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

Suppose the data consists of n samples $(y_i, x_i)_{i=1}^n$ with p features. The model can be written in matrix notation as :

$$y = X\beta + \epsilon$$

where

- X is an $n \times p$ matrix of regressors
- β is a $p \times 1$ vector of unknown parameters
- ϵ is a vector of normal random errors with mean 0

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

$$\hat{\beta}^{LS} \in argmin \ \underbrace{\frac{1}{2n}\|y - X\beta\|^2}_{f(\beta)}$$

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

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

Notation

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

$$X^T X = \begin{pmatrix} x_1^T \\ . \\ . \\ x_p^T \end{pmatrix} (x_1 ... x_p),$$

The Gram matrix is equivalent to :

$$[X^T X]_{j,j'} = [\langle x_j, x_{j'} \rangle]_{(j,j') \in [1,p]^2}$$

Remark

Most of the times, we scale features.

We have : $\bar{X_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ (1)

To center explanatory variables, we use the equation (1) to build the centered vector X_c

$X_c = X$ - $(\bar{X_1}1_n, ..., \bar{X_p}1_n)$ where $1_n = (1, ..., 1)$

Then we obtain $\bar{X_c} = O_n$

To reduce explanatory variables, we use :

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

Let X_r be the reduced vector, then :

$$X_{r_j} = \frac{X_j - \bar{X_j}1_n}{\hat{\sigma}_j}$$

First Order Optimality Conditions

We can verify the first order optimality condition because $\nabla f(\hat{\beta}^{LS}) = 0$
Note that f is a C^{∞} function, then differentiable

Remark

f is a convex function so a local minimum is a global one.

Conclusion

$\hat{\beta}^{LS}$ satisfy the following equations of orthogonality :

- $\frac{X^T X}{n}\hat{\beta}^{LS} - \frac{X^T y}{n} = 0$
- $\iff \mathrm{X}^T(\frac{X\hat{\beta}^{LS} - y}{n}) = 0$
- $\iff \mathrm{X}^T(y - X\hat{\beta}^{LS}) = 0$
- $\iff \langle X_j, y - X\beta \rangle = 0$ for j in 1:p

Attention

If $p < n$ so rank(X) $\leq n <$ p Then $\hat{\beta}^{LS}$ is not unique

Interpretation

- Each explanatory feature is orthogonal to the residuals $\Gamma = y - X\hat{\beta}^{LS}$ With $\hat{\beta}^{LS}$ a solution of the linear $p \times p$ system :

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

Remarks

- If $\hat{\Sigma}$ is invertible, the solution of the linear system is unique
- $\hat{\Sigma}$ is invertible $\Rightarrow$ $\hat{\Sigma}$ is positive definite
- If $\hat{\Sigma}$ invertible, so $rank(\hat{\Sigma}) = p$
- we assume that we have a full rank column e.g. :

$$rank(X) = dim(Vect(X_1, ..., X_p)) \leq n$$

Remark

- If $rank(X) = p$, so $\hat{\Sigma}$ is invertible and :

$$\hat{\beta}^{LS} = \hat{\Sigma}^{-1}\frac{X^T y}{n} = (\frac{X^T y}{n})^{-1}\frac{X^T y}{n}$$

so :

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

Notice

- In practice it is exceptional to invert $\hat{\Sigma}$ because one solves many linear systems

Data analysis

Goal

We want to build some ordinary least squares models of prediction with 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"

Data analysis

Solution

We convert features into ordinal features.

Prediction principle

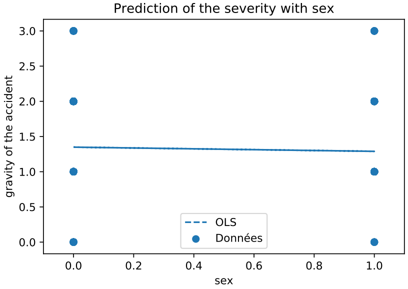
Calculate the coefficients β on a training sample and predict on a test sample the feature of interest. 0 is the value for male and 1 is the value for female.

Out[192]:

OLS Regression Results

Dep. Variable:	grave_quanti	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.002
Method:	Least Squares	**F-statistic:**	101.6
Date:	Sat, 03 Oct 2020	**Prob (F-statistic):**	7.12e-24
Time:	17:58:35	**Log-Likelihood:**	-63570.
No. Observations:	64515	**AIC:**	1.271e+05
Df Residuals:	64513	**BIC:**	1.272e+05
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

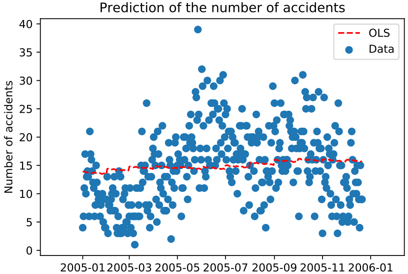
Predict the number of accidents with the date (day, month and year) that is an ordinal feature with periodic component. Results are also very bad.

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

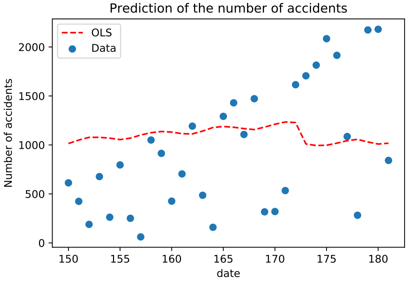


Table of Contents

Reminder

Let $\Sigma \in \mathbb{R}^{p \times p}$.
If $\Sigma^T = \Sigma$ then Σ is diagonalizable.

Theorem

For all matrix $M \in \mathbb{R}^{m_1 \times m_2}$ of 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 :

$$M = Udiag(s_1 ... s_r)U^T$$

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

Note that : $M = \sum_{j=1}^{r} s_j u_j v_j^T$ with : $U = [u_1, ..., u_r]$ et $V = [v_1 ... v_r]$

Definition

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

$$M^+ = Vdiag(\frac{1}{s_1}...\frac{1}{s_r})U^T = \sum_{j=1}^{r} \frac{1}{s_j} v_j u_j^T$$

Remark : 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.