INFO-F-422: Statistical foundations of machine learning

Linear regression

Gianluca Bontempi

Machine Learning Group
Computer Science Department
mlg.ulb.ac.be

Beyond parameter estimation

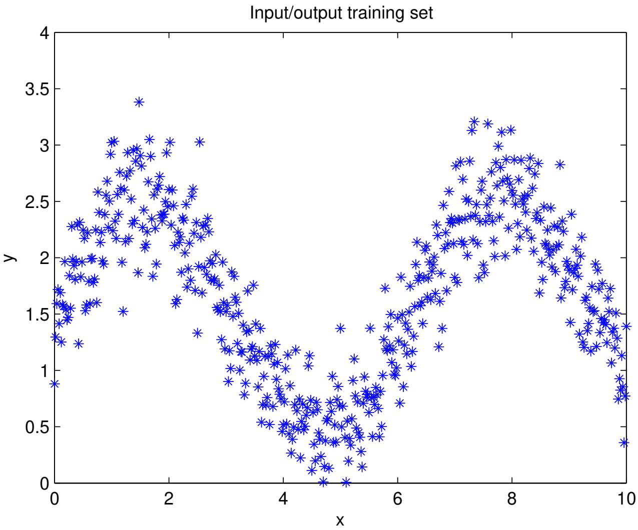
So far we focused on estimation of parameters of univariate distributions. Although the basic notions remain the same the main goal of estimation is to address more complex tasks.

- **Estimation of parameters of multivariate distributions:** consider for example multivariate gaussians.
- **Estimation of distribution and density functions:** this problem is also known as clustering or unsupervised learning.
- **Estimation of more complex functionals.**
- **Estimation of discrete conditional distributions:** this is also known as pattern recognition or pattern classification.
- **Estimation of regression curves.**

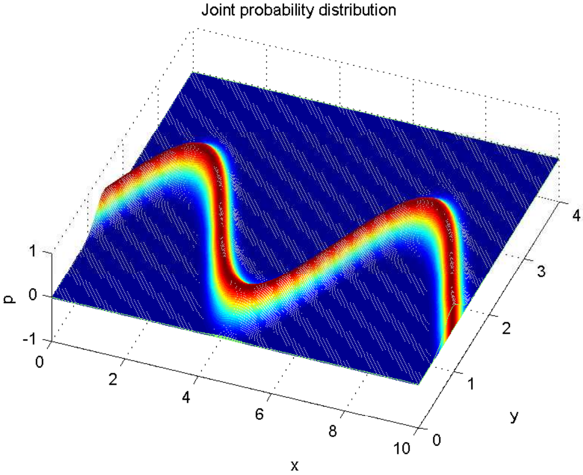
Recap of previous episodes...

- Model dependency between $\mathbf{x}$ and continuous or ordered $\mathbf{y}$ (e.g. study hours vs exam marks).
- Dataset (matrix) D_N of N i.i.d. samples
- Target: $\theta = E[\mathbf{y}|\mathbf{x} = x]$ or $\theta = \beta_1$ in a linear regression then
  1. We have to define an estimator $\hat{\boldsymbol{\theta}} = g(\mathbf{D}_N)$, i.e. the algorithm to convert data into estimates,
  2. We have to assess the accuracy (in terms of bias and variance) of such algorithm (e.g. to compare with others)
- For simple estimation tasks the bias and variance can be analytically derived.
- When this is not the case, resampling strategies (e.g. bootstrap) may help.
- Assessing an algorithm is an estimation problem itself!

Bivariate continuous random scatterplot



Bivariate density distribution



Prediction problems

Consider this set of prediction problems

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack, on the basis of demographic, diet and clinical measurements.
- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.
- Identify the risk factors for breast cancer, based on clinical, demographic and genetic variables.
- Classify the category of a text email (spam or not) on the basis of its text content.
- Characterize the mechanical property of a steel plate on the basis of its physical and chemical composition.

Input/output problems

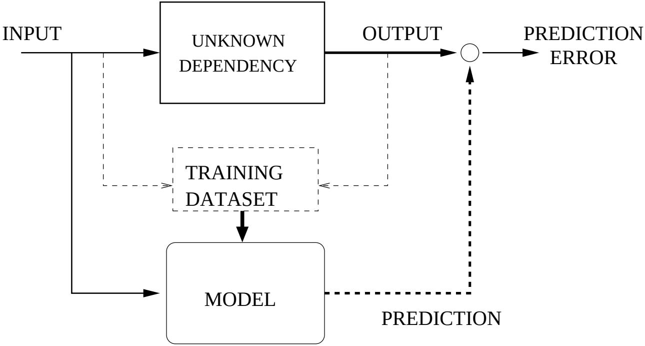
All the previous examples are characterized by

1. An outcome measurement, also called **output**, usually quantitative (like a stock price) or categorical (like heart attack/no heart attack).
2. a set of **features** or **inputs**, also quantitative or categorical, that we wish to use to predict the output.

