27411 Chemometrics and Biological Data Analysis

Lecture 2: MLR - Multiple Linear Regression (eNote3)

Per Bruun Brockhoff

DTU Compute, Statistics and Data Analysis
Building 324, Room 220
Danish Technical University
2800 Lyngby – Danmark
e-mail: perbb@dtu.dk

DTU Compute
Department of Applied Mathematics and Computer Science

Overview

DTU Compute
Department of Applied Mathematics and Computer Science

Regression Analysis- overview

- Description (Estimation)
- Inference (Hypothesis testing, confidence bands)
- Prediction (of future values)
- Validation/model diagnostics/data check.

DTU Compute
Department of Applied Mathematics and Computer Science

Simple Linear Regression

- Model: $y_i = b_0 + b_1 x_i + \varepsilon_i$
- "Guesses" Y given knowledge of X: $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$
  - Error in guessing: $\sum_i (y_i - \hat{y}_i)^2$
- Guessing Y WITHOUT any X-knowledge: $y_i^{guess} = \bar{y}$
  - Error in guessing: $\sum_i (y_i - \bar{y})^2$ ("usual" Y-variance)

DTU Compute
Department of Applied Mathematics and Computer Science

Explained variance

$$R^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- R^2 = squared correlation between y_is and $\hat{y}_i$s
- R^2 = squared correlation between y_is and x_is
  - $(\hat{b}_0, \hat{b}_1)$ is chosen to maximize R^2
  - $(\hat{b}_0, \hat{b}_1)$ is chosen to minimize squared residual variation.
  - Least Squares (LS) estimates.

Multiple Linear Regression (MLR)

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + \varepsilon_i, \; i = 1, \ldots, n$$

E.g. $n = 6$ objects and $p = 3$ x-variables:

$$\begin{aligned}
y_1 &= b_0 + b_1 x_{11} + b_2 x_{12} + b_3 x_{13} + \varepsilon_1 \\
y_2 &= b_0 + b_1 x_{21} + b_2 x_{22} + b_3 x_{23} + \varepsilon_2 \\
y_3 &= b_0 + b_1 x_{31} + b_2 x_{32} + b_3 x_{33} + \varepsilon_3 \\
y_4 &= b_0 + b_1 x_{41} + b_2 x_{42} + b_3 x_{43} + \varepsilon_4 \\
y_5 &= b_0 + b_1 x_{51} + b_2 x_{52} + b_3 x_{53} + \varepsilon_5 \\
y_6 &= b_0 + b_1 x_{61} + b_2 x_{62} + b_3 x_{63} + \varepsilon_6
\end{aligned}$$

Multiple Linear Regression (MLR)

Or equivalently:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \\ 1 & x_{51} & x_{52} & x_{53} \\ 1 & x_{61} & x_{62} & x_{63} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}$$

Or equivalently:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon}$$

Assumptions of MLR

- Model structure is OK:
  $\mathrm{E}(y_i | x_{i1}, x_{i2}, \ldots, x_{ip}) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$
  - Errors $\varepsilon_1, \ldots, \varepsilon_n$
  - are independent
- have homogeneous variance: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- The $n \times p$-matrix $\boldsymbol{X}$ is of full rank:
  - No x-variable can be perfectly expressed by a linear combination of the others. (No *perfect redundancy*)
  - A consequence: The number of objects must exceed the number of variables: $n > p$

Regression Analysis

- Description: $\hat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ (Unbiased: $\mathrm{E}(\hat{b}) = b$)
- Inference: $\mathrm{Var}(\hat{\boldsymbol{b}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (Minimal)
- Prediction: $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{b}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i1} + \hat{b}_2 x_{i2} + \cdots + \hat{b}_p x_{ip}$$

- Validation: "Look at" $e_i = y_i - \hat{y}_i$

Matrix of cross products ($n = 6$, $p = 3$)

$$\boldsymbol{X}'\boldsymbol{X} = \left[\begin{array}{cccccc} 1 & 1 & 1 & 1 & 1 & 1 \\ \hline x_{11} & x_{21} & x_{31} & x_{41} & x_{51} & x_{61} \\ x_{12} & x_{22} & x_{32} & x_{42} & x_{52} & x_{62} \\ x_{13} & x_{23} & x_{33} & x_{43} & x_{53} & x_{63} \end{array}\right] \left[\begin{array}{c|ccc} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \\ 1 & x_{51} & x_{52} & x_{53} \\ 1 & x_{61} & x_{62} & x_{63} \end{array}\right]$$

$$= \left[\begin{array}{c|ccc} 6 & \sum_i x_{i1} & \sum_i x_{i2} & \sum_i x_{i3} \\ \hline \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \sum_i x_{i1}x_{i3} \\ \sum_i x_{21} & \sum_i x_{i2}x_{i1} & \sum_i x_{i2}^2 & \sum_i x_{i2}x_{i3} \\ \sum_i x_{31} & \sum_i x_{i3}x_{i1} & \sum_i x_{i3}x_{i2} & \sum_i x_{i3}^2 \end{array}\right]$$

