t \varepsilon_{t-1}) &= 0, && \varepsilon_t \text{ are uncorrelated.} \end{aligned}$$

- The assumption of uncorrelated or independent errors for time series data is often not appropriate:

$$E(\varepsilon_t \varepsilon_{t-1}) \neq 0. \tag{8}$$

- Sources of autocorrelation:
  - The regression model is not complete. One or more important regressors are not included.

Autocorrelated errors II: effects of autocorrelated errors

- If all assumptions are valid, the estimated regression coefficients are unbiased, efficient, and consistent.

Autocorrelated errors II: effects of autocorrelated errors

- If all assumptions are valid, the estimated regression coefficients are unbiased, efficient, and consistent.
- If not, the regression coefficients are no longer minimum variance estimates: estimates are inefficient.

Autocorrelated errors II: effects of autocorrelated errors

- If all assumptions are valid, the estimated regression coefficients are unbiased, efficient, and consistent.
- If not, the regression coefficients are no longer minimum variance estimates: estimates are inefficient.
- The residual mean square may seriously underestimate σ^2.
    - The standard errors of the regression coefficients may be too small.

Autocorrelated errors II: effects of autocorrelated errors

- If all assumptions are valid, the estimated regression coefficients are unbiased, efficient, and consistent.
- If not, the regression coefficients are no longer minimum variance estimates: estimates are inefficient.
- The residual mean square may seriously underestimate σ^2.
  - The standard errors of the regression coefficients may be too small.
  - The confidence intervals are shorter than they really should be.

Autocorrelated errors II: effects of autocorrelated errors

- If all assumptions are valid, the estimated regression coefficients are unbiased, efficient, and consistent.
- If not, the regression coefficients are no longer minimum variance estimates: estimates are inefficient.
- The residual mean square may seriously underestimate σ^2.
  - The standard errors of the regression coefficients may be too small.
  - The confidence intervals are shorter than they really should be.
  - Misleading test statistics: indicating significance for insignificant results.

Autoregressive model

- If there are autocorrelations in the residuals, the linear regression model has to be transformed.

Autoregressive model

- If there are autocorrelations in the residuals, the linear regression model has to be transformed.
- An autoregressive model is applied to estimate the degree of autocorrelation.

Autoregressive model

- If there are autocorrelations in the residuals, the linear regression model has to be transformed.
- An autoregressive model is applied to estimate the degree of autocorrelation.
- Second-order autoregressive model for the residuals ε_t at time t:

$$\varepsilon_t = \rho_1 \, \varepsilon_{t-1} + \rho_2 \, \varepsilon_{t-2} + a_t, \tag{9}$$

where a_t is a random variable, and ρ_1 and ρ_2 are the autocorrelation parameter.

$\rho_1 + \rho_2 < 1$

$\rho_2 - \rho_1 < 1$

$-1 < \rho_2 < 1$

Autoregressive model

- If there are autocorrelations in the residuals, the linear regression model has to be transformed.
- An autoregressive model is applied to estimate the degree of autocorrelation.
- Second-order autoregressive model for the residuals ε_t at time t:

$$\varepsilon_t = \rho_1 \, \varepsilon_{t-1} + \rho_2 \, \varepsilon_{t-2} + a_t, \tag{9}$$

  where a_t is a random variable, and ρ_1 and ρ_2 are the autocorrelation parameter.

  $\rho_1 + \rho_2 < 1$

  $\rho_2 - \rho_1 < 1$

  $-1 < \rho_2 < 1$

- ρ_1 and ρ_2 are used to transform the model according to Tiao et al. (1990).

Transformation of the model

(see Tiao et al. 1990, Appendix A)

- Transformation of the response variable:

$$y'_t = y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} \tag{10}$$

Transformation of the model

(see Tiao et al. 1990, Appendix A)

- Transformation of the response variable:

$$y'_t = y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} \tag{10}$$

- The transformed model (for a simple example):

$$\begin{aligned} y'_t &= \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \varepsilon_t - \\ &\quad \rho_1(\beta_0 + \beta_1 x_{(t-1)1} + \beta_2 x_{(t-1)2} + \varepsilon_{t-1}) - \\ &\quad \rho_2(\beta_0 + \beta_1 x_{(t-2)1} + \beta_2 x_{(t-2)2} + \varepsilon_{t-2}) \\ &= \beta_0(1 - \rho_1 - \rho_2) + \beta_1(x_{t1} - \rho_1 x_{(t-1)1} - \rho_2 x_{(t-2)1}) + \\ &\quad \beta_2(x_{t2} - \rho_1 x_{(t-1)2} - \rho_2 x_{(t-2)2}) + \varepsilon_t - \rho_1 \varepsilon_{t-1} - \rho_2 \varepsilon_{t-2} \end{aligned}$$

$$y'_t = \beta'_0 + \beta'_1 x'_{t1} + \beta'_2 x'_{t2} + a_t \tag{11}$$

Transformation of the model II

- The uncertainties:
  For the first run of the least square regression (lsqr) the uncertainties are set to 1: $\sigma = 1$. They are used prior to the lsqr to normalise the Matrix $\boldsymbol{X}$ and the response variable $\boldsymbol{y}$.