Assumption: input variables provide some explanation for the variability of the output.

We collected a set of input/output data (**training set**), we use statistical methods to build a **prediction model** (**learner**) to predict the outcome for **new unseen objects**.

Supervised learning



Supervised learning because of the presence of the outcome variable which guides the learning process.
Collecting a set of training data is like having a teacher suggesting the correct answer for each input.

Regression and classification

According to the type of output, two prediction tasks:

- **Regression**: quantitative outputs, e.g. real or integer numbers
- **Classification (or pattern recognition)**: qualitative or categorical outputs which take values in a finite set of classes (e.g. black, white and red) where there is no explicit ordering. Qualitative variables are also referred to as **factors**.

The simple linear model

The simplest regression model is is the linear model

$$\mathbf{y} = \beta_0 + \beta_1 x + \mathbf{w}$$

where

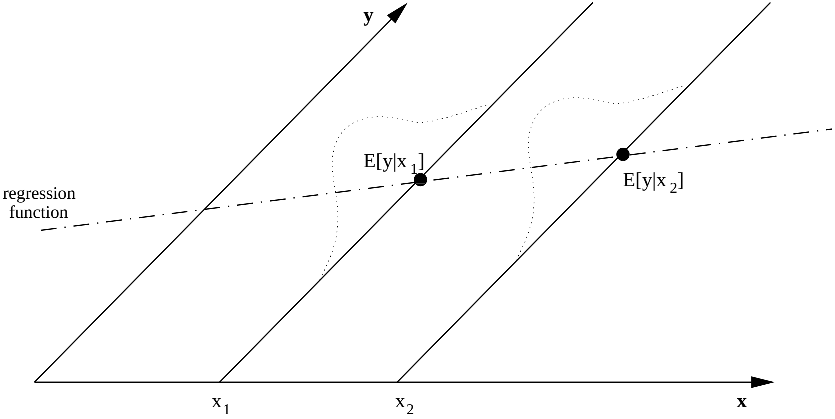
- $x \in \mathbb{R}$ is the regressor (or independent) variable,
- $\mathbf{y} \in \mathbb{R}$ is the measured response (or dependent) variable,
- β_0 is the intercept, β_1 is the slope
- $E[\mathbf{w}] = 0$ where $\mathbf{w}$ is the *model error*

This implies that

$$\begin{aligned} E[\mathbf{y}|x] &= f(x) = \beta_0 + \beta_1 x \\ \mathrm{Var}\,[\mathbf{y}|x] &= \mathrm{Var}\,[\mathbf{w}] \end{aligned}$$

Linear regression function

The function $f(x) = E[\mathbf{y}|x]$ is also known as *regression function*.



What does “linear” mean?

In the following we will intend as *linear model* each input/output relationships which is *linear in the model parameters* and not necessarily in the dependent variables. This means that

1. any value of the response variable y is described by a linear combination of a series of parameters (regression slopes, intercept)
2. no parameter appears as an exponent or is multiplied or divided by another parameter.

Example of linear models

According to our definition of linear model, then

- $y = \beta_0 + \beta_1 x$ is a linear model
- $y = \beta_0 + \beta_1 x^2$ is again a linear model. Simply by making the transformation $X = x^2$, the model can be put in in the linear form $y = \beta_0 + \beta_1 X$
- $y = B_0 x^{\beta_1}$ can be studied as a linear model between $Y = \log(y)$, $X = \log(x)$ and $\beta_0 = \log(B_0)$ thanks to the equality

$$\log(y) = \log(B_0) + \beta_1 \log(x) \Leftrightarrow Y = \beta_0 + \beta_1 X$$

- the relationship $y = \beta_0 + \beta_1 \beta_2^x$ cannot be linearized.
- let $\mathbf{z}$ a categorical variable taking 4 possible values $\{c_1, \ldots, c_4\}$. It is possible to model a linear dependence with $\mathbf{y}$ by creating four binary variables $\mathbf{x}_j$ such that $\mathbf{x}_j = 1 \Leftrightarrow \mathbf{z} = c_j$.

Model estimation

- Suppose that N pairs of observations $D_N = \{\langle x_i, y_i \rangle\}$, $i = 1, \dots, N$ are available.
- Let us assume that data are generated by the following stochastic process

$$y_i = \beta_0 + \beta_1 x_i + w_i, \qquad i = 1, \dots, N$$

  where

  1. the $\mathbf{w}_i$ are iid realizations of the r.v. $\mathbf{w}$ having mean zero and constant variance $\sigma_{\mathbf{w}}^2$ (homoscedasticity),
  2. the x_i are non random and observed with negligible error

- Then the unknown parameters (also known as *regression coefficients*) β_0 and β_1 can be estimated by the *least-squares method*.

