Course 02403 Introduction to Mathematical Statistics

Lecture 9: Linear regression: model reduction, residuals, multicollinearity and polynomial regression

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Agenda

1. Succesive testing
2. Linear regression as an LM
   - The t-test and ANOVA
3. Checking assumptions
   - Residuals
   - Influencial observations
4. Multicollinearity
5. Polynomial regression

Overview

Nested models/hypotheis

Two models are said tobe nested, if the simpler model can be formulated by fixing a number of parameters in the more complicated model.

Example: Consider the model

$$M_2: \quad Y_i = \beta_0 + x_{1,i}\beta_1 + x_{2,i}\beta_2 + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma^2)$$

A nested model is

$$M_1: \quad Y_i = \beta_0 + x_{1,i}\beta_1 + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma^2)$$

And further the model

$$M_0: \quad Y_i = \beta_0 + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma^2)$$

is neted to both M_2 and M_1. And we write

$$M_0 \subset M_1 \subset M_2$$

Type I partioning of variation

Consider a series of nested hypothesis

$$H_0 \subset H_1 \subset \cdots \subset H_M \subset \mathbb{R}^n$$

corresponding to the design matrices

$$\begin{aligned}
\boldsymbol{X}_0 &= \mathbf{1} \\
\boldsymbol{X}_1 &= [\mathbf{1} \quad \tilde{\boldsymbol{X}}_1] \\
&\vdots \\
\boldsymbol{X}_i &= [\boldsymbol{X}_{i-1} \quad \tilde{\boldsymbol{X}}_i] \\
&\vdots \\
\boldsymbol{X}_M &= [\boldsymbol{X}_{M-1} \quad \tilde{\boldsymbol{X}}_M],
\end{aligned}$$

and corresponding projection matrices $\boldsymbol{H}_i = \boldsymbol{X}_i(\boldsymbol{X}_i^T \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i^T$.

Type I partioning of variation

Paramters of the model is: $\boldsymbol{\beta} = [\beta_0, \tilde{\boldsymbol{\beta}}_1^T, ..., \tilde{\boldsymbol{\beta}}_M^T]^T$

The total variation can be writen as

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y} + \sum_{i=1}^{M} \boldsymbol{Y}^T(\boldsymbol{H}_i - \boldsymbol{H}_{i-1})\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}.$$

If $\tilde{\boldsymbol{\beta}}_j = \boldsymbol{0}$ for $j > i$ then

$$F_j = \frac{\boldsymbol{Y}^T(\boldsymbol{H}_j - \boldsymbol{H}_{j-1})\boldsymbol{Y}/df_j}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}/df_{SSE}} \sim F(1, df_j); \quad j > i$$

and hypothesis tests can be based on that distribution.

- $df_j = Trace(\boldsymbol{H}_j - \boldsymbol{H}_{j-1}) = Rank(\tilde{\boldsymbol{X}}_j)$
- $df_{SSE} = Trace(\boldsymbol{I} - \boldsymbol{H}_M) = n - Rank(\boldsymbol{X}_M)$
- $\hat{\sigma}^2 = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}/df_{SSE}$, is a central estimator.

Type III partioning of variation

The type I test is conditional, meaning that the test at level j is conditioned on level $j, j+1, .., M$ already been removed from the model.

Type I: The order of effects

In the type I partioning of variation the order in which effects are entered matters.

In the Type III test are formulated as if every effect was entered last

$$\begin{aligned} \boldsymbol{X}_M &= [\mathbf{1} \quad \tilde{\boldsymbol{X}}_1 \quad \cdots \quad \tilde{\boldsymbol{X}}_M] \\ \boldsymbol{X}_{-i} &= [\mathbf{1} \quad \tilde{\boldsymbol{X}}_1 \quad \cdots \quad \tilde{\boldsymbol{X}}_{i-1} \quad \tilde{\boldsymbol{X}}_{i+1} \quad \cdots \quad \tilde{\boldsymbol{X}}_M] \end{aligned}$$

with the partioning af variation

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_{-i}\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{H}_M - \boldsymbol{H}_{-i})\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}$$

and the test statistics $(H_0 : \tilde{\boldsymbol{\beta}}_i = \mathbf{0})$

$$F_i = \frac{\boldsymbol{Y}^T(\boldsymbol{H}_M - \boldsymbol{H}_{-i})\boldsymbol{Y}/(p - p_i)}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_M)\boldsymbol{Y}/(n - p)} \sim F(p - p_i, n - p).$$

Items on a scale

Two items are put on a scaled, first separately then together and the deviation from a nominal weight is reported. Let μ_i denote the weight of item i, there is a fairly clear hierarchy of hypothesis