$$
\begin{aligned}
X_{tj} &= \frac{x_{tj}}{\sigma_t} \\
Y_t &= \frac{y_t}{\sigma_t}
\end{aligned}
$$

Transformation of the model II

- The uncertainties:
  For the first run of the least square regression (lsqr) the uncertainties are set to 1: $\sigma = 1$. They are used prior to the lsqr to normalise the Matrix $\boldsymbol{X}$ and the response variable $\boldsymbol{y}$.

$$X_{tj} = \frac{x_{tj}}{\sigma_t}$$
$$Y_t = \frac{y_t}{\sigma_t}$$

- Update the uncertainties (Box and Jenkins, 1970):

$$\sigma_t = \sqrt{\left(\frac{1-\rho_2}{1+\rho_2}\right) \frac{\sigma_\varepsilon^2}{[(1-\rho_2)^2 - \rho_1^2]}} \tag{12}$$

  where σ_ε, the standard deviation of the residuals, is derived for each individual month.

Transformation of the model II

- The uncertainties:
  For the first run of the least square regression (lsqr) the uncertainties are set to 1: $\sigma = 1$. They are used prior to the lsqr to normalise the Matrix $\boldsymbol{X}$ and the response variable $\boldsymbol{y}$.

$$X_{tj} = \frac{x_{tj}}{\sigma_t}$$
$$Y_t = \frac{y_t}{\sigma_t}$$

- Update the uncertainties (Box and Jenkins, 1970):

$$\sigma_t = \sqrt{\left(\frac{1-\rho_2}{1+\rho_2}\right) \frac{\sigma_\varepsilon^2}{[(1-\rho_2)^2 - \rho_1^2]}} \tag{12}$$

  where σ_ε, the standard deviation of the residuals, is derived for each individual month.

- After the transformation the lsqr is run again.

Outline

The uncertainties of the regression coefficients

- The uncertainties of the regression parameters are given with the diagonal elements of the covariance matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \boldsymbol{C} = \begin{pmatrix} C_{11} & \dots & \dots & \dots \\ \dots & C_{22} & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \dots & C_{kk} \end{pmatrix}$$

- The values for $\sqrt{C_{jj}}$ are stored an extra output file.

The t test statistics

- With a Student's t test the following Hypotheses H_0 and H_1 are tested:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

  If H_0 is rejected, basis function x_j has a significant influence on $\boldsymbol{y}$.

- The t test statistics are calculated from the uncertainties and the estimated regression coefficients:

$$t_{0j} = \frac{\hat{\beta}_j}{\sqrt{C_{jj}}}$$

Outline

The linear regression model

- A variable at time t, y_t can be modelled in the following way:

$$
\begin{aligned}
y(t) \quad = \quad & \beta_{offs} \quad * offset_{(N=m)} + \\
& \beta_{tr} \quad * trend(t)_{(N=m)} + \\
& \beta_{qbo} \quad * QBO(t)_{(N=m)} + \\
& \beta_{qbo\_or} * QBO\_orthog(t)_{(N=m)} + \\
& \beta_{sfl} \quad * solar(t)_{(N=0)} + \\
& \beta_{ens} \quad * ENSO(t)_{(N=n)} + \\
& \beta_{vol} \quad * Volcano(t)_{(N=m)} + \\
& \cdots + \\
& \varepsilon(t), \quad t = 1, n
\end{aligned}
\tag{13}
$$

The linear regression model

- A variable at time t, y_t can be modelled in the following way:

$$
\begin{aligned}
y(t) \quad = \quad & \beta_{offs} \quad * offset_{(N=m)} + \\
& \beta_{tr} \quad * trend(t)_{(N=m)} + \\
& \beta_{qbo} \quad * QBO(t)_{(N=m)} + \\
& \beta_{qbo\_or} * QBO\_orthog(t)_{(N=m)} + \\
& \beta_{sfl} \quad * solar(t)_{(N=0)} + \\
& \beta_{ens} \quad * ENSO(t)_{(N=n)} + \\
& \beta_{vol} \quad * Volcano(t)_{(N=m)} + \\
& \cdots + \\
& \varepsilon(t), \quad t = 1, n
\end{aligned}
\tag{13}
$$

- The model can easily be expanded with more basis functions.

