



■ FACULTÉ DE DROIT,
D'ÉCONOMIE
ET DE GESTION

Rapport Data Mining 2

Bède MASSALLA

Master 1 Economie Appliquée, Ingénierie Economique et Evaluation

Année académique 2019-2020

Professeur: **Christophe DANIEL**

I-Introduction

L'objet de ce travail est de comprendre le bien être des pays de l'OCDE, c'est pourquoi nous allons expliquer la variable SEV (satisfaction à l'égard de vie) en en fonction des variables qui peuvent intervenir dans l'explication de cette dernière, pour simplifier les choses notre variable principale SEV serait définie par Y et les variables explicatives qui interviennent dans l'explication de cette variable sont résumées dans la matrice X, le travail sera donc de modifier les variables initiales X et les transformer en composantes, qui peuvent être vue comme des variables construits comme des combinaisons linéaires des des variables initiales.

II-Description des variables

II.1-Dictionnaire des variables:

LSESB	Logement sans équipement sanitaire de base
CL	Coût de logement
IMT	Insécurité sur le marché du travail
TE	Taux d'emploi
TCLD	Taux de chômage à longue durée
NI	Niveau d'instruction
AS	Année de scolarité
PA	Pollution atmosphérique
QE	Qualité de l'eau
EV	Espérance de vie
SSQMS N	Se sentir en sécurité quand on marche seul la nuit
TH	Taux d'homicide

SEV	Satisfaction à l'égard de vie
-----	-------------------------------

II.2-Statistique des variables:

```
> summary(datamining)
```

LSESB		CL		IMT		TE		TCLD		NI	
Min.	: 0.000	Min.	:15.00	Min.	: 1.500	Min.	:43.00	Min.	: 0.030	Min.	:37.00
1st Qu.:	0.150	1st Qu.:	20.00	1st Qu.:	2.600	1st Qu.:	65.00	1st Qu.:	1.280	1st Qu.:	74.50
Median :	0.600	Median :	21.00	Median :	4.000	Median :	69.00	Median :	2.020	Median :	81.00
Mean :	3.415	Mean :	20.87	Mean :	5.462	Mean :	67.72	Mean :	3.168	Mean :	77.15
3rd Qu.:	4.250	3rd Qu.:	23.00	3rd Qu.:	5.350	3rd Qu.:	73.50	3rd Qu.:	3.855	3rd Qu.:	87.50
Max.	:37.000	Max.	:26.00	Max.	:26.500	Max.	:86.00	Max.	:16.950	Max.	:95.00

AS		PA		QE		EV		SSQMSN		TH	
Min.	:14.80	Min.	: 3.00	Min.	:54.00	Min.	:57.40	Min.	:36.10	Min.	: 0.200
1st Qu.:	16.45	1st Qu.:	9.50	1st Qu.:	74.50	1st Qu.:	78.35	1st Qu.:	61.25	1st Qu.:	0.600
Median :	17.30	Median :	14.00	Median :	84.00	Median :	81.10	Median :	70.20	Median :	1.000
Mean :	17.37	Mean :	13.41	Mean :	82.23	Mean :	79.55	Mean :	68.63	Mean :	2.951
3rd Qu.:	18.10	3rd Qu.:	17.00	3rd Qu.:	91.50	3rd Qu.:	82.20	3rd Qu.:	79.05	3rd Qu.:	1.700
Max.	:21.20	Max.	:28.00	Max.	:99.00	Max.	:83.90	Max.	:87.70	Max.	:27.600

SEV	
Min.	:4.800
1st Qu.:	5.900
Median :	6.600
Mean :	6.528
3rd Qu.:	7.250
Max.	:7.500

III-Corrélation des variables: le modèle que nous allons réaliser notre étude est un modèle linéaire, dans le cadre de ce dernier une forte corrélation entre les régresseurs conduit à un problème de quasi-colinéarité ou de colinéarité, cela rend donc l'estimation de paramètre complexe voire impossible.

```
> round(cor(datamining),3)
```

