Regression Models in Statistics

A Short Guide Through a Set of Abbreviations Containing the Letters L, M, G, and A

Gero Walter

Department of Statistics
Ludwig-Maximilians-Universität München (LMU)

September 5th, 2011

Concept and Scope

Linear Regression:

$$y_i = x_i^{\mathsf{T}}\beta + \varepsilon_i \text{ with } \mathrm{E}[\varepsilon_i] = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2,$$

or $$\mathrm{E}[y_i] = x_i^{\mathsf{T}}\beta = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots$$

Concept and Scope

Linear Regression:

$$
\begin{aligned}
& y_i = x_i^\top \beta + \varepsilon_i \text{ with } \mathrm{E}[\varepsilon_i] = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2, \\
\text{or} \quad & \mathrm{E}[y_i] = x_i^\top \beta = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots
\end{aligned}
$$

- modeling: determine & quantify the influence of each predictor variable $x_{i1}, x_{i2}, \ldots$ on the response variable y_i
  - tests on estimated regression parameters $\beta_1, \beta_2, \ldots$
  - model / variable selection (separate procedures / simultaneous with estimation)

Concept and Scope

Linear Regression:

$$y_i = x_i^T \beta + \varepsilon_i \text{ with } \mathrm{E}[\varepsilon_i] = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2,$$

or $$\mathrm{E}[y_i] = x_i^T \beta = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots$$

- modeling: determine & quantify the influence of each predictor variable $x_{i1}, x_{i2}, \ldots$ on the response variable y_i
  - tests on estimated regression parameters $\beta_1, \beta_2, \ldots$
  - model / variable selection (separate procedures / simultaneous with estimation)
- prediction of the response variable y_{n+1} given x_{n+1}
  - categorical response $\hat{=}$ classification
  - "supervised learning" in machine learning
  - provide enough model flexibility, but prevent overfitting

Generalizations of Linear Regression

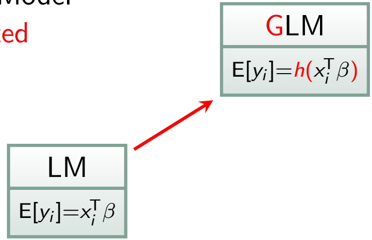
LM: Linear Model

LM
$\mathrm{E}[y_i] = x_i^T \beta$

Generalizations of Linear Regression

LM: Linear Model

G: Generalized



binary, categorical, ordinal, count data (. . .) response modeled by response function

Generalizations of Linear Regression

LM: Linear Model

G: Generalized

M: Mixed

GLM

$E[y_i] = h(x_i^T \beta)$

binary, categorical, ordinal, count data (. . .) response modeled by response function

LM

$E[y_i] = x_i^T \beta$

LMM

$E[y_{ij}] = x_{ij}^T \beta + u_{ij}^T \gamma_i$

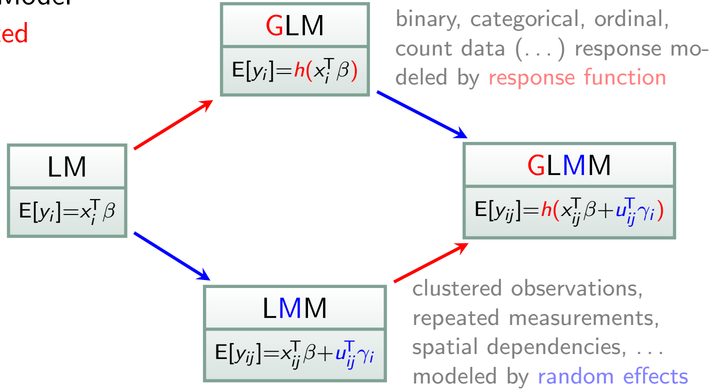
clustered observations, repeated measurements, spatial dependencies, . . . modeled by random effects