The seasonal variability

- The basis functions can be expanded into $N = m$ pairs of sine and cosine functions to account for the seasonal variability:

$$\beta_j\ x_{tj} = \beta_{j0}\ x_{tj} \quad + \quad \sum_{k=1}^{m} [\ \beta_{j(2k-1)}\ sin(2\pi kt/365.25) + \beta_{j(2k)}\ cos(2\pi kt/365.25)\ ]\ x_{tj}$$

The seasonal variability

- The basis functions can be expanded into $N = m$ pairs of sine and cosine functions to account for the seasonal variability:

$$\beta_j\ x_{tj} = \beta_{j0}\ x_{tj} \quad + \quad \sum_{k=1}^{m} [\ \beta_{j(2k-1)}\ sin(2\pi kt/365.25) + \beta_{j(2k)}\ cos(2\pi kt/365.25)\ ]\ x_{tj}$$

- For example the trend with $m = 1$:

$$\beta_{tr}\ tr_{tj} = \beta_{tr0}\ tr_{tj} \quad + \quad [\ \beta_{tr1}\ sin\ (2\pi t/365.25) + \beta_{tr2}\ cos(2\pi t/365.25)\ ]\ tr_{tj}$$

Outline

Treatment of the basis functions

- If a basis function has a mean and a trend > 0:
  - the offset basis function is not accounting for all of the offset,
  - the trend basis function is not accounting for all of the trend.

Treatment of the basis functions

- If a basis function has a mean and a trend > 0:
  - the offset basis function is not accounting for all of the offset,
  - the trend basis function is not accounting for all of the trend.
- The mean or the mean and the trend can be removed:

$$
\begin{aligned}
x_t &= x_t - \overline{x}, && \text{TREATMENT='RemoveMean'},\\
x_t &= x_t - b\,t, && \text{TREATMENT='RemoveTrend'}\\
x_t &= x_t - b\,t - a, && \text{TREATMENT='RemoveTrendAndMean'}
\end{aligned}
$$

Treatment of the basis functions

- If a basis function has a mean and a trend > 0:
  - the offset basis function is not accounting for all of the offset,
  - the trend basis function is not accounting for all of the trend.
- The mean or the mean and the trend can be removed:

$$
\begin{aligned}
x_t &= x_t - \overline{x}, && \text{TREATMENT='RemoveMean'},\\
x_t &= x_t - b\,t, && \text{TREATMENT='RemoveTrend'}\\
x_t &= x_t - b\,t - a, && \text{TREATMENT='RemoveTrendAndMean'}
\end{aligned}
$$

- Remove the seasonal cycle:

$$
x_t = x_t - \overline{x}_{t,ltm}, \quad \text{TREATMENT='Deseason'}
$$

Treatment of the basis functions

- If a basis function has a mean and a trend > 0:
  - the offset basis function is not accounting for all of the offset,
  - the trend basis function is not accounting for all of the trend.
- The mean or the mean and the trend can be removed:

$$
\begin{aligned}
x_t &= x_t - \overline{x}, && \text{TREATMENT='RemoveMean'},\\
x_t &= x_t - b\,t, && \text{TREATMENT='RemoveTrend'}\\
x_t &= x_t - b\,t - a, && \text{TREATMENT='RemoveTrendAndMean'}
\end{aligned}
$$

- Remove the seasonal cycle:

$$
x_t = x_t - \overline{x}_{t,ltm}, \quad \text{TREATMENT='Deseason'}
$$

- Whether or not the basis function should be treated, depends on the question to answer with the linear regression model (Bodeker et al., 1998, JGR).

Fill matrix $\boldsymbol{X}$ with treated basis functions

- Create the design matrix $\boldsymbol{X}$ from the treated basis functions (regressors).

Fill matrix $\boldsymbol{X}$ with treated basis functions

- Create the design matrix $\boldsymbol{X}$ from the treated basis functions (regressors).
- Each column represents a time series $t = 1, 2, \ldots, n$..
- The number of columns depends on the number of basis functions and the number of fourier pair expansions used for them.

$$\boldsymbol{X} = \begin{bmatrix} off1 & off2_1 & off3_1 & qbo1_1 & \ldots & tr3_1 \\ off1 & off2_2 & off3_2 & qbo1_2 & \ldots & tr3_2 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ off1 & off2_n & off3_n & qbo1_n & \ldots & tr3_n \end{bmatrix}$$

Main tasks

- Fill the design Matrix $\boldsymbol{X}$ with treated basis functions.

Main tasks

- Fill the design Matrix $\boldsymbol{X}$ with treated basis functions.
- Perform the first run of the least square regression subroutine.

