Statistical learning for biological data

David Causeur
Department of Statistics and Computer Science
Agrocampus Ouest
IRMAR CNRS UMR 6625
http://www.agrocampus-ouest.fr/math/causeur/

Course objectives

Expertise in regression modeling for biological issues

- Nonlinear and nonparametric regression;
- Handling high-throughput profiles of explanatory variables;
- Model choice;
- Functional data analysis.

Course objectives

Mathematical vs Applied statistics

- Statistical theory is reduced to its essentials
- Solving problems by data analysis using `R`

Course objectives

By the end, students are expected to be able to:

- Implement methods for high-dimensional regression;
- Compare procedures based on statistical arguments;
- Assess the prediction performance of a learning algorithm;
- Apply these key insights using statistical software.

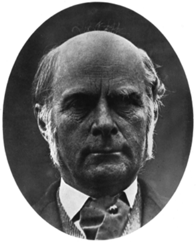
Pre-requisites/assignments

- Regression
  - Assumptions of linear regression modeling?
  - Ordinary Least squares fitting?
- Model assessment
  - R^2?
  - AIC?
- Testing
  - t-test?
  - F-test?
- Statistical software: `R`
  - `glm(y~x,...)`?
  - `anova(glm(...))`?

Assignments: 1-hour written exam (all documents permitted)

Outline

Understanding life mechanisms

F. Galton 1822-1911	R. Fisher 1890-1962	W. Gosset (*Student*) 1876-1937
		

Issue in life sciences: understanding phenotypical variations

In agricultural sciences: understanding yields variations

Regression modeling

Wheat yield (Y) modeling

Wheat production profile (x)

- Variety
- Chemical inputs
- Soil composition
- ...

For a given profile $x = (x_1, \ldots, x_p)'$ with phenotype Y

$$\mathbb{E}_x(Y) = f(x)$$

f: regression function

Range of regression modeling

- *Y can take various forms.* Among them:
  - → The reference framework. Y on a continuous scale.
    $\mathbb{E}_x(Y)$ is a 'mean' Y value for the profile x
  - → The 'classification' framework. $Y \in \{y_1, \ldots, y_K\}$ is a K-class variable.
    $\mathbb{E}_x(Y)$ is a $K-$vector of class probabilities $\mathbb{P}_x(Y = y_k)$ for the profile x
- *f also*

→ f known up to some unknown parameters
$f(x; \beta_0, \beta_1) = \beta_0 + \beta_1 x$, $f(x; \beta_0, \beta_1) = \beta_0 x^{\beta_1}$, ...

→ f fully unknown
$f(x)$ is 'regular' (continuous, differentiable, ...)

The model selection issue

General framework for the course:

- One response variable Y
- Many explanatory variables $x = (x_1, \dots, x_p)$
- **Data**: n independent joint observations $(x_i, Y_i),\ i = 1, \dots n$

Central question: What is the best model to predict Y using $x,\ j = 1, \dots, p$?

Sub-question: How to compare the prediction ability of two models?

Subsub-question: How to fit a model?

Illustration with a real-valued response

LMP: Lean Meat Percentage of a pig carcass

- LMP requires complete dissection
  → impossible on the slaughter-line
  → LMP is predicted by fat and muscle depths

- Different devices to measure tissue depths

Invasive probe

Scanning device

Prediction of the LMP

Linear regression model

$$\begin{aligned}\mathbb{E}_x(Y) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \\ &\varepsilon = Y - \mathbb{E}_x(Y) \sim \mathcal{N}(0, \sigma),\end{aligned}$$

where

- Y is the LMP of a pig
- $x = (x_1, \ldots, x_p)'$ is the 'tissue depths' profile of this pig
- β_0 and $\beta = (\beta_1, \ldots, \beta_p)'$ are the regression parameters
- σ is the residual standard deviation.