Least squares formulation

The method of least squares is designed to provide

1. estimations $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1
2. the fitted values of the response y

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \dots, N$$

so that the **residual sum of squares**

$$\text{SSE}_{\text{emp}}(b_0, b_1) = \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2$$

is minimized.
See Shiny script `leastsquares.R`.

Least-squares solution

Since the error function $\text{SSE}_{\text{emp}}(b_0, b_1)$ is a quadratic function of the coefficients b_0 and b_1, the minimization of the error function has a unique solution which can be found in closed form.

This is called the *least-squares solution*

$$\{\hat{\beta}_0, \hat{\beta}_1\} = \arg \min_{\{b_0, b_1\}} \text{SSE}_{\text{emp}}(b_0, b_1) = \arg \min_{\{b_0, b_1\}} \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2$$

Empirical risk

From SSE_{emp} we can define the term

$$\widehat{\text{MISE}}_{\text{emp}} = \min_{\{b_0, b_1\}} \text{SSE}_{\text{emp}}(b_0, b_1) = \\ = \frac{\text{SSE}_{\text{emp}}(\hat{\beta}_0, \hat{\beta}_1)}{N} = \frac{\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{N}$$

is called the **empirical risk** or **training error**.

Note that the term SSE_{emp} is a function of the training set and as such it can be considered as a realization of a random variable.

Least squares solution

It can be shown that the least-squares solution is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}, \quad \bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

and

$$S_{xy} = \sum_{i=1}^{N} (x_i - \bar{x}) y_i$$

$$S_{xx} = \sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{i=1}^{N} (x_i - \bar{x}) x_i$$

Properties of the least-squares estimators

If the dependency underlying the data is linear then the estimators are unbiased. Since x is nonrandom and $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$

$$E_{\mathbf{D}_N}[\hat{\boldsymbol{\beta}}_1] = E_{\mathbf{D}_N}\left[\frac{S_{xy}}{S_{xx}}\right] = \sum_{i=1}^{N} \frac{(x_i - \bar{x})E[\mathbf{y}_i]}{S_{xx}}$$

$$= \frac{1}{S_{xx}}\left(\sum_{i=1}^{N}[(x_i - \bar{x})\beta_0] + \sum_{i=1}^{N}[(x_i - \bar{x})\beta_1 x_i]\right) = \frac{\beta_1 S_{xx}}{S_{xx}} = \beta_1$$

Also it can be shown that

$$\mathrm{Var}\left[\hat{\boldsymbol{\beta}}_1\right] = \frac{\sigma_{\mathbf{w}}^2}{S_{xx}}$$

$$E[\hat{\boldsymbol{\beta}}_0] = \beta_0$$

$$\mathrm{Var}\left[\hat{\boldsymbol{\beta}}_0\right] = \sigma_{\mathbf{w}}^2\left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Properties of the least-squares estimators (II)

- It can be shown that the **error mean-square**

$$\hat{\sigma}^2_{\mathbf{w}} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-2}$$

  is an unbiased estimator of $\sigma^2_{\mathbf{w}}$ *under the (strong) assumption that the linear model is correct*.
- The denominator is often referred to as the *residual degrees of freedom*, also denoted by df.
- The degree of freedom can be be seen as the number N of samples reduced by the numbers p of parameters estimated (slope and intercept).
- The estimate of the variance $\sigma^2_{\mathbf{w}}$ allows the estimation of the variance of the intercept and slope, respectively.

Sample correlation coefficient

The usual estimator of the correlation

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{Cov}[\mathbf{x}, \mathbf{y}]}{\sqrt{\mathrm{Var}\,[\mathbf{x}]\,\mathrm{Var}\,[\mathbf{y}]}}$$

between two r.v. $\mathbf{x}$ and $\mathbf{y}$ is

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Note that since

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

the following relation holds

$$\hat{\rho}^2 = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}}$$

Interval of confidence

- **Under the assumption of normal distribution** of **w** , we have then

$$\text{Prob}\left\{ -t_{\alpha/2,N-2} < \frac{(\hat{\beta}_1 - \beta_1)}{\hat{\sigma}} \sqrt{S_{xx}} < t_{\alpha/2,N-2} \right\} = 1 - \alpha$$

  where $t_{\alpha/2,N-2}$ is the upper $\alpha/2$ critical point of the $\mathcal{T}$-distribution with $N-2$ degrees of freedom.

- Equivalently we can say that with probability $1-\alpha$, the real parameter β_1 is covered by the interval described by

$$\hat{\beta}_1 \pm t_{\alpha/2,N-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Interval of confidence (II)

In a similar fashion the $100(1-\alpha)\%$ confidence interval on β_0 is

$$\hat{\beta}_0 \pm t_{\alpha/2,N-2}\hat{\sigma}\sqrt{\frac{1}{N}+\frac{\bar{x}^2}{S_{xx}}}$$

where $t_{\alpha/2,N-2}$ is the upper $\alpha/2$ critical point of the Student distribution with $N-2$ degrees of freedom.