Main tasks

- Fill the design Matrix $\boldsymbol{X}$ with treated basis functions.
- Perform the first run of the least square regression subroutine.
- Analyse the residuals and transform the times series:
  - Run the second order autoregressive model on the residuals.
  - Transform the model according to the autocorrelation coefficients ρ_1 and ρ_2.
  - Update the standard deviation σ_t.

Main tasks

- Fill the design Matrix $\boldsymbol{X}$ with treated basis functions.
- Perform the first run of the least square regression subroutine.
- Analyse the residuals and transform the times series:
  - Run the second order autoregressive model on the residuals.
  - Transform the model according to the autocorrelation coefficients ρ_1 and ρ_2.
  - Update the standard deviation σ_t.
- Perform the second run of the least square regression subroutine.

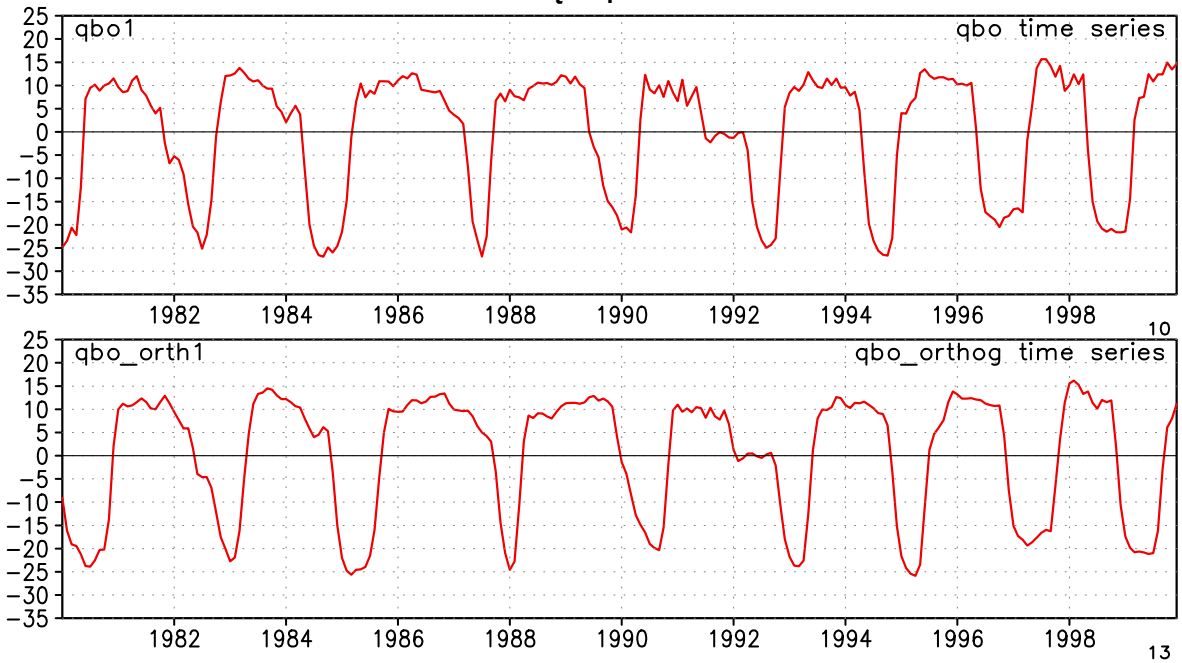
Orthogonal version of the QBO

- To account for the different phases of the QBO in lower and higher levels of the tropical stratosphere two QBO basis functions are used.

Orthogonal version of the QBO

- To account for the different phases of the QBO in lower and higher levels of the tropical stratosphere two QBO basis functions are used.
- An orthogonal version is created from the time series of the QBO. QBO_orthog is orthogonal to QBO if the dot product vanishes.

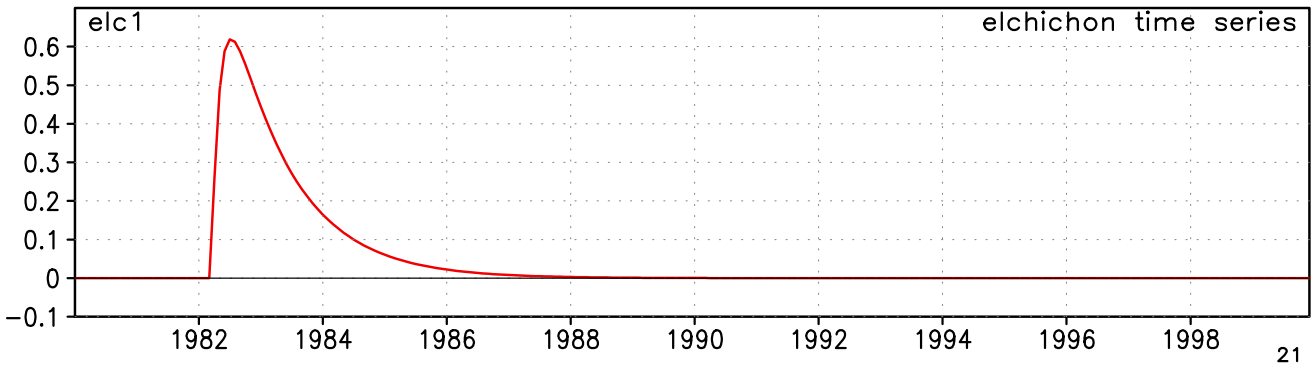
$$\boldsymbol{qbo} \cdot \boldsymbol{qbo\_orthog} = \sum_{t=1}^{n} qbo_t \;\; qbo\_orthog_t = 0$$

