HMMA 307 : Advanced Linear Modeling

Chapter 1 : Linear regression

Emma Santinelli Mégane Diéval Yassine Sayd

https://github.com/MegDie/advanced_lm_introduction

Université de Montpellier

Table of Contents

Table of Contents

Model

Observations: n samples $(y_i, x_i)_{i=1,\dots,n}$ with p features.
The model can be written in matrix notation as :

$$\boxed{y = X\beta + \varepsilon}$$

where

- $X = [\mathbf{x}_1, \dots, \mathbf{x}_p] = [x_1^\top, \dots, x_n^\top]^\top$ is an $n \times p$ matrix of covariates/features
- β is a $p \times 1$ vector of unknown parameters
- ε is a vector of *i.i.d.* random normal errors with mean 0

(Ordinary) Least squares: $\hat{\beta}^{\mathrm{LS}}$

The **LS** estimator is any coefficient vector $\hat{\beta}^{\mathrm{LS}} \in \mathbb{R}^p$ such that :

$$\hat{\beta}^{\mathrm{LS}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \underbrace{\frac{1}{2n}\|y - X\beta\|^2}_{f(\beta)}$$

and

$$f(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \frac{1}{2n}(X\beta)_i)^2 = \beta^\top \frac{X^\top X}{2n}\beta + \frac{1}{2n}\|y\|^2 - \langle y, X\beta\rangle$$

where $\langle y, X\beta\rangle = y^\top X\beta = \beta^\top X^\top y = \langle \beta, X^\top y\rangle$

Rem: $1/2$ is convenient for optimization (computing gradients), and $1/n$ for if $n \to \infty$ (p fixed) then the objective function convergences to something like $\mathbb{E}\left[\frac{1}{2n}(y_\infty - x_\infty^\top \beta\right]^2$.

Gram Matrix

Notation

The matrix $\hat{\Sigma} = \frac{X^\top X}{n}$ is called the **Gram** matrix.

$$X^\top X = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_p^\top \end{pmatrix} (\mathbf{x}_1, \ldots, \mathbf{x}_p)$$

Elementwise Gram matrix : $[X^\top X]_{j,j'} = [\langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle]_{(j,j') \in [\![1,p]\!]^2}$

Standarization: centering

Feature centering[1]:

Compute the columns sample means:

$$\bar{\mathbf{x}}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \qquad (*)$$

Use Equation (*) to get the centered matrix:

$$\mathrm{centering}(X) := \bar{X} = X - [\bar{\mathbf{x}}_1 \mathbb{1}_n, \ldots, \bar{\mathbf{x}}_p \mathbb{1}_n]$$

where $\mathbb{1}_n = (1, \ldots, 1)^\top$

Rem: $\bar{X}$ has columns with zero means

(1) the average is performed along samples!

Standardization: scaling

Feature scaling (reduction):

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{\mathbf{x}}_j)^2$$

and then one can define:

$$S_X = \mathrm{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_p)$$

To get the standardized matrix:

$$\begin{aligned}\mathrm{stdzing}(X) &= \mathrm{center}(X) \cdot S_X^{-1} \\ &= \left[\frac{\mathbf{x}_a - \bar{\mathbf{x}}_a \mathbb{1}_n}{\hat{\sigma}_a}, \ldots, \frac{\mathbf{x}_p - \bar{\mathbf{x}}_p \mathbb{1}_n}{\hat{\sigma}_p}\right]\end{aligned}$$

Rem: `sklearn`[2] convention is $1/n$ (could have been $1/(n-1)$)

---

[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Optimization

First Order Optimality Conditions

Since f is differentiable over $\mathbb{R}^p$, the following holds:

$$\nabla f(\hat{\beta}^{\mathrm{LS}}) = 0$$

Rem: If f is even C^∞ a function

Rem: f is a convex function so a local minimum is a global one

Conclusion: $\hat{\beta}^{\mathrm{LS}}$ satisfies the following equations of orthogonality :

$$\begin{aligned}
\nabla f(\hat{\beta}^{\mathrm{LS}}) = 0 &\iff \frac{X^\top X}{n}\hat{\beta}^{\mathrm{LS}} - \frac{X^\top y}{n} = 0 \\
&\iff X^\top\left(\frac{X\hat{\beta}^{\mathrm{LS}} - y}{n}\right) = 0 \\
&\iff X^\top(y - X\hat{\beta}^{\mathrm{LS}}) = 0 \\
&\iff \langle \mathbf{x}_j, y - X\beta \rangle = 0, \forall j \in [\![1, p]\!]
\end{aligned}$$