Variance of the response

- Let

$$\hat{\mathbf{y}} = \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 x = \bar{\mathbf{y}} - \hat{\boldsymbol{\beta}}_1 \bar{x} + \hat{\boldsymbol{\beta}}_1 x = \bar{\mathbf{y}} + \hat{\boldsymbol{\beta}}_1 (x - \bar{x})$$

  be the estimator of the regression function value in x.

- Under the linear hypothesis, we have for a specific $x = x_0$

$$E[\hat{\mathbf{y}}|x_0] = E[\hat{\boldsymbol{\beta}}_0] + E[\hat{\boldsymbol{\beta}}_1]x_0 = \beta_0 + \beta_1 x_0 = E[\mathbf{y}|x_0]$$

- Since

$$\mathrm{Var}\left[\hat{\boldsymbol{\beta}}_1\right] = \frac{\sigma^2_{\mathbf{w}}}{S_{xx}}$$

  and $\mathrm{Cov}[\bar{\mathbf{y}}, \hat{\boldsymbol{\beta}}_1] = 0$, the variation of $\hat{\mathbf{y}}$ at x_0 if repeated data collection and consequent regressions were conducted is

$$\mathrm{Var}\left[\hat{\mathbf{y}}|x_0\right] = \mathrm{Var}\left[\bar{\mathbf{y}} + \hat{\boldsymbol{\beta}}_1 (x_0 - \bar{x})\right] = \sigma^2_{\mathbf{w}} \left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

  where $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$.

Multiple linear dependency

- Consider a linear relation between an independent variable $x \in \mathcal{X} \subset \mathbb{R}^n$ and a dependent random variable $\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}$

$$\mathbf{y} = \beta_0 + \beta_1 x_{\cdot 1} + \beta_2 x_{\cdot 2} + \cdots + \beta_n x_{\cdot n} + \mathbf{w}$$

  where $\mathbf{w}$ represents a random variable with mean zero and constant variance $\sigma^2_{\mathbf{w}}$.

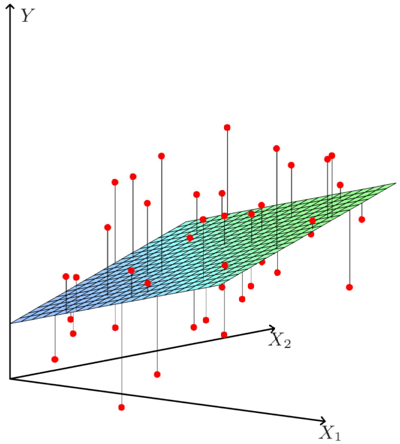
- In matrix notation the equation can be written as:

$$\mathbf{y} = x^T \beta + \mathbf{w}$$

  where x stands for the $[p \times 1]$ vector $x = [1, x_{\cdot 1}, x_{\cdot 2}, \ldots, x_{\cdot n}]^T$, $\beta = [\beta_0, \ldots, \beta_n]^T$ is the vector of parameters and $p = n + 1$ is the total number of model parameters.

- NB: in the following $x_{\cdot j}$ will denote the jth variable of the vector x, while x_i will denote the ith observation of the vector x.

The multiple linear regression model with $n = 2$



(excerpt from "The Elements of Statistical Learning " book)

The multiple linear regression model

Consider N observations $D_N = \{\langle x_i, y_i \rangle : i = 1, \dots, N\}$, where $x_i = (1, x_{i1}, \dots, x_{in})$, generated according to the previous model. We suppose that the following multiple linear relation holds

$$Y = X\beta + W$$

where Y is the $[N \times 1]$ response vector, X is the $[N \times p]$ *data matrix*, whose j^{th} column of X contains readings on the j^{th} regressor, β is the $[p \times 1]$ vector of parameters

$$Y = \begin{bmatrix} y_{\mathbf{1}} \\ y_{\mathbf{2}} \\ \vdots \\ y_N \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{\mathbf{11}} & x_{\mathbf{12}} & \dots & x_{\mathbf{1}n} \\ 1 & x_{\mathbf{21}} & x_{\mathbf{22}} & \dots & x_{\mathbf{2}n} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{N\mathbf{1}} & x_{N\mathbf{2}} & \dots & x_{Nn} \end{bmatrix} = \begin{bmatrix} x_{\mathbf{1}}^T \\ x_{\mathbf{2}}^T \\ \vdots \\ x_N^T \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_{\mathbf{0}} \\ \beta_{\mathbf{1}} \\ \vdots \\ \beta_n \end{bmatrix} \quad W = \begin{bmatrix} w_{\mathbf{1}} \\ w_{\mathbf{2}} \\ \vdots \\ w_N \end{bmatrix}$$

where w_i are assumed uncorrelated, with mean zero and constant variance $\sigma^2_{\mathbf{w}}$ (homogeneous variance). Then $\mathrm{Var}\left[\mathbf{w}_1, \dots, \mathbf{w}_N\right] = \sigma^2_{\mathbf{w}} I_N$.