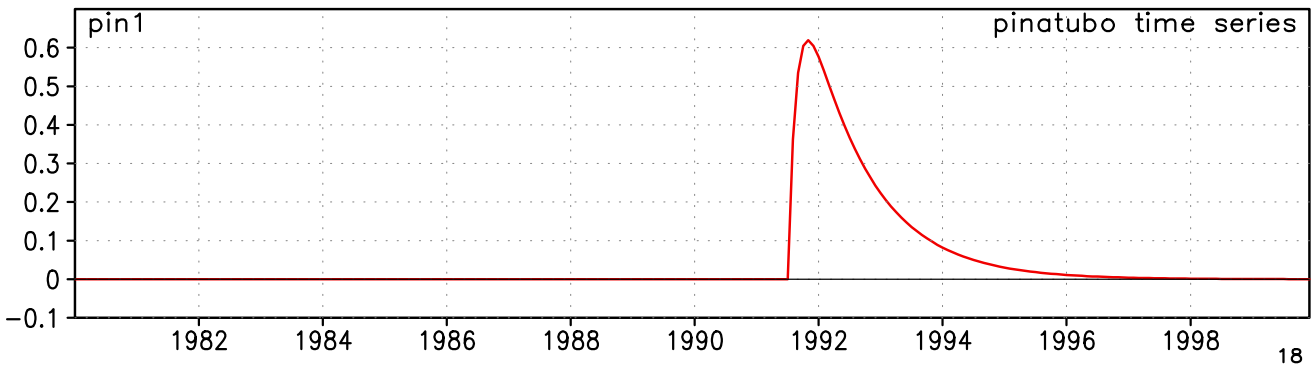


Volcanic basis functions

- rapid initial perturbation followed by an exponential relaxation (Bodeker et al., 1998):

$$y_t = \exp^{(Onset - DecimalDate)} \left(1 - \exp^{6(Onset - DecimalDate)}\right)$$





Outline

Main namelist to control model behaviour.

```
&MLREG_INP
  OFILE        = 'mlreg-output', ! -> output name.
  START_DATE   = '19800101',     ! -> start date of data processing
  END_DATE     = '20051231',     ! -> end date of data processing
  TLAB         = 'ALL',          ! -> use complete time series
  LAUTO        = T,              ! -> account for autocorrelation
                                 !    of the residuals
  ICOEFF       = 1,              ! -> write out 1. regression coeff.
  LOFFSET      = T,              ! -> use offset
  LTREND       = T,              ! -> use trend
  LQBO         = T,              ! -> use QBO
  LQBO_ORTHOG  = T,              ! -> use orthogonal QBO
  LENSO        = T,              ! -> use ENSO
  LPINATUBO    = T,              ! -> use Pinatubo volcano
  LELCHICHON   = T,              ! -> use El Chichon volcano
  LAGUNG       = F,              ! -> use Agung volcano
  LNAO         = F,              ! -> use NAO
  LLAG         = F,              ! -> use time lagged input
  LNAM         = F,              ! -> use Northern annular mode
  LSAM         = F,              ! -> use Southern annular mode
  LEXTRA1      = F,              ! -> \
  LEXTRA2      = F,              ! ->   use extra defined
  LEXTRA3      = F               ! -> /
/
```

namelist to describe the input data time series.

```
&DATA_INP
  LAREA      = T,      ! -> create and area weighted mean over
                       !    latitudes, specified by LAT_S, LAT_E
  LZM        = F,      ! -> create zonal mean prior to regression
  LINP_OUT   = T,      ! -> write out the used input data
  LFIT_OUT   = T,      ! -> write out fitted data for each single
                       !    basis function
  MISSING    = ,       ! -> value for missing data
  TFILTER    = F, 'RMEAN', 5, 0.,0.,0., ! -> characterise time filter
  LON_S      = 25.,    ! -> start longitude; default: all longitudes
  LON_E      = -25.,   ! -> end   longitude
  LAT_S      = 25.,    ! -> start latitude;  default: all latitudes
  LAT_E      = -25.,   ! -> end   latitude
  SDATE      = '',     ! -> start date of input data file
  EDATE      = '',     ! -> end date of input data file
  LEVEL      = 10.,    ! -> requested pressure level(s);
                       !    default: all pressure levels
  CODE       = '',     ! -> code of the variable in the netCDF file
  IFILE      = '',     ! -> netCDF input data file
  LDESEASON = F        ! -> deseasonalise the input time series
                       !    prior to regression
/
```