	LSESB	CL	IMT	TE	TCLD	NI	AS	PA	QE	EV	SSQMSN	TH	SEV
LSESB	1.000	-0.300	0.609	-0.515	0.409	-0.316	-0.370	0.328	-0.519	-0.896	-0.656	0.441	-0.524
CL	-0.300	1.000	-0.022	0.182	0.115	0.067	0.156	-0.233	0.201	0.219	0.182	-0.177	0.198
IMT	0.609	-0.022	1.000	-0.790	0.846	-0.594	-0.217	0.270	-0.495	-0.565	-0.443	0.155	-0.580
TE	-0.515	0.182	-0.790	1.000	-0.712	0.633	0.396	-0.529	0.709	0.520	0.610	-0.265	0.684
TCLD	0.409	0.115	0.846	-0.712	1.000	-0.394	-0.194	0.253	-0.348	-0.400	-0.296	0.041	-0.582
NI	-0.316	0.067	-0.594	0.633	-0.394	1.000	0.222	-0.091	0.426	0.333	0.473	-0.465	0.321
AS	-0.370	0.156	-0.217	0.396	-0.194	0.222	1.000	-0.465	0.569	0.418	0.529	-0.415	0.486
PA	0.328	-0.233	0.270	-0.529	0.253	-0.091	-0.465	1.000	-0.538	-0.331	-0.435	0.022	-0.559
QE	-0.519	0.201	-0.495	0.709	-0.348	0.426	0.569	-0.538	1.000	0.512	0.742	-0.484	0.606
EV	-0.896	0.219	-0.565	0.520	-0.400	0.333	0.418	-0.331	0.512	1.000	0.712	-0.566	0.520
SSQMSN	-0.656	0.182	-0.443	0.610	-0.296	0.473	0.529	-0.435	0.742	0.712	1.000	-0.684	0.569
TH	0.441	-0.177	0.155	-0.265	0.041	-0.465	-0.415	0.022	-0.484	-0.566	-0.684	1.000	-0.143
SEV	-0.524	0.198	-0.580	0.684	-0.582	0.321	0.486	-0.559	0.606	0.520	0.569	-0.143	1.000

les résultats de ce tableau nous montre clairement que les variables **EV** et **LSESB** sont très corrélés (0,896), ainsi que **TCLD** et **IMT** (0,845),

IV-Détection de la multicolinéarité des variables

la multicolinéarité engendre une instabilité des coefficients, il convient donc de la détecter

```
call:
lm(formula = SEV ~ LSESB + CL + IMT + TE + TCLD + NI + AS + PA +
    QE + EV + SSQMSN + TH, data = datamining)

Residuals:
    Min       1Q   Median       3Q      Max
-0.97819 -0.18778  0.09458  0.19861  0.88374

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.350103    5.028589  -0.467   0.6441
LSESB         0.005172    0.033920   0.152   0.8800
CL            0.046927    0.042530   1.103   0.2800
IMT           0.038773    0.050635   0.766   0.4507
TE           -0.002937    0.028648  -0.103   0.9191
TCLD         -0.102832    0.053410  -1.925   0.0652 .
NI            0.006111    0.009727   0.628   0.5353
AS            0.090134    0.081957   1.100   0.2815
PA           -0.008707    0.022078  -0.394   0.6965
QE            0.016504    0.015854   1.041   0.3075
EV            0.042213    0.050641   0.834   0.4121
SSQMSN       0.019958    0.013969   1.429   0.1650
TH            0.064037    0.031289   2.047   0.0509 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5233 on 26 degrees of freedom
Multiple R-squared:  0.6807,    Adjusted R-squared:  0.5333
F-statistic: 4.619 on 12 and 26 DF,  p-value: 0.0005347
```

le $R^2=0,5333$, 53% de la variation de $\ln Y$ est expliquée par les 12 régresseurs. cependant aucune variable n'apparaît significatif.

V-Facteur d'inflation de la variance (VIF): une recherche approfondie de la multicolinéarité nécessite l'examen de la valeur de R^2 obtenu en faisant la régression de chaque régression par

rapport aux autres. le vif permet donc une analyse de la relation entre les régresseurs et une valeur de $VIF > 10$ indique un sérieux problème de colinéarité.

```
> vif(RegModel.1)
      LSESB      CL      IMT      TE      TCLD      NI      AS
6.858675  1.429053  9.061823  7.480032  5.344102  3.274106  1.763195
      PA      QE      EV      SSQMSN      TH
2.291589  4.011616  7.624368  4.590707  3.956203
> mean(vif(RegModel.1))
[1] 4.807122
```

le tableau nous montre que certaines variables comme **LSESB**, **IMT**, **TE**, **TCLD** et **EV** sont supérieures à la valeur moyenne, alors que les variables **CL**, **NI**, **AS**, **PA**, **QE**, **SSQMSN** et **TH** sont inférieures à la valeur moyenne. cependant aucune valeur est > 10 , donc le problème de colinéarité n'est pas aussi sérieux que ce que l'on peut imaginer.