The least-squares solution

We seek the **the least-squares estimator** $\hat{\boldsymbol{\beta}}$ such that

$$\hat{\beta} = \arg\min_b \sum_{i=1}^{N} (y_i - x_i^T b)^2 = \arg\min_b \left((Y - Xb)^T (Y - Xb) \right)$$

Given $\hat{\boldsymbol{\beta}}$ we obtain

$$\text{SSE}_{\text{emp}} = \left((Y - X\hat{\boldsymbol{\beta}})^T (Y - X\hat{\boldsymbol{\beta}}) \right) = e^T e$$

where SSE_{emp} represents the **residual sum of squares** for linear models and e is the $[N \times 1]$ vector of residuals. We define also the **empirical (or training) error** quantity

$$\widehat{\text{MISE}}_{\text{emp}} = \frac{\text{SSE}_{\text{emp}}}{N}$$

The vector $\hat{\beta}$ must satisfy

$$\frac{\partial}{\partial \hat{\beta}} [(Y - X\hat{\beta})^T (Y - X\hat{\beta})] = 0 \Leftrightarrow -2X^T (Y - X\hat{\beta}) = 0$$

Normal equations

Differentiating the residual sum of squares we obtain the *least-squares normal equations*

$$(X^T X)\hat{\beta} = X^T Y$$

As a result, assuming X is of full column rank

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

where the $X^T X$ matrix is a positive definite symmetric $[p \times p]$ matrix which plays an important role in multiple linear regression. The predicted values for the training set are

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y$$

where $H = X(X^T X)^{-1} X^T$ is also known as the Hat matrix.

In R notation:

```
betahat=solve(t(X)%*%X) %*%t(X)%*%Y
```

R function lm

```
 summary(lm(Y~X))

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40141 -0.14760 -0.02202  0.03001  0.43490

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.09781    0.11748   9.345 6.26e-09
X            0.02196    0.01045   2.101   0.0479

(Intercept) ***
X           *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.2167 on 21 degrees of freedom
Multiple R-Squared: 0.1737,     Adjusted R-squared: 0.1343
F-statistic: 4.414 on 1 and 21 DF,  p-value: 0.0479
```

Analysis of the LS estimate

If the linear dependency assumption holds

- If $E[\mathbf{w}] = 0$ then $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of β.
- The *residual mean square* estimator

$$\hat{\boldsymbol{\sigma}}^2 = \frac{(Y - X\hat{\boldsymbol{\beta}})^T (Y - X\hat{\boldsymbol{\beta}})}{N - p}$$

  is an unbiased estimator of the error variance $\sigma_{\mathbf{w}}^2$.
- If the $\mathbf{w}_i$ are uncorrelated and have common variance, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma_{\mathbf{w}}^2 (X^T X)^{-1}$$

- R script `bv_mult.R`.

Variance of the prediction

- The prediction $\hat{\mathbf{y}}$ for a generic input value $x = x_0$ is unbiased

$$E[\hat{\mathbf{y}}|x_0] = x_0^T \beta$$

- The variance of the prediction $\hat{\mathbf{y}}$ for a generic input value $x = x_0$ is given by

$$\text{Var}[\hat{\mathbf{y}}|x_0] = \sigma_{\mathbf{w}}^2 x_0^T (X^T X)^{-1} x_0$$

- Assuming a normal error $\mathbf{w}$, the $100(1-\alpha)\%$ confidence bound for the regression value $E[\hat{\mathbf{y}}|x = x_0]$ is given by

$$\hat{y}(x_0) \pm t_{\alpha/2, N-p} \hat{\sigma}_{\mathbf{w}} \sqrt{x_0^T (X^T X)^{-1} x_0}$$

where $t_{\alpha/2,N-p}$ is the upper $\alpha/2$ percent point of the t-distribution with $N-p$ degrees of freedom and the quantity $\hat{\sigma}_{\mathbf{w}} \sqrt{x_0^T (X^T X)^{-1} x_0}$ is the *standard error of prediction* for multiple regression.

Generalization error of the linear model

- The linear predictor

  $$\hat{y} = x^T \hat{\beta}$$

  has been estimated by using the training dataset $D_N = \{\langle x_i, y_i \rangle : i = 1, \dots, N\}$. Then $\hat{\boldsymbol{\beta}}$ is a r.v.
- Now we want to use it to predict for a given input x the future output $\mathbf{y}(x)$.
- The future output $\mathbf{y}(x)$ is independent of the training set $\mathbf{D}_N$.
- Which precision can we expect from $\hat{y}(x_i) = x_i^T \hat{\beta}$ on average?
- A measure of error is the MSE

  $$\text{MSE}(x) = E_{\mathbf{D}_N, \mathbf{y}}[(\mathbf{y}(x) - x^T \hat{\boldsymbol{\beta}})^2] = \sigma_{\mathbf{w}}^2 + E_{\mathbf{D}_N}[(x^T \beta - x^T \hat{\boldsymbol{\beta}})^2]$$

  where $\mathbf{y}(x_i)$ is **independent** of $\mathbf{D}_N$ and the integrated version

  $$\text{MISE} = \int_{\mathbf{X}} \text{MSE}(x) p(x) dx$$

- How can we estimate this quantity?