Filter characterisation:

The namelist entry TFILTER characterises the time filtering of the basis function.
TFILTER is a derived TYPE and has the following elements:

```
TYPE filter
    ! lfilter = .TRUE.    - Apply a time filter.
    !
    LOGICAL             :: lfilter
    !
    ! fkind = 'rmean'        - simple running mean with equal weights.
    ! fkind = 'rmean_gauss'  - running mean with gaussian weights.
    ! fkind = 'butterworth'  - second order butterworth filter.
    !
    CHARACTER(len=16) :: fkind
    !
    INTEGER  :: nrmean    ! number of time steps for running mean
    REAL(sp) :: variance  ! variance used for gaussian weights
    !
    ! The cutoff frequencies of the butterworth filter are
    ! calculated from the time periods lowt and hight:
    ! lowf = 2*Pi/lowt   - lower cutoff frequency
    ! uppf = 2*Pi/uppt   - upper cutoff frequency.
    !
    REAL(sp) :: lowt
    REAL(sp) :: uppt
  END TYPE filter
```

namelist to describe the offset and trend basis function.

```
&OFFSET_INP
  NFOUR     = 0,      ! -> number of fourier pair expansions
  NSPH      = 0,      ! -> number of spherical harmonic expansions
  TREATMENT = 'None'  ! -> always none for offset basis function
/
&TREND_INP
  NFOUR     = 0,
  NSPH      = 0,
  TREATMENT = 'RemoveMean'
/
```

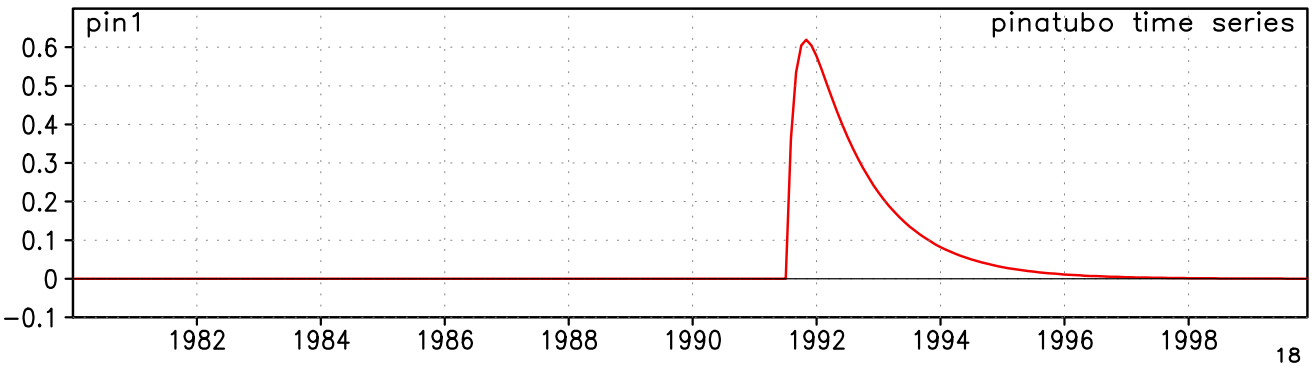
Possible treatments for other basis functions:

```
TREATMENT = 'RemoveMean'
TREATMENT = 'RemoveTrend'
TREATMENT = 'RemoveTrendAndMean'
TREATMENT = 'Deseason'
TREATMENT = 'DeseasonAndRemoveTrend'
TREATMENT = 'DeseasonAndRemoveMean'
TREATMENT = 'DeseasonAndRemoveTrendAndMean'
```

namelist to describe the volcanic basis function.

```
&PINATUBO_INP
  NFOUR     = 0,      ! -> number of fourier pair expansions
  NSPH      = 0,      ! -> number of spherical harmonic expansions
  TSHIFT    = 0.,     ! -> time shift; default: t = -9.99E30
  TREATMENT = 'None'
/
```

Without specifying the parameter TSHIFT, the best optimal shift will be estimated within the program. This can slow down the whole process.



namelist to describe the QBO basis function.

