Interpolation

Thomas Romary

Centre de géosciences, Equipe géostatistique

thomas.romary@minesparis.psl.eu

Interpolation

- Objective : predict the variable of interest at new locations
- Application example : cartography
- How to? : learn a predictive model

Some predictive models

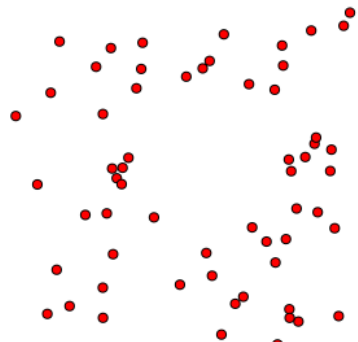
1. Regression models
   - Linear regression
   - ANOVA (Analysis of variance)
   - Local polynomial regression
   - ...
2. Interpolation methods
   - nearest-neighbor
   - N-nearest-neighbors
   - Inverse distance

Linear regression

- Data $Z = (z(x_1), \ldots, z(x_n))$ independent, X matrix of p explanatory variables (or predictors)
- Model $Z = X\beta + \varepsilon$,
  where ε is a vector of i.i.d. variables, centered with variance σ^2
- Solution $\widehat{\beta} = (X'X)^{-1}X'Z$
- Prediction $\forall x \in \mathcal{X}$, $Z(x) = X(x)'\widehat{\beta}$
- `python` function : `ols` in `statsmodels`
  prediction with the method `predict`

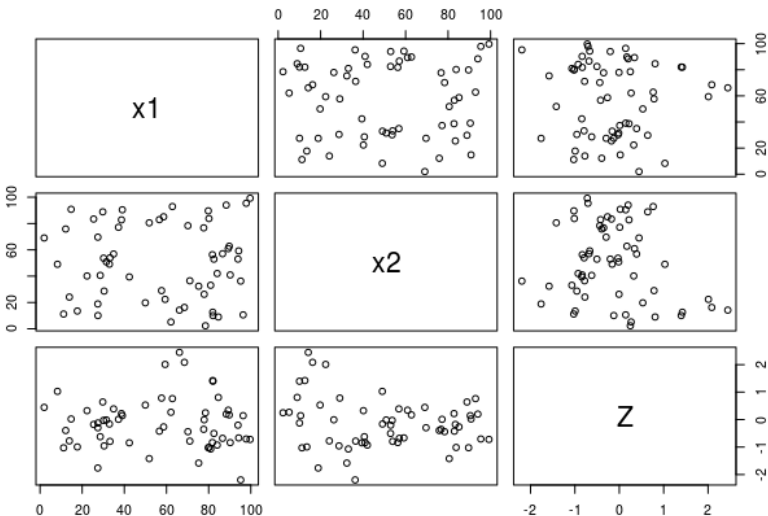
Linear Regression

Illustration



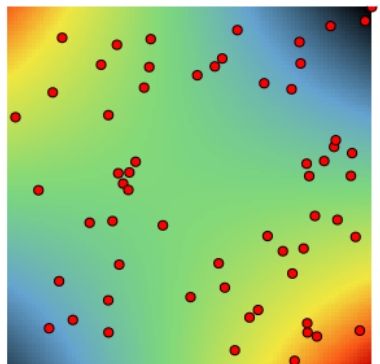
Linear Regression

Illustration



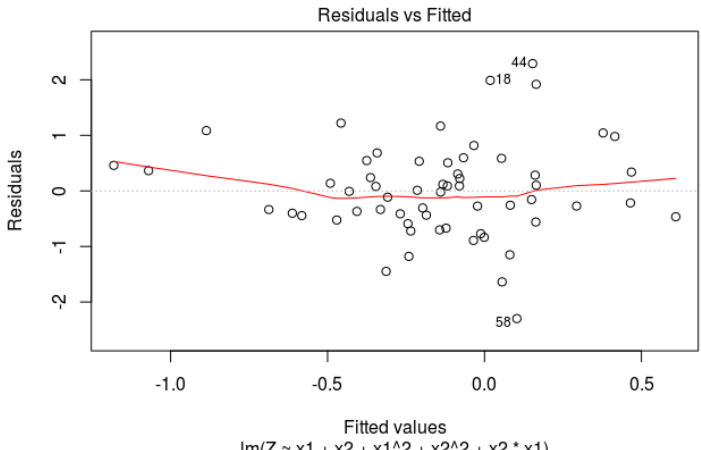
Linear Regression

Illustration – 2nd order polynomial



Linear Regression

Illustration — $R^2 = 0.21$



Reminder : coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(z_i - \widehat{z}_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z}_i)^2}$$

is an adjustment score

When the goal is to predict, we rely on *predictive scores*, computed on a validation set of data (not used for the fitting), e.g. the mean squared error :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(z_i - z_i^\star)^2$$