VI-Sélection du modèle

la méthode appliquée pour sélectionner les variables explicatives de notre modèle est celle de stepwise, car cette méthode apparaît plus efficace dans la mesure où elle permet à chaque étape un test de Student ou de Fichier pour éviter qu'une variable non significative se retrouve dans le modèle.


```

      Df Sum of Sq    RSS   AIC
- QE      1    0.26168  7.8304 -50.616
<none>                7.5687 -49.942
- CL      1    0.52370  8.0924 -49.333
- AS      1    0.61770  8.1864 -48.882
- SSQMSN  1    1.30843  8.8771 -45.723
- TH      1    1.56107  9.1298 -44.629
- TCLD    1    2.62787 10.1966 -40.319

Step:   AIC=-50.62
SEV ~ CL + TCLD + AS + SSQMSN + TH

      Df Sum of Sq    RSS   AIC
<none>                7.8304 -50.616
- CL      1    0.62799  8.4584 -49.608
- AS      1    0.97367  8.8041 -48.046
- TH      1    1.61033  9.4407 -45.323
- SSQMSN  1    2.69718 10.5276 -41.073
- TCLD    1    3.10159 10.9320 -39.603

Call:
lm(formula = SEV ~ CL + TCLD + AS + SSQMSN + TH, data = datamining)

Coefficients:
(Intercept)          CL          TCLD           AS          SSQMSN           TH
    0.86580     0.05591    -0.08513     0.13843     0.03208     0.05399

```

la méthode de stepwise nous permet de garder notre modèle les variables suivantes:

CL: coût de logement

TCLD: taux de chômage à longue durée

PA: pollution atmosphérique

SSQMSN: se sentir en sécurité quand on marche seul la nuit

TH: taux d'homicides

VII-Regression PCR

La régression sur composantes principales est conduite grâce à la fonction pcr. Ici nous avons au maximum 5 composantes principales. La procédure de validation croisé nous permet de garder que les 5 composantes car l'erreur de prévision est plus petite avec ces 5 composantes.

En gardant les 5 composantes nous pouvons expliquer 94,85% de la variabilité des variables exogènes, et 60,67% de la variabilité de la variable endogène.

Table PCR(a)

```

> datamining <- pcr(Y ~ X, ncomp = 5, validation="cv")
> summary(datamining)
Data:   X dimension: 39 12
        Y dimension: 39 1
Fit method: svdpc
Number of components considered: 5

VALIDATION: RMSEP
Cross-validated using 10 random segments.
      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV          0.7761  0.6524  0.6388  0.6084  0.5549  0.5445
adjcv       0.7761  0.6492  0.6350  0.5980  0.5494  0.5388

TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps
X       60.62   79.18   86.73   91.95   94.85
Y       37.71   45.54   57.16   59.25   60.67

```

le graphique nous à permis de tracer la statistique de l'erreur quadratique moyenne, et on voit bien comment les résultats du graphiques confirment que l'erreur de prévision est minimale avec 5 composantes.

Graphique PCR(a)