Generalizations of Linear Regression

LM: Linear Model

G: Generalized

M: Mixed



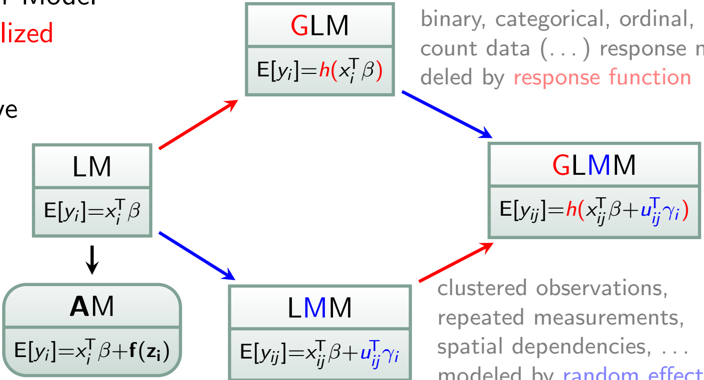
Generalizations of Linear Regression

LM: Linear Model

G: Generalized

M: Mixed

A: Additive



binary, categorical, ordinal, count data (...) response modeled by response function

clustered observations, repeated measurements, spatial dependencies, ... modeled by random effects

univariate smoothing: z_i has nonlinear influence on y_i. functional form estimated via basis functions approach

Generalizations of Linear Regression

LM: Linear Model
G: Generalized
M: Mixed
A: Additive

binary, categorical, ordinal, count data (...) response modeled by response function

GLM: $\mathrm{E}[y_i] = h(x_i^T \beta)$

LM: $\mathrm{E}[y_i] = x_i^T \beta$

GAM: $\mathrm{E}[y_i] = h(x_i^T \beta + \mathbf{f}(\mathbf{z_i}))$

GLMM: $\mathrm{E}[y_{ij}] = h(x_{ij}^T \beta + u_{ij}^T \gamma_i)$

AM: $\mathrm{E}[y_i] = x_i^T \beta + \mathbf{f}(\mathbf{z_i})$

LMM: $\mathrm{E}[y_{ij}] = x_{ij}^T \beta + u_{ij}^T \gamma_i$

clu... re... sp... mo...

GAMM: $\mathrm{E}[y_{ij}] = h(x_{ij}^T \beta + u_{ij}^T \gamma_i + \mathbf{f}(\mathbf{z_i}))$

AMM: $\mathrm{E}[y_{ij}] = x_{ij}^T \beta + u_{ij}^T \gamma_i + \mathbf{f}(\mathbf{z_i})$

univariate smoothing: z_i has nonlinear influence on y_i. functional form estimated via basis functions approach

Examples For Further Intricacies

- linear/additive predictor approach can be used to model other quantities of interest
  - proportional hazard/Cox models: $\lambda_i(t) = \lambda_0(t)\exp(x_i^T\beta)$
  - quantile regression: modeling quantiles of the response distribution
- varying coefficients: $\beta_2 \rightarrow \beta_2(t)$
  (or depending on other variables than t)
- estimating also the response function $h(\cdot)$ in GLMs/GAMs
- correcting for measurement errors in the predictors
- $p \gg n$ (gene expression data: 100 obs. for 500 000 variables)
- functional data (e.g., from mass spectrometry)
- ...

Some Estimation Techniques

- least squares (yawn. . .)
- robust methods (L_1 regression, . . .)
- maximum likelihood
  - **A**Ms: penalized ML
  - shrinkage estimators (ridge, lasso, . . .)
  - quasi-likelihood / generalized estimation equations (GEE)
- boosting, support vector machine, . . . (from machine learning)
- Bayesian (empirical / full: penalization $\hat{=}$ prior)

Some Estimation Techniques

- least squares (yawn. . .)
- robust methods (L_1 regression, . . .)
- maximum likelihood
  - **A**Ms: penalized ML
  - shrinkage estimators (ridge, lasso, . . .)
  - quasi-likelihood / generalized estimation equations (GEE)
- boosting, support vector machine, . . . (from machine learning)
- Bayesian (empirical / full: penalization $\hat{=}$ prior)
- NPI