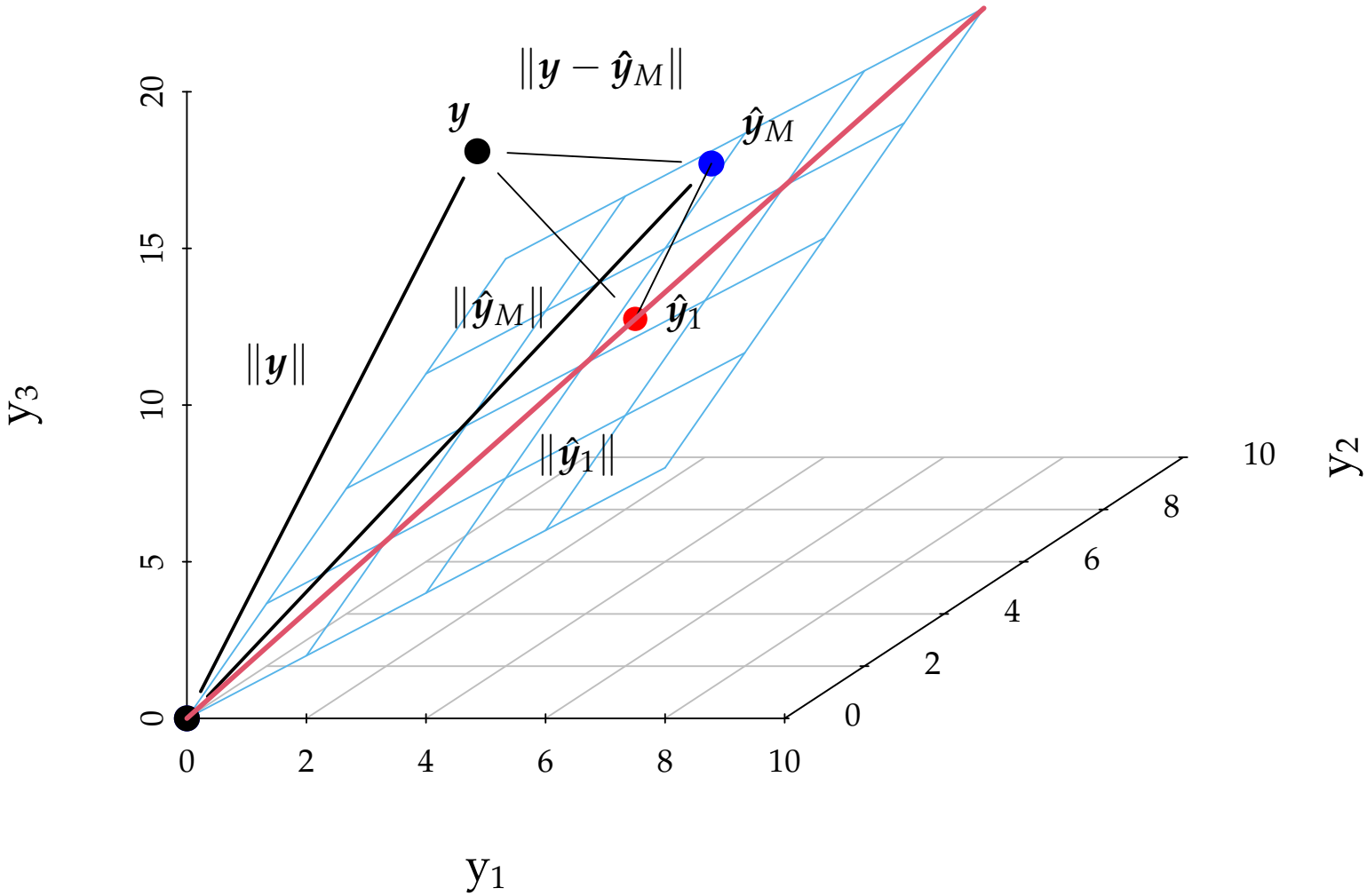
$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = 0 \\ H_1: &\quad \mu_1 = \mu_2 = \mu \\ H_2: &\quad \mu_1 \neq \mu_2. \end{aligned}$$

In terms of design matrices this can be formulated as

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}; \quad \boldsymbol{X}_2 = \begin{bmatrix} 1 & -\frac{1}{2} \\ 1 & \frac{1}{2} \\ 2 & 0 \end{bmatrix}$$

Parameter interpretation β_0: average weight of items, β_1: deviation from average.

Items on a scale: geometri



Example: Skive fjord

Using Type I and Type III partioning of variation to test models for log chlorophyll.