The standard QBO is given by monthly mean radiosonde derived winds (Naujokat 1986). The input is given as an unformatted binary file. All information about the structure of the file must be given in the namelist.

```
&QBO_INP
   NFOUR     = 2,
   NSPH      = 0,
   TREATMENT = 'RemoveTrendAndMean',
   TFILTER   = F, 'RMEAN', 5, 0.,0.,0.,
   SDATE     = '19530101', ! -> start date of the data file
   EDATE     = '20081231', ! -> end date of the data file
   MISSING   = -99999.,    ! -> value for missing data
   LEVEL     = 100.,   70.,   50.,   40.,
               30.,   20.,   15.,   10.,
   WORK      = 50.,      ! -> pressure level WORK for QBO definition
   IFILE     = 'qbo-195301-200812.dat' ! -> unformatted binary file
/
```

namelist to describe the QBO basis function: netCDF input

It is possible to use a netCDF file as QBO input file. This is important when model data should be analysed.

```
&QBO_INP
   NFOUR     = 2,
   NSPH      = 0,
   TREATMENT = 'RemoveTrendandMean',
   TFILTER   = F, 'RMEAN', 5, 0.,0.,0.,
   MISSING   = -9.999E30, ! -> value for missing data
   LAT_S     = 5.,      ! -> calculate area weighted mean from
   LAT_E     = -5.,     !    LAT_S to LAT_E, to define the QBO
   WORK      = 50.,     ! -> use pressure level WORK for
                        !    QBO definition
   CODE      = 'ua',    ! -> name of the variable within the
                        !    netCDF file
   IFILE     = ""       ! -> netCDF input file
/
```

namelist to describe the SFLUX, ENSO, NAO basis functions.

The namelist structure is the same for the SFLUX, ENSO, and NAO basis functions. It is possible to input binary or ASCII data.
Example for ENSO:

```
&ENSO_INP
   NFOUR     = 2,
   NSPH      = 0,
   TREATMENT = 'RemoveTrendandMean',
   TFILTER   = F, 'RMEAN', 5, 0.,0.,0., ! -> characterise time filter
   SDATE     = '18710101',
   EDATE     = '20081201',
   MISSING   = -99.9,
   DTYPE     = 'BIN',                ! -> unformatted binary input
   IFILE     = 'data/ENSO/nino34/nino34_index.dat'
/

&ENSO_INP
   NFOUR     = 2,
   NSPH      = 0,
   TREATMENT = 'RemoveTrendandMean',
   TFILTER   = F, 'RMEAN', 5, 0.,0.,0., ! -> characterise time filter
   MISSING   = -99.9,
   DTYPE     = 'TAB',                  ! -> ASCII input
   IFILE     = 'data/ENSO/nino34/nino34_index_tab.dat'
/
```

namelist to describe the EXTRA basis functions.

There is an additional entry LAB to label the basis function. Again, it is possible to input binary or ASCII data.
Example for QBO according to Randel and Wu:

```
&EXTRA_INP
   LAB        = 'qbo',       ! -> label the basis function
   NFOUR      = 0,
   NSPH       = 0,
   TREATMENT  = 'DeseasonAndRemoveMean',
   TFILTER    = F, 'RMEAN', 5, 0.,0.,0.,
   MISSING    = -99999.,
   DTYPE      = 'TAB',
   IFILE      = '/home/kunze/data/mlreg/data/QBO/qbo1_tab.dat'
/

&EXTRA_INP
   LAB        = 'qbo_orth', ! -> label the basis function
   NFOUR      = 0,
   NSPH       = 0,
   TREATMENT  = 'DeseasonAndRemoveMean',
   TFILTER    = F, 'RMEAN', 5, 0.,0.,0.,
   MISSING    = -99999.,
   DTYPE      = 'TAB',
   IFILE      = '/home/kunze/data/mlreg/data/QBO/qbo2_tab.dat'
/
```

Format of the ASCII input file.

The format of the standard ASCII input file consists of two columns:

- column 1: decimal date.
- column 2: floating point data value.

For example:

`data/SFLUX/solar_flux_monthly_tab.dat`

```
 1947.0416 -99.99
 1947.125 202.7
 1947.2084 235.7
 1947.2916 264.1
 1947.375 261.2
 1947.4584 226.6
 1947.5416 215.2
 1947.625 231.2
 1947.7084 199.7
 1947.7916 209.0
 1947.875 179.8
 1947.9584 176.40001
 1948.0416 155.7
 1948.125 134.3
 1948.2084 135.5
 1948.2916 208.1
 1948.375 226.5
 ....
```

Command line arguments.

All namelists are stored in one single namelist file. The name of the namelist file is given on the command line:

```
Syntax:  mlreg -nl  <namelist>

-nl   <namelist>   execute with the namelist file <namelist>
-wnl  <namelist>   create a default namelist file <namelist>
```

Program output.