The expected empirical error

- Is the empirical risk a good estimate of the generalization error?
- The expectation of the residual sum of squares can be written as[1]

$$E_{\mathbf{D}_N}\left[\widehat{\mathbf{MISE}}_{\text{emp}}\right] = E_{\mathbf{D}_N}\left[\frac{\sum_{i=1}^N(\mathbf{y}_i - x_i^T\hat{\beta})^2}{N}\right] =$$

$$= \frac{N-p}{N}\, E_{\mathbf{D}_N}\left[\frac{\sum_{i=1}^N(\mathbf{y}_i - x_i^T\hat{\beta})^2}{N-p}\right] = \frac{N-p}{N}\sigma_{\mathbf{w}}^2$$

- This is the expectation of the error made by a linear model trained on D_N to predict the value of the output in D_N.

[1] derivation in the handbook

MISE error

- Let us compute now the expected prediction error of a linear model trained on D_N when this is used to predict a set of outputs distributed according to the same linear dependency but independent of the training set.
- It can be shown[2] that in case of linear dependency

$$\mathrm{MISE} = E_{\mathbf{D}_N,\mathbf{y}}\left[\frac{\sum_{i=1}^{N}(\mathbf{y} - x_i^T\hat{\beta})^2}{N}\right] = \frac{N+p}{N}\sigma_{\mathbf{w}}^2$$

Note that in the MISE formula the $\mathbf{y}$ distribution is independent of $\mathbf{D}_N$ and then of $\hat{\beta}$.

[2] derivation in the handbook

MISE error

Then it follows that the empirical error returns a biased estimate of MISE, that is

$$E_{\mathbf{D}_N}[\widehat{\mathbf{MISE}}_{\text{emp}}] = \frac{N-p}{N}\sigma_{\mathbf{w}}^2 \neq \text{MISE} = \frac{N+p}{N}\sigma_{\mathbf{w}}^2$$

If we replace $\widehat{\text{MISE}}_{\text{emp}}$ with

$$\widehat{\text{MISE}}_{\text{emp}} + 2\frac{p}{N}\sigma_{\mathbf{w}}^2$$

we obtain an unbiased estimator of the quantity MISE (see R file `ee.R`).
Nevertheless, this estimator requires an estimate of the noise variance.

The PSE and the FPE

- Given an a priori estimate $\hat{\sigma}^2_{\mathbf{w}}$ we have the **Predicted Square Error (PSE)** criterion

$$\mathrm{PSE} = \widehat{\mathrm{MISE}}_{\mathrm{emp}} + 2\hat{\sigma}^2_{\mathbf{w}} p/N$$

- Taking as estimate of $\sigma^2_{\mathbf{w}}$

$$\hat{\sigma}^2_{\mathbf{w}} = \frac{1}{N-p}\mathrm{SSE}_{\mathrm{emp}}$$

  we have the **Final Prediction Error (FPE)**

$$\mathrm{FPE} = \frac{1+p/N}{1-p/N}\widehat{\mathrm{MISE}}_{\mathrm{emp}}$$

- See the R script `fpe.R`

Dual linear formulation

Consider a linear regression problem with $[N, n]$ input matrix X and $[N, 1]$ output vector y. The conventional least-squares solution is the $[n, 1]$ parameter vector

$$\hat{\beta} = (X'X)^{-1}X'y$$

and the prediction for a new $[n, 1]$ vector x is returned by

$$\hat{y} = \hat{\beta}'x = \langle \hat{\beta}, x \rangle$$

The dual formulation is

$$\hat{\beta} = (X'X)^{-1}X'y = X'\underbrace{X(X'X)^{-2}X'y}_{\alpha} = X'\alpha = \sum_{i=1}^{N} \alpha_i x_i$$

where α is a $[N, 1]$ vector and x_i is the $[n, 1]$ vector which represents the ith observation.

Recursive least-squares

- The least-squares estimator for a training set of N samples has the form:

$$\hat{\beta}_{(N)} = (X_{(N)}^T X_{(N)})^{-1} X_{(N)}^T Y_{(N)}$$

- The subscript (N) is added to denote the number of samples used for the estimation.
- Suppose that a new sample $\langle x_{N+1}, y_{N+1} \rangle$ becomes available. Instead of recomputing the estimate $\hat{\beta}_{(N+1)}$ by using all the $N+1$ available data we want to derive $\hat{\beta}_{(N+1)}$ as an update of $\hat{\beta}_{(N)}$.
- This is the problem solved by the **recursive least-squares (RLS) identification**.