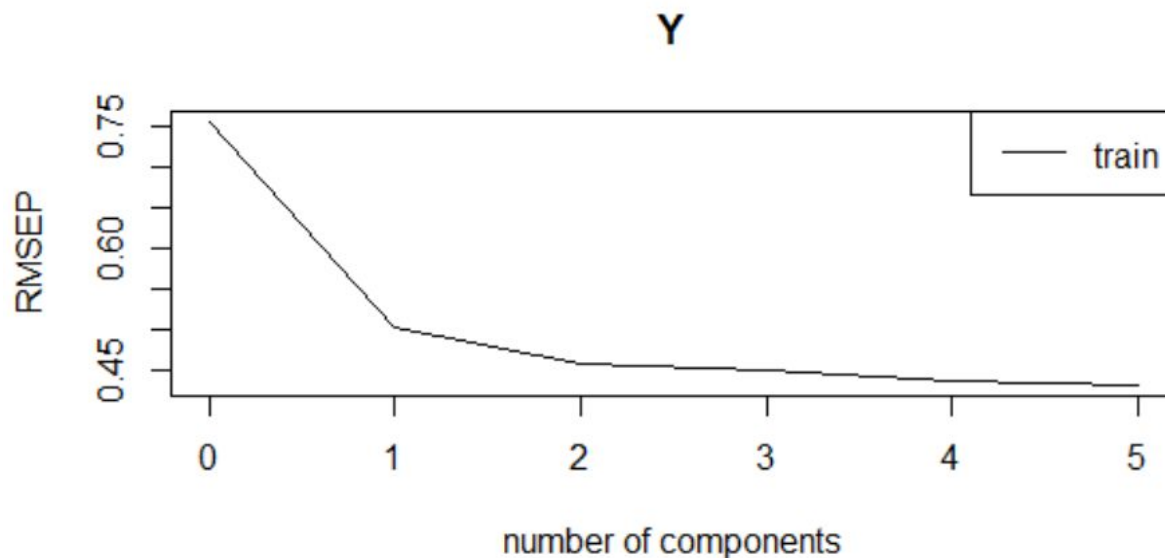


Table PCR(b)

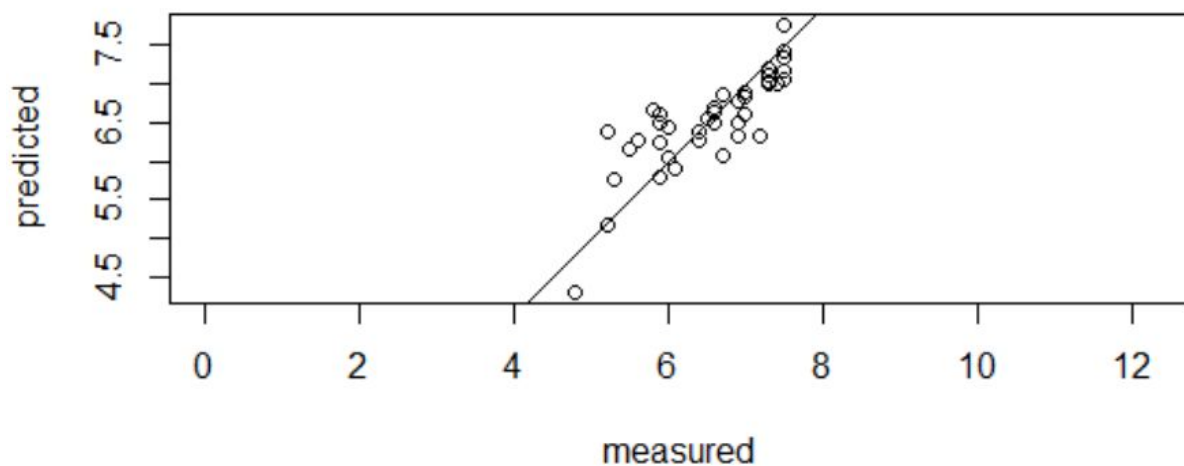
cette présente les composantes et leurs contribution dans l'explication dans la variances

```
> explvar(datamining.pcr)
  Comp 1   Comp 2   Comp 3   Comp 4   Comp 5
47.625022 11.333815 10.640060  4.769818  4.945267
> |
```

Graphique PCR(b)

ce graphique présente les prévisions

Y, 4 comps, train



VIII-Régression PLS

VIII.1-Sélection du nombre de composante:

VIII.1.1-Méthode de Jackknife:

En appliquant le principe de jackknife, on ne peut pas retenir 5 composantes car l'erreur de prévision n'est pas minimisée, les résultats du tableau semblent indiquer que en retenant 3 composante le nombre de composante optimal sera atteint.

```
> datamining.pls <- pls(Y ~ X, ncomp = 5, validation="LOO")
> summary(datamining.pls)
Data:      X dimension: 39 12
          Y dimension: 39 1
Fit method: kernelpls
Number of components considered: 5

VALIDATION: RMSEP
Cross-validated using 39 leave-one-out segments.
      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV          0.7761  0.6014  0.5691  0.5509  0.7120  1.07
adjCV       0.7761  0.6010  0.5682  0.5500  0.7085  1.06

TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps
X      59.55  75.73  86.22  90.20  93.41
Y      45.29  57.43  61.13  62.66  64.62
>
```

le principe de validation croisée nous permet ici de retenir donc 3 composantes, cela nous permet d'expliquer 86% de variance de X(des variables explicatives) et 61% de la variance de Y (la variable à expliquer).

```
> datamining.pls <- pls(Y ~ X, ncomp = 3, validation="CV")
> summary(datamining.pls)
Data:      X dimension: 39 12
          Y dimension: 39 1
Fit method: kernelpls
Number of components considered: 3