Overview

The general linear model

The general linear model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

with $\boldsymbol{X} \in \mathbb{R}^{n \times p}$

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \\
\hat{\sigma}^2 &= \frac{1}{n-p} \boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{Y} = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\
\boldsymbol{H} &= \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T.
\end{aligned}$$

The partial t-test and ANOVA

The partial t-test

$$t_{obs,i} = \frac{\hat{\beta}_i}{\sqrt{(\hat{\mathbf{\Sigma}}_\beta)_{ii}}} \sim t(n-p),$$

and the Type III ANOVA are equivalent in the sence that

$$t_{obs,i}^2 = F_i$$

and $t_{obs,t}^2 \sim F(1, n-1)$ and $F_i \sim F(1, n-1)$.

Test for total homogeneity

Given a linear regression model

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Statistical software ususally report the test for total homogeneity

$$H_0: \quad \beta_1 = \ldots = \beta_p = 0$$

corresponding to the data being described only by a mean value, or that at least one of the parameters is significantly different from zero.

Linear transformations of input/regressors

The general linear model is invariant to linear trainsformation of the desing matrix, i.e. if $\boldsymbol{T}$ is invertible then

$$\boldsymbol{H} = \tilde{\boldsymbol{H}}$$

where $\boldsymbol{H}$ is based on $\boldsymbol{H}$ and $\tilde{\boldsymbol{H}}$ is based on $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{T}$.

Eg. it does not matter if temperature is measure in Celsius or Fahrenheit, or time is measured in minutes or days, as long as the intercept is included in the model.

Example: Skive fjord

Perform the test for total homogeneity.

Overview

Residuals

In the general linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

The assumption are $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ or $\varepsilon_i \sim N(0, \sigma^2)$ and iid.

We do not observe $\boldsymbol{\epsilon}$ but rather

$$\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

and therefore

$$\boldsymbol{r} \sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))$$

and hence r_i is not iid. e.g. $V[r_i] = \sigma^2(1 - h_{ii})$.

Residuals

Definition (Standardized residuals)

Standardized residuals are defined as

$$r_i^{rs} = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

standardized residuals are identicaly distrubuted, with a complicated distribution.

Definition (Studentized residuals)

Studentized residuals are defined as

$$r_i^{rt} = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \sim t(n-p-1),$$

where $\hat{\sigma}^2_{(i)}$ is the estimate of the variance, excluding the i'th observation.

Studentized are hence more suited for QQ-plots.

Influencial observations

Observations can be unusual in (at least) two ways

- Observations might be far from the prediction.
- The input might be far from other input.

The first point is measured by the residual, the second point can be measured by the effect a change in the observation have in the fitted value

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}} = \boldsymbol{H},$$

and the values of h_{ii} (also called leverage) should be monitored.

Example: Skive fjord

Perform residual analysis of the model.

Overview

Example: perfect collinearity

Consider the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma^2)$$

asume that $x_2 = a + bx_1$, then

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_1 + \beta_2 (a + bx_1) + \varepsilon_i \\ =& \beta_0 + \beta_2 a + (\beta_1 + \beta_2 b) x_1 + \varepsilon_i. \end{aligned}$$

I.e. 2 (not 3) mean value parameters ($(\beta_0 + \beta_2 a)$, and $(\beta_1 + \beta_2 b)$).

In this case the empirical correlation between x_1 and x_2 will be 1.

Multicollinearity in practice

In practice we usually do not observe perfect collinearity, but rather correlations close to ± 1, and we should check

- Pairwise correlations between explanatory variables, e.g. pairwise scatter plots.
- The problems can be trace back to taking the inverse of $\boldsymbol{X}^T\boldsymbol{X}$, hence the condition number of that matrix might also be examined.

Simple actions to avoid (strong) multicollinarity

- Design your experiment in a "good" way
- Instead of the regressor x_i use $x_i - \bar{x}$

In addition principal componet analysis can be used to remove multicollinearity all together (i.e. orthogonal design), at price of interpretability.

Overview

Basis function regression

Linearity of the general linear model refer to the parameters, not the regressor. Hence the model

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_i)\beta_j; \quad \varepsilon_i \sim N(0, \sigma^2)$$

where $f_j(\cdot)$ are known function, also belong to the family of general linear models.