If a single new example $\langle x_{N+1}, y_{N+1} \rangle$ is added to the training set where x_{N+1} is a $[1, p]$ vector , the X matrix acquires a new row and $\hat{\beta}_{(N+1)}$ can be expressed as:

$$\hat{\beta}_{(N+1)} = \left(\left[\begin{array}{c} X_{(N)} \\ x_{N+1} \end{array} \right]^T \left[\begin{array}{c} X_{(N)} \\ x_{N+1} \end{array} \right] \right)^{-1} \left[\begin{array}{c} X_{(N)} \\ x_{N+1} \end{array} \right]^T \left[\begin{array}{c} Y_{(N)} \\ y_{N+1} \end{array} \right]$$

By denoting the $[p, p]$ matrix

$$S_{(N)} = (X_{(N)}^T X_{(N)})$$

we have

$$S_{(N+1)} = (X_{(N+1)}^T X_{(N+1)}) = \left(\left[X_{(N)}^T x_{N+1}^T \right] \left[\begin{array}{c} X_{(N)} \\ x_{N+1} \end{array} \right] \right) =$$

$$= \left(X_{(N)}^T \; X_{(N)} + x_{N+1}^T x_{N+1} \right) = S_{(N)} + x_{N+1}^T x_{N+1}$$

$$\begin{bmatrix} X_{(N)} \\ x_{N+1} \end{bmatrix}^T \begin{bmatrix} Y_{(N)} \\ y_{N+1} \end{bmatrix} = X_{(N)}^T Y_{(N)} + x_{N+1}^T y_{N+1}$$

and

$$S_{(N)} \hat{\beta}_{(N)} = (X_{(N)}^T X_{(N)}) \left[(X_{(N)}^T X_{(N)})^{-1} X_{(N)}^T Y_{(N)} \right] = X_{(N)}^T Y_{(N)}$$

imply

$$\begin{aligned} S_{(N+1)} \hat{\beta}_{(N+1)} = \begin{bmatrix} X_{(N)} \\ x_{N+1} \end{bmatrix}^T \begin{bmatrix} Y_{(N)} \\ y_{N+1} \end{bmatrix} &= S_{(N)} \hat{\beta}_{(N)} + x_{N+1}^T y_{N+1} = \\ = \left(S_{(N+1)} - x_{N+1}^T x_{N+1} \right) \hat{\beta}_{(N)} + x_{N+1}^T y_{N+1} &= \\ = S_{(N+1)} \hat{\beta}_{(N)} - x_{N+1}^T x_{N+1} \hat{\beta}_{(N)} + x_{N+1}^T y_{N+1} & \end{aligned}$$

or equivalently

$$\hat{\beta}_{(N+1)} = \hat{\beta}_{(N)} + S_{(N+1)}^{-1} x_{N+1}^T (y_{N+1} - x_{N+1} \hat{\beta}_{(N)})$$

1st Recursive formulation

Then, we have the following recursive formulation

$$\begin{cases} S_{(N+1)} & = S_{(N)} + x_{N+1}^T x_{N+1} \\ \gamma_{(N+1)} & = S_{(N+1)}^{-1} x_{N+1}^T \\ e & = y_{N+1} - x_{N+1}\hat{\beta}_{(N)} \\ \hat{\beta}_{(N+1)} & = \hat{\beta}_{(N)} + \gamma_{(N+1)} e \end{cases}$$

- $\hat{\beta}_{(N+1)}$ can be expressed as a function of the old estimate $\hat{\beta}_{(N)}$ and the new sample $\langle x_{N+1}, y_{N+1} \rangle$.
- This formulation requires the inversion of the $[p \times p]$ matrix $S_{(N+1)}$.
- This operation is computationally expensive but, fortunately, using a matrix inversion theorem, an incremental formula for S^{-1} can be found.

The matrix inversion formula

- Let us consider the four matrices F, G, H and K and the matrix $F + GHK$. Assume that the inverses of the matrices F, G and $(F + GHK)$ exist. Then

$$(F + GHK)^{-1} = F^{-1} - F^{-1}G\left(H^{-1} + KF^{-1}G\right)^{-1}KF^{-1}$$

- Consider the case where F is a $[n \times n]$ square nonsingular matrix, $G = z$ where z is a $[n \times 1]$ vector, $K = z^T$ and $H = 1$. Then the formula simplifies to

$$(F + zz^T)^{-1} = F^{-1} - \frac{F^{-1}zz^TF^{-1}}{1 + z^TFz}$$

where the denominator in the right hand term is a scalar.

