**Advanced Statistics for Machine Learning**

**Exam 2024**

**Samd Guizani**

**(A23 – SPOC)**

# Exercise 1

## Question 1

This problem is a constrained least square optimization of a multiple linear regression. So, the aim is to find a vector of coefficients such that:

*N.B.: (1/2) coefficient in J(β) is introduced to avoid carrying multiplication by 2 when later taking the derivative.*

Where *R* is a q x (p+1) matrix and *r* is a vector of size q, q being the number of constraints and also the rank of *R*. So, *R* and *r* define a set of q linear constraints on the coefficients of the vector *β* (*R* holds the weights on the constrained coefficients and *r* holds the weighted sum to reach).

First, let us check the existence and uniqueness of a solution:

* The set is a **closed set** as it is defined by equality constraints.
* ***J(β)* is continuous** and ***α*-convex** as it is a squared norm.

We conclude that there exists at least a minimum of *J* on *K*.

Moreover:

* *K* is a **convex set**, indeed considering 2 points *β1* and *β2* in *K* i.e. and considering any point , we get:

So, *βθ* belongs to *K*.

* ***J(β)* is strictly convex** (being *α*-convex)

We conclude that there exists at most one minimum of *J* on *K*.

To find the minimum, since J and the linear constraints are differentiable and the constraints are regular, we use the Lagrange multipliers to find the minimum of *J*:

Where *λ* is a vector of size q.

The derivatives of *J* and *F* are:

Then:

To minimize *J(β)*, we must find the critical point of the Lagrangian, such that:

From the 1st equation, we deduce an expression of :

Substituting in the 2nd equation, we deduce an expression of *λ*:

Finally, injecting the expression of *λ* in the expression of , we obtain:

## Question 2

In the Ozone dataset, 2 variables are categorical: pluie (with 2 levels “Pluie” and “Sec”) and vent (with 4 levels "Est”, “Nord”, “Ouest” and “Sud”). To facilitate calculations and comparison of standard R software lm function with our computation of multiple linear regression (constrained or not), the categorical variables have been preprocessed using a one-hot key encoding. To illustrate the process, the 6 first observations are reported hereafter, before and after one-hot key encoding:

*Before pre-processing*:A screenshot of a computer

Description automatically generated

*After pre-processing*:A screenshot of a computer

Description automatically generated

### Model involving all the explanatory variables

Applying R software lm function, we obtain a model predicting maxO3 from all the explanatory variables. The summary is:

> LM = lm(formula = maxO3 ~ ., data = Datase .... [TRUNCATED]

> summary(LM)

Call:

lm(formula = maxO3 ~ ., data = Dataset\_ozone)

Residuals:

Min 1Q Median 3Q Max

-51.814 -8.695 -1.020 7.891 40.046

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16.26536 15.94398 1.020 0.3102

T9 0.03917 1.16496 0.034 0.9732

T12 1.97257 1.47570 1.337 0.1844

T15 0.45031 1.18707 0.379 0.7053

Ne9 -2.10975 0.95985 -2.198 0.0303 \*

Ne12 -0.60559 1.42634 -0.425 0.6721

Ne15 -0.01718 1.03589 -0.017 0.9868

Vx9 0.48261 0.98762 0.489 0.6262

Vx12 0.51379 1.24717 0.412 0.6813

Vx15 0.72662 0.95198 0.763 0.4471

maxO3v 0.34438 0.06699 5.141 1.42e-06 \*\*\*

ventNord 0.53956 6.69459 0.081 0.9359

ventOuest 5.53632 8.24792 0.671 0.5037

ventSud 5.42028 7.16180 0.757 0.4510

pluieSec 3.24713 3.48251 0.932 0.3534

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.51 on 97 degrees of freedom

Multiple R-squared: 0.7686, Adjusted R-squared: 0.7352

F-statistic: 23.01 on 14 and 97 DF, p-value: < 2.2e-16

We can calculate , the unconstrained least square solution as following:

> hbeta = solve(t(X) %\*% X) %\*% t(X) %\*% Y

> print(hbeta)

[,1]

(Intercept) 16.26535597

T9 0.03916979

T12 1.97257424

T15 0.45030800

Ne9 -2.10975486

Ne12 -0.60559218

Ne15 -0.01717804

Vx9 0.48260889

Vx12 0.51379495

Vx15 0.72662334

maxO3v 0.34437835

ventNord 0.53956395

ventOuest 5.53631722

ventSud 5.42028442

pluieSec 3.24713025

### Model with constraint

If we apply a constraint on the variables T9, T12 and T15 such that , we obtain the following solution:

> hbeta\_c =

+ hbeta +

+ solve(t(X) %\*% X) %\*%

+ t(R) %\*%

+ solve(R %\*% solve(t .... [TRUNCATED]

> print(hbeta\_c)

[,1]

(Intercept) 59.42206714

T9 -1.87215990

T12 1.17078960

T15 0.70137031

Ne9 -2.97625939

Ne12 -1.37664117

Ne15 0.08932548

Vx9 -0.16837184

Vx12 0.53454939

Vx15 0.77110593

maxO3v 0.46895526

ventNord -5.14140434

ventOuest 3.94210314

ventSud 6.50590711

pluieSec 5.28638433

### Models’ comparison

# Exercise 2

## Question 1

## Question 2