The most simple choice of basis functions are polynomial

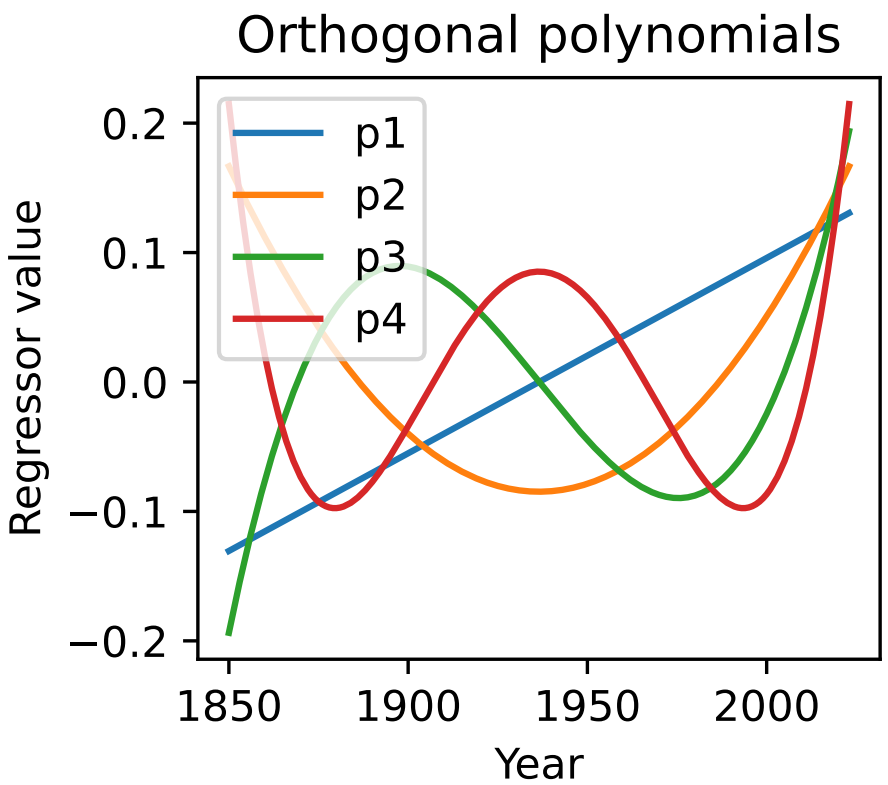
$$Y_i = \beta_0 + \sum_{j=1}^{p} x_i^j \beta_j + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma^2)$$

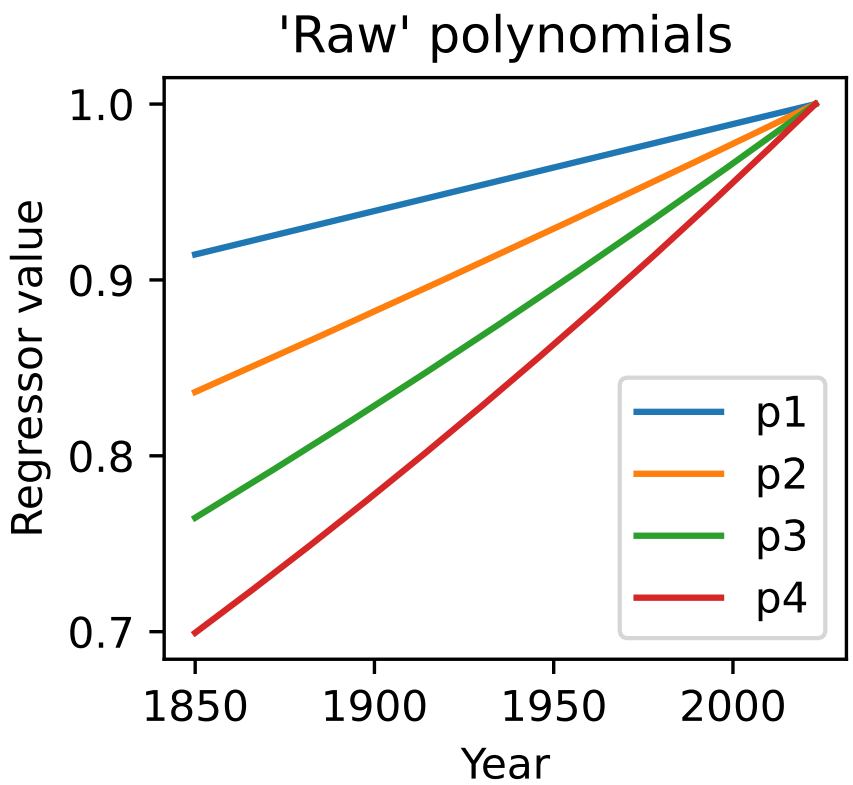
the term x_i^j is often replaced by (close to) orthogonal polynomials of order j. But other basis functions (like *sine*, *cosine* pairs) are also common.

Basis function regression, some notes

- Be carefull with extrapolation.
- Use appropriate basis function (e.g. *sine*, *cosine* pairs for periodic variations)
- Be ware of collinarity, replacing x_i^j with $(x_i - \bar{x})^j$ usually takes you a long way, but orthogonal polynomials can be constructed.

Orthogonal polynomials





Example: Skive fjord

Make a model of chlorophyll as a 3rd order polynomial function of temperature and global radiation. Test for model reduction.

Summary

1. Succesive testing
2. Linear regression as an LM
   - The t-test and ANOVA
3. Checking assumptions
   - Residuals
   - Influencial observations
4. Multicollinearity
5. Polynomial regression