Matrix of cross products

IF the variables are mean-centered AND we omit the first column (of ones):

$$\begin{aligned} \left(\boldsymbol{X}'\boldsymbol{X}\right)_{jk} &= \sum_i (x_{ij} - \bar{x}_{\cdot j})(x_{ik} - \bar{x}_{\cdot k}) \\ &= (n-1)\mathrm{Corr}(\boldsymbol{x}_j, \boldsymbol{x}_k)\sqrt{\mathrm{Var}(\boldsymbol{x}_j)}\sqrt{\mathrm{Var}(\boldsymbol{x}_k)} \end{aligned}$$

Matrix of cross products

IF the variables are mean-centered AND standardized: (still omitting the first column):

$$\begin{aligned} \left(\boldsymbol{X}'\boldsymbol{X}\right)_{jk} &= \sum_i \frac{(x_{ij} - \bar{x}_{\cdot j})(x_{ik} - \bar{x}_{\cdot k})}{\sqrt{\mathrm{Var}(\boldsymbol{x}_j)}\sqrt{\mathrm{Var}(\boldsymbol{x}_k)}} \\ &= (n-1)\mathrm{Corr}(\boldsymbol{x}_j, \boldsymbol{x}_k) \end{aligned}$$

Inference

- $\hat{\sigma}^2 = \sum_i (y_i - \hat{y}_i)^2/(n-p-1)$
- Testing the model: $H_0 : b_1 = b_2 = \cdots = b_p$

$$F = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \hat{y}_i)^2/(n-p-1)}, \sim F(p, n-p-1)$$

- Testing individual parameters: $H_0 : b_k = 0$

$$t = \frac{\hat{b}_k}{SE(\hat{b}_k)}, \sim t(n-p-1)$$

$$SE((\hat{b}_k))^2 = \; k\text{th diagonal element of } \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Centering and scaling in MLR

- X-Centering changes $\hat{b}_0$ BUT leaves everything else UNCHANGED
- X-scaling changes $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_p$ by the scaling factors
  - Variance and inference is unchanged!
- MLR is *invariant* to centering and scaling

Interpretation of parameters

- Each parameter expresses the ADDITIONAL effect of the variable AFTER the effects of all other variables have been removed
- Effects of other varibles are *corrected for*
- Expresses the expected change in y given a change of the individual variable leaving all other variables fixed.

Interpretation of parameters

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \varepsilon_i$$

$\hat{b}_1$ can be found by:

- Regress $\boldsymbol{y}$ on $\boldsymbol{x}_2$, compute residuals: e_i^y
- Regress $\boldsymbol{x}_1$ on $\boldsymbol{x}_2$, compute residuals: e_i^x
- Regress e^y on e^x.

Validation: Assumptions

- Model structure is satisfactory
- Homogeneous variance: $\text{Var}(\varepsilon_i) = \sigma^2$
- Independent errors
- (Data is OK)

Validation: Tools

- Residual investigations
- Plots and tests
- Investigating models that allows for EXTENDED structure
  - Interactions
  - Non-linearities
- Cross-Validation and Jacknifing
- Checking for outliers and influential observations

Residuals

- Residuals: $e_i = y_i - \hat{y}_i$
- Studentized Residuals I: $e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$
- Studentized Residuals II: $e_i^* = \frac{e_i}{\hat{\sigma}(i)\sqrt{1-h_{ii}}}$
- Leverage of observation i: $h_{ii} = \left(X(X^tX)^{-1}X^t\right)_{ii}$
  - Express the (potential) influence of the observation
  - Measures the relative X-position of the observation
- $\hat{\sigma}(i)$: Estimated in a model fit WITHOUT using the i'th observation.

Influential/outlying observations

- Residuals vs. leverage
- Another method: DFBETA: measure the change of beta when leaving out an observation.
- Outliers: Extreme y-values. Look at all plots!

Additional structure

- Ptentially: plot residuals vs. each of the X-variables
- Include e.g. x_i^2 (and maybe also x_i^3) in the model and check whether these terms are significant!
- Include interaction terms: Allowing e.g. the effect of ELEVATION to depend on COUNTY:
  - Include the product of COUNTY and ELEVATION in the model

Resampling

- Cross-Validation (CV)
- Jackknifing (Leave-on-out CV)
- Bootstrapping
- A good generic approach:
  - Split the data into a TRAINING and a TEST set.
  - Use Cross-validation on the TRAINING data
  - Check the model performance on the TEST-set
  - MAYBE: REPEAT all this many times (Repeated Double Cross Validation)

Multi-collinearity!!!

- IF too much redundant X-information: model fit becomes UNSTABLE!
  - Poor prediction performance!
- Classic approach: Variable selection!
- Better approach: "Biased regression methods"
  - Principal Component Regression (PCR)
  - Partial Least Squares Regression (PLSR)
  - Ridge and Lasso Regression (RR)
  - Many others exist