**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 (MLR). 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 :

being the unconstrained solution of the minimization problem.

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 = Dataset\_ozone)

> 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 (with their 95% confidence intervals), the unconstrained least square solution, as following:

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

> print(hbeta\_ci)

Est. Coef. Lower Bound Upper Bound

(Intercept) 16.26535597 -15.3790310 47.9097430

T9 0.03916979 -2.2729470 2.3512865

T12 1.97257424 -0.9562915 4.9014400

T15 0.45030800 -1.9057023 2.8063183

Ne9 -2.10975486 -4.0148008 -0.2047090

Ne12 -0.60559218 -3.4364784 2.2252941

Ne15 -0.01717804 -2.0731403 2.0387842

Vx9 0.48260889 -1.4775515 2.4427692

Vx12 0.51379495 -1.9614872 2.9890771

Vx15 0.72662334 -1.1627861 2.6160327

maxO3v 0.34437835 0.2114188 0.4773379

ventNord 0.53956395 -12.7473509 13.8264788

ventOuest 5.53631722 -10.8335269 21.9061613

ventSud 5.42028442 -8.7939071 19.6344759

pluieSec 3.24713025 -3.6646975 10.1589580

### 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(X) %\*% X) %\*% t(R)) %\*%

+ (r - R %\*% hbeta)

> print(hbeta\_c\_ci)

Est. Coef. Lower Bound Upper Bound

(Intercept) 59.42206714 25.6494312 93.1947031

T9 -1.87215990 -4.3397785 0.5954587

T12 1.17078960 -1.9550576 4.2966368

T15 0.70137031 -1.8130939 3.2158346

Ne9 -2.97625939 -5.0094295 -0.9430893

Ne12 -1.37664117 -4.3979192 1.6446369

Ne15 0.08932548 -2.1049109 2.2835618

Vx9 -0.16837184 -2.2603631 1.9236195

Vx12 0.53454939 -2.1072083 3.1763071

Vx15 0.77110593 -1.2453760 2.7875879

maxO3v 0.46895526 0.3270535 0.6108570

ventNord -5.14140434 -19.3219330 9.0391244

ventOuest 3.94210314 -13.5286978 21.4129040

ventSud 6.50590711 -8.6642624 21.6760766

pluieSec 5.28638433 -2.0902997 12.6630683

### Models’ comparison

The 2 models can be compared using statistical metrics as presented in the following table:

|  |  |  |
| --- | --- | --- |
|  | Unconstrained MLR | Constrained MLR |
| R² | 0.7686 | 0.7364 |
| Adjusted R² | 0.7352 | 0.6983 |
| Residuals Standard Error | 14.51 | 15.48 |

As expected, the constrained model shows lower R² values and higher Residuals Standard Error. Indeed, the unconstrained least square minimization yields a set of coefficients , that allows the predictions to be as close as possible to the observed datapoints. Hence, the Residuals Standard Error is the lowest and the R²/Adjusted R² are the highest. On the other hand, constraining the coefficients of the variables T9, T12 and T15, yields a set of coefficients for which the predictions are more distant from the observed dataset, compared to the unconstrained solution. Consequently, Residuals Standard Error grows and R²/Adjusted R² are lower.

On the following figure, the 2 models can be graphically compared by plotting the predicted response of each model against the actual response.



It is to be noticed that the constrained and unconstrained models are very close, despite the restriction imposed on the coefficients of the variables T9, T12 and T15. This may be explained by the correlations that exist among the explanatory variables which can be evaluated through the correlation matrix of the explanatory variables:

A screen shot of a computer

Description automatically generated

Noticeably, variables T9, T12 and T15 are correlated negatively with Ne9, Ne12 and Ne15 and positively with maxO3v. Hence, the constraint on the coefficients of T9, T12 and T15 get “distributed” on other variables. This finally leads to a different model (refer to section 2a and 2b where the coefficients of the 2 models are reported) which overall is not far from the unconstrained solution.

# Exercise 2

## Question 1

## Question 2