To fit the regression model = to estimate the βs

Data needed to fit the model

Sample of independent units

Units	Y	x_1	x_2	$\ldots$	x_p
1	Y_1	x_{11}	x_{12}	$\ldots$	x_{1p}
2	Y_2	x_{21}	x_{22}	$\ldots$	x_{1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	Y_n	x_{n1}	x_{n2}	$\ldots$	x_{np}

- Import `pig` data in the `R` session

The reference fitting method: least-squares

Fitting principle: searching for the 'closest' model from data

$$\sum_{i=1}^{n}\Big(\overbrace{Y_i - \left[\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right]}^{\varepsilon_i = i\text{th residual}}\Big)^2 = \sum_{i=1}^{n}\varepsilon_i^2$$

A very convenient closed-form solution ... provided S_x^{-1} exists

$$\hat{\beta} = S_x^{-1} s_{xy}, \; \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}_1 - \ldots - \hat{\beta}_p \bar{x}_p$$

where S_x is the sample $p \times p$ variance matrix of the $x-$profile and s_{xy} is the sample $p-$vector of covariances between Y and the $x-$profile.

- Least-squares fitting in R

Assessment of the fit

Closeness between observed Y and fitted values $\hat{Y}$:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

using the residual sum-of-squares:

$$\text{RSS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

▶ Residual sum-of-squares in R

Assessment of the fit

Comparison with the null model:

$$\mathcal{M}_0 \quad : \quad Y = \beta_0 + \varepsilon, \ \text{ with } \mathrm{RSS}_0 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

using the R^2 coefficient:

$$\begin{aligned} R^2 &= \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}_0}, \\ &= \mathrm{Cor}^2(Y, \hat{Y}) \qquad \text{[alternatively]} \end{aligned}$$

▶ R^2 coefficient in R

Assessment of the fit

Closeness between observed Y and fitted values $\hat{Y}$:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

using the residual sum-of-squares:

$$\text{RSS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

▶ Residual sum-of-squares in R

Illustration with a K-class response

How to guess the place where coffee is produced from a physico-chemical profile?

- Y, the production site, takes six possible values y_k;
- Five physico-chemical variables x_j: concentrations in
  - Chlorogenic acids (CGA),
  - Cafeine,
  - Trigonelline,
  - Fat and
  - dry matter

Model for probabilities

Multinomial Logistic Linear Regression model

$$
\begin{aligned}
\log \frac{\mathbb{P}_x(Y = y_2)}{\mathbb{P}_x(Y = y_1)} &= \beta_0^{(2)} + \beta_1^{(2)} x_1 + \ldots + \beta_p^{(2)} x_p, \\
\log \frac{\mathbb{P}_x(Y = y_3)}{\mathbb{P}_x(Y = y_1)} &= \beta_0^{(3)} + \beta_1^{(3)} x_1 + \ldots + \beta_p^{(3)} x_p, \\
&\vdots \\
\log \frac{\mathbb{P}_x(Y = y_6)}{\mathbb{P}_x(Y = y_1)} &= \beta_0^{(6)} + \beta_1^{(6)} x_1 + \ldots + \beta_p^{(6)} x_p,
\end{aligned}
$$

where $\beta_0^{(k)}$ and $\beta^{(k)} = (\beta_1^{(k)}, \ldots, \beta_p^{(k)})'$ are the regression parameters

Maximum-likelihood (ML) estimation

Fitting principle: searching for the 'closest' model from data

'closest': the 'deviance' perspective

$$\begin{aligned} \ell_{x,y}(\beta) &= \mathbb{P}_{x_1}(Y = y_1)\ldots\mathbb{P}_{x_n}(Y = y_n), && \text{[Likelihood]} \\ \mathcal{D}_{x,y}(\beta) &= -2\log \ell_{x,y}(\beta), && \text{[Deviance]} \end{aligned}$$

Minimization of the deviance: No closed-form solution ... an iterative fitting algorithm is needed.

▶ Model fitting in R

Assessment of the fit

Closeness between estimated probabilities and observed classes:

- Using the explained deviance:

$$\mathcal{D} \;=\; \mathcal{D}_{x,y}(\hat{\beta}_0) - \mathcal{D}_{x,y}(\hat{\beta}).$$

  where $\mathcal{D}_{x,y}(\hat{\beta}_0)$ is the residual deviance of the null model.
- Comparing fitted and observed classes:

  Bayes rule: fitted class is the class with maximal estimated probability.

▶ Model assessment in R