Reminder : coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(z_i - \widehat{z}_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z}_i)^2}$$

is an adjustment score

When the goal is to predict, we rely on *predictive scores*, computed on a validation set of data (not used for the fitting), e.g. the mean squared error :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(z_i - z_i^\star)^2$$

ANOVA

- The ANOVA (analysis of variance) is a particular case of the linear regression model, where the predictors are the indicators of being in a given set $A_i,\ i = 1, ..p$
- $X_{ij} = (\mathbb{1}_{Z_i \in A_j})$
- Model $Z = X\beta + \varepsilon$,
  where ε is a vector of i.i.d. variables, centered with variance σ^2
- Solution $\widehat{\beta} = (X'X)^{-1}X'Z$
- Prediction $\forall x \in \mathcal{X},\ Z(x) = X(x)'\widehat{\beta}$
- `python` function : `ols` in `statsmodels`
  `stats.anova_lm` for the analysis
  prediction with the method `predict`

Others

- Thin plate splines
  `thin-plate-spline` package
- Random forests
  `RandomForestRegressor` from package `sklearn.ensemble`
- Support vector regression
  `svm` from package `sklearn`
- ...

Remark

All these methods assume the independance between the residuals!

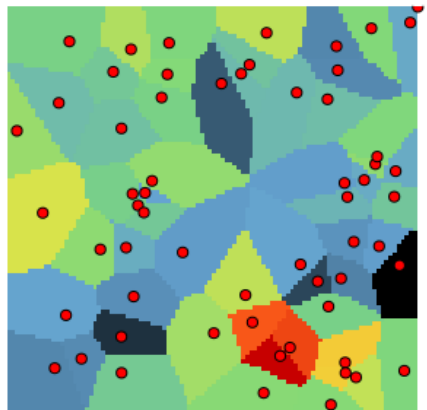
Nearest neighbor

- Objective : map the phenomenon under study
- Principle : we generally affect a weight equal to 1 to the nearest data point
- Model : $\forall x \in \mathcal{X}$,

$$Z(x) = \sum_{i=1}^{n} z(x_i) \mathbb{1}_{\|x_i - x\| = \min_{j=1,\dots,n}(\|x_j - x\|)}$$

Nearest neighbor

Illustration



N nearest neighbors

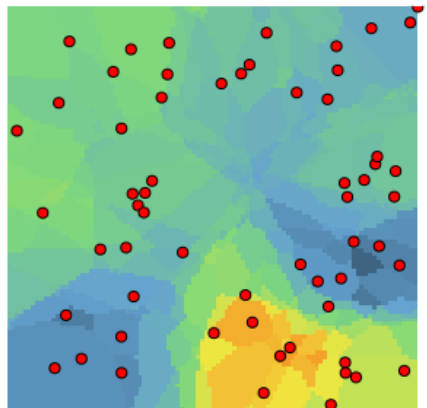
- Objective : map the phenomenon under study
- Principle : we generally affect a weight equal to $1/N$ to the N nearest data points
- Model : $\forall x \in \mathcal{X}$,

$$Z(x) = \sum_{i=1}^{n} \frac{1}{N} z(x_i) \mathbb{1}_{i \in V_x}$$

where V_x is the set of the N nearest neighbors to x

5 nearest neighbors

Illustration



Inverse distance

- Objective : map the phenomenon under study
- Principle : we affect a weight proportional to $\omega_i(x) = 1/\|x - x_i\|^\alpha$ to the data points
- Model : $\forall x \in \mathcal{X} \setminus \{x_1, \ldots, x_n\}$,

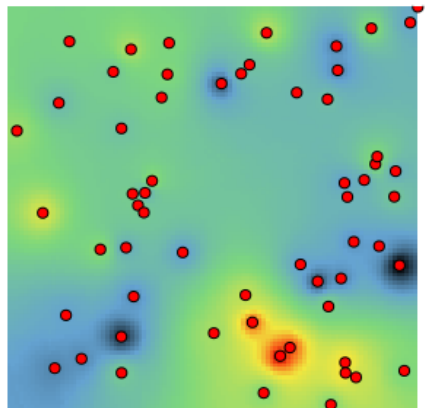
$$Z(x) = \frac{1}{\sum_{i=1}^n \omega_i(x)} \sum_{i=1}^n \omega_i(x) z(x_i) \mathbb{1}_{i \in V_x}$$

where V_x is a neighborhoud of x for instance defined by a maximum distance

$$Z(x_i) = z(x_i), i = 1, \ldots, n$$

Inverse distance

Illustration



Summary

- Regression methods require predictors that are well adapted to the problem (not always available or in sufficent number)
- The interpolation methods do not take into account the struture of the data