High dimension warning

When $p > n$ **and** $\mathrm{rank}(X) \leq n$, then, $\hat{\beta}^{\mathrm{LS}}$ is not unique

Rem: this happens when $\hat{\Sigma}$ is singular

Normal equations

Interpretation

Each feature is orthogonal to the residuals $r = y - X\hat{\beta}^{\mathrm{LS}}$:

$$\forall j \in [\![1, p]\!], \langle r, \mathbf{x}_j \rangle = 0$$

The LS vector $\hat{\beta}^{\mathrm{LS}}$ is a solution of a $p \times p$ linear system β:

$$\hat{\Sigma}\beta = \frac{X^\top y}{n}$$

Rem:

- $\hat{\Sigma}$ is invertible $\Rightarrow$ the solution of the linear system is unique
- $\hat{\Sigma}$ is invertible $\Rightarrow$ $\hat{\Sigma}$ is positive definite
- $\hat{\Sigma}$ invertible $\Rightarrow \mathrm{rank}(\hat{\Sigma}) = p$
- we assume that we have a full rank column e.g. :
  $$\mathrm{rank}(X) = \dim(\mathrm{vect}(X_1, \dots, X_p)) \leq n$$

The full column rank case

Theorem

If $\mathrm{rank}(X) = p$, then $\hat{\Sigma}$ is invertible and one has

$$\hat{\beta}^{\mathrm{LS}} = (X^\top X)^{-1} X^\top y$$

Proof:

$$\hat{\beta}^{\mathrm{LS}} = \hat{\Sigma}^{-1} \frac{X^\top y}{n} = \left(\frac{X^\top X}{n} \right)^{-1} \frac{X^\top y}{n}$$

<u>Rem</u>: In practice you hardly ever invert $\hat{\Sigma}$, but rather solve a linear system (inverting = solving p systems here, when one is enough)

Data analysis

Motivation

Using ordinary least squares models on two datasets:

- Bicycle accidents
- Count data of bicycles

We propose to estimate the severity of accidents by the feature "sexe". The problem is that the features are qualitative:

- Modalities of the feature to predict: "0 - Indemne", "1 - Blessé léger", "Blessé hospitalisé", and "3 - Tué"
- Modalities of the feature "sexe": "M" and "F"

Source: see associated code

Data analysis

First: convert features into ordinal features.

Prediction principle

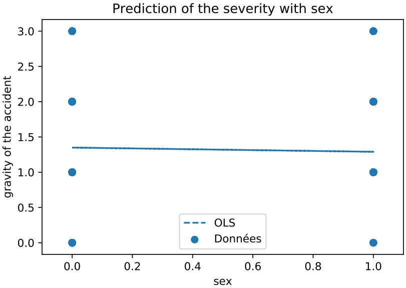
Calculate the coefficients β on a training sample and predict on a test sample the feature of interest. 0 is the value for male and 1 is the value for female.

Out[192]:

OLS Regression Results

Dep. Variable:	grave_quanti	**R-squared:**	0.00
Model:	OLS	**Adj. R-squared:**	0.00
Method:	Least Squares	**F-statistic:**	101.
Date:	Sat, 03 Oct 2020	**Prob (F-statistic):**	7.12e-2
Time:	17:58:35	**Log-Likelihood:**	-63570
No. Observations:	64515	**AIC:**	1.271e+0
Df Residuals:	64513	**BIC:**	1.272e+0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3492	0.003	458.306	0.000	1.343	1.355
sex_quanti	-0.0595	0.006	-10.079	0.000	-0.071	-0.048



Data analysis

Conclusion: The prediction is very bad on qualitative features. We notice that the R^2 is closed to 0 and it's mostly the same with the others qualitative features. With this dataset, the OLS model is not efficient for qualitative features.

Prediction of a quantitative feature

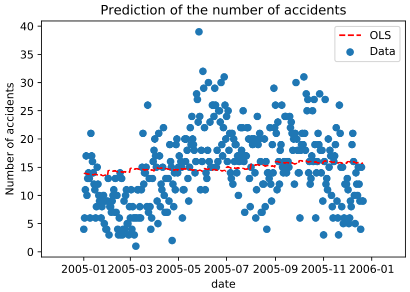
Predict the number of accidents with the date (day, month and year) that is an ordinal feature with periodic component. Results are also very bad.