2nd Recursive formulation

Once defined

$$V_{(N)} = S_{(N)}^{-1} = (X_{(N)}^T X_{(N)})^{-1}$$

we have $(S_{(N+1)})^{-1} = (S_{(N)} + x_{N+1}^T x_{N+1})^{-1}$ and

$$V_{(N+1)} = V_{(N)} - \frac{V_{(N)} x_{N+1}^T x_{N+1} V_{(N)}}{1 + x_{N+1} V_{(N)} x_{N+1}^T}$$

This leads to a second recursive formulation:

$$\begin{cases} V_{(N+1)} & = V_{(N)} - \frac{V_{(N)} x_{N+1}^T x_{N+1} V_{(N)}}{1+x_{N+1} V_{(N)} x_{N+1}^T} \\ \gamma_{(N+1)} & = V_{(N+1)} x_{N+1}^T \\ e & = y_{N+1} - x_{N+1} \hat{\beta}_{(N)} \\ \hat{\beta}_{(N+1)} & = \hat{\beta}_{(N)} + \gamma_{(N+1)} e \end{cases}$$

RLS initialization

RLS needs the initial values $\hat{\beta}_{(0)}$ and $V_{(0)}$. One way to avoid choosing these initial values is to collect the first N data points, to solve $\hat{\beta}_{(N)}$ and $V_{(N)}$ directly from

$$
\begin{aligned}
V_{(N)} &= (X_{(N)}^T X_{(N)})^{-1} \\
\hat{\beta}_{(N)} &= V_{(N)} X_{(N)}^T Y_{(N)}
\end{aligned}
$$

and to start iterating from the $N + 1^{\text{th}}$ point. Otherwise, in case of a generic initialization $\hat{\beta}_{(0)}$ and $V_{(0)}$ we have the following relations

$$
\begin{aligned}
V_{(N)} &= (V_{(0)} + X_{(N)}^T X_{(N)})^{-1} \\
\hat{\beta}_{(N)} &= V_{(N)} (X_{(N)}^T Y_{(N)} + V_{(0)}^{-1} \hat{\beta}_{(0)})
\end{aligned}
$$

RLS initialization (II)

- A common choice is to put

$$V_{(0)} = aI, \qquad a > 0$$

- Since $V_{(0)}$ represents the variance of the estimator to choose a very large a is equivalent to consider the initial estimation of β as very uncertain. In Bayesian terms this means that our priori is very weak.
- By setting a equal to a large number the RLS algorithm will diverge very rapidly from the initialization $\hat{\beta}_{(0)}$.
- Therefore, we can force the RLS variance and parameters to be arbitrarily close to the ordinary least-squares values, regardless of $\hat{\beta}_{(0)}$.
- However, in practice $\hat{\beta}_{(0)}$ is usually put equal to a zero vector.

RLS with forgetting factor

Consider the two situations

- the phenomenon underlying the data is linear but non stationary
- the phenomenon underlying the data is stationary and nonlinear but can be approximated by a linear model *locally in time*. This is of large use in adaptive control problems.

In these cases it is useful not give always the same importance to all the historical data but assign higher weights to more recent data (and forget older data).
RLS techniques can deal with these situations by a formulation with forgetting factor $\mu < 1$.

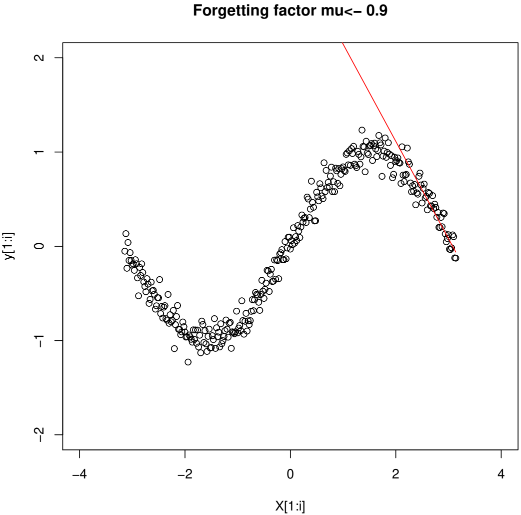
RLS with forgetting factor (II)

$$\begin{cases} V_{(N+1)} & = \frac{1}{\mu}\left(V_{(N)} - \frac{V_{(N)}x_{N+1}^T x_{N+1} V_{(N)}}{1+x_{N+1}V_{(N)}x_{N+1}^T}\right) \\ \gamma_{(N+1)} & = V_{(N+1)}x_{N+1}^T \\ e & = y_{N+1} - x_{N+1}\hat{\beta}_{(N)} \\ \hat{\beta}_{(N+1)} & = \hat{\beta}_{(N)} + \gamma_{(N+1)}e \end{cases}$$

- The smaller μ, the higher the forgetting.
- Note that for $\mu = 1$ we have the conventional RLS formulation.

TP: R script for RLS

R script `lin_rls.R`


Forgetting factor mu<– 0.9
y[1:i]
X[1:i]