All results are stored in netCDF data files. A GrADS control file is provided for each netCDF file.

```
ta_EMAC_1_196001_200012_coeff.ctl    ! -> regression coefficients
ta_EMAC_1_196001_200012_coeff.nc
ta_EMAC_1_196001_200012_coeff_ts.ctl ! -> time resolved coeff.
ta_EMAC_1_196001_200012_coeff_ts.nc
ta_EMAC_1_196001_200012_inp.ctl      ! -> input data and data fit
ta_EMAC_1_196001_200012_inp.nc
ta_EMAC_1_196001_200012_prob.ctl     ! -> probabilities of t test
ta_EMAC_1_196001_200012_prob.nc
ta_EMAC_1_196001_200012_prob_ts.ctl
ta_EMAC_1_196001_200012_prob_ts.nc
ta_EMAC_1_196001_200012_regr.ctl     ! -> basis functions
ta_EMAC_1_196001_200012_regr.nc
ta_EMAC_1_196001_200012_tval.ctl     ! -> t test statistics
ta_EMAC_1_196001_200012_tval.nc
ta_EMAC_1_196001_200012_tval_ts.ctl
ta_EMAC_1_196001_200012_tval_ts.nc
ta_EMAC_1_196001_200012_unc.ctl      ! -> uncertainties
ta_EMAC_1_196001_200012_unc.nc
ta_EMAC_1_196001_200012_unc_ts.ctl
ta_EMAC_1_196001_200012_unc_ts.nc
```

Expand the results in time

- The regression coefficients can be expanded in time to get the seasonal variations of the specific influence:

$$y_{tj} = \beta_{j0} + \sum_{k=1}^{m} [\, \beta_{j(2k-1)} \, sin(2\pi kt/365.25) + \beta_{j(2k)} \, cos(2\pi kt/365.25) \,]$$

Program output.

```
File: ta_EMAC_1_196001_200012_coeff.ctl

dset ^ta_EMAC_1_196001_200012_coeff.nc
dtype netcdf
.....
vars    49
off1=>off1  31 t,z,y,x [offset]
off2=>off2  31 t,z,y,x [offset]*sin( 2*Pi*Decimal Year )
off3=>off3  31 t,z,y,x [offset]*cos( 2*Pi*Decimal Year )
off4=>off4  31 t,z,y,x [offset]*sin( 4*Pi*Decimal Year )
off5=>off5  31 t,z,y,x [offset]*cos( 4*Pi*Decimal Year )
off6=>off6  31 t,z,y,x [offset]*sin( 6*Pi*Decimal Year )
off7=>off7  31 t,z,y,x [offset]*cos( 6*Pi*Decimal Year )
tr1=>tr1    31 t,z,y,x [trend]
tr2=>tr2    31 t,z,y,x [trend]*sin( 2*Pi*Decimal Year )
tr3=>tr3    31 t,z,y,x [trend]*cos( 2*Pi*Decimal Year )
tr4=>tr4    31 t,z,y,x [trend]*sin( 4*Pi*Decimal Year )
tr5=>tr5    31 t,z,y,x [trend]*cos( 4*Pi*Decimal Year )
tr6=>tr6    31 t,z,y,x [trend]*sin( 6*Pi*Decimal Year )
tr7=>tr7    31 t,z,y,x [trend]*cos( 6*Pi*Decimal Year )
......
```

Program output.

```
File: ta_EMAC_1_196001_200012_inp.nc

data=>data    31 t,z,y,x input data time series
dfit1=>dfit1  31 t,z,y,x complete fit of the regression model
dfit2=>dfit2  31 t,z,y,x second complete fit of the regression model
dfit3=>dfit3  31 t,z,y,x third complete fit of the transformed data
off1=>off1    31 t,z,y,x [offset]
off2=>off2    31 t,z,y,x [offset]*sin( 2*Pi*Decimal Year )
off3=>off3    31 t,z,y,x [offset]*cos( 2*Pi*Decimal Year )
......
res1=>res1    31 t,z,y,x residuals
res2=>res2    31 t,z,y,x residuals after accounting for autocorrelatio
```

References

- D. C. Montgomery, E. A. Peck, G. G. Vining
  Introduction to linear regression analysis.
  John Wiley & sons, 2001.

- G. E. Bodeker, I. S. Boyd, and W. A. Matthews
  Trends and variability in vertical ozone and temperature profiles measured by ozonesondes at Lauder, New Zealand: 1986–1996.
  J. of Geophys. Res., 103, D22, 28661-28681, 1998.

- G. E. P. Box, and G. M. Jenkins
  Time Series Analysis Forecasting and Control.
  Holden-Day, Merrifield, Va., 1970.

- G. C. Tiao et al.
  Effects of autocorrelation and temporal sampling schemes on estimates of trend and spatial correlation.
  J. Geophys. Res., 95, 20507–20517, 1990.