Out[99]:

OLS Regression Results

Dep. Variable:	accidents	R-squared:	0.059
Model:	OLS	Adj. R-squared:	0.058
Method:	Least Squares	F-statistic:	62.23
Date:	Sun, 04 Oct 2020	Prob (F-statistic):	3.83e-63
Time:	13:39:42	Log-Likelihood:	-16186.
No. Observations:	5000	AIC:	3.238e+04
Df Residuals:	4994	BIC:	3.242e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	703.6676	44.309	15.881	0.000	616.803	790.533
day	-0.0118	0.010	-1.180	0.238	-0.031	0.008
month	0.1591	0.028	5.740	0.000	0.105	0.213
year	-0.3438	0.022	-15.610	0.000	-0.387	-0.301
periodic_day	-0.0580	0.123	-0.471	0.638	-0.300	0.183
periodic_month	-0.2805	0.131	-2.140	0.032	-0.537	-0.024



Data analysis

Same thing on the second dataset

Prediction of the number of bicycles in a day with the date and the total number of bicycles. We introduce also periodic components.

Out[82]:

OLS Regression Results

Dep. Variable:	Day_total	**R-squared:**	0.256
Model:	OLS	**Adj. R-squared:**	0.246
Method:	Least Squares	**F-statistic:**	25.33
Date:	Sun, 04 Oct 2020	**Prob (F-statistic):**	3.53e-10
Time:	15:02:24	**Log-Likelihood:**	-1119.3
No. Observations:	150	**AIC:**	2245.
Df Residuals:	147	**BIC:**	2254.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	345.6597	69.136	5.000	0.000	209.031	482.288
num	5.7071	0.802	7.113	0.000	4.122	7.293
sinus_num	19.0227	49.163	0.387	0.699	-78.134	116.180

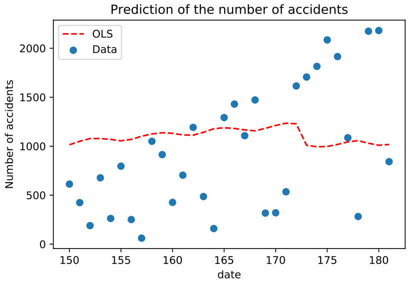


Table of Contents

SVD

Reminder: Let $\Sigma \in \mathbb{R}^{p \times p}$, if $\Sigma^\top = \Sigma$ then Σ is diagonalizable.

Theorem

For all matrix $M \in \mathbb{R}^{m_1 \times m_2}$ of $\operatorname{rank}(r)$, there exist two orthogonal matrix $U \in \mathbb{R}^{m_1 \times r}$ and $V \in \mathbb{R}^{m_2 \times r}$ such that :

$$M = U \operatorname{diag}(s_1, \dots, s_r) U^\top$$

where $s_1 \geq s_2 \geq \cdots \geq s_r \geq 0$ are the singular values of M.

Rem: $M = \sum_{j=1}^r s_j u_j v_j^\top$ with : $U = [u_1, \dots, u_r]$ et $V = [v_1 \dots v_r]$

Pseudo-inverse

Definition

For $M \in \mathbb{R}^{m_1 \times m_2}$, a pseudoinverse of M is defined as a matrix M^+ satisfying :

$$M^+ = V \operatorname{diag}\left(\frac{1}{s_1} \dots \frac{1}{s_r}\right) U^\top = \sum_{j=1}^{r} \frac{1}{s_j} v_j u_j^\top$$

Rem: If M is invertible, its pseudoinverse is its inverse. That is, $A^+ = A^{-1}$

Bibliography

[1] Joseph Salmon, *Modéle linéaire avancé : introduction*, 2019, http://josephsalmon.eu/enseignement/Montpellier/HMMA307/Introduction.pdf.

[2] Francois Portier and Anne Sabourin, *Lecture notes on ordinary least squares*, 2019, https://perso.telecom-paristech.fr/sabourin/mdi720/main.pdf

[3] *Ordinary least squares*, 2020, https://en.wikipedia.org/wiki/Ordinary_least_squares.

[4] *Singular value decomposition*, 2020, https://en.wikipedia.org/wiki/Singular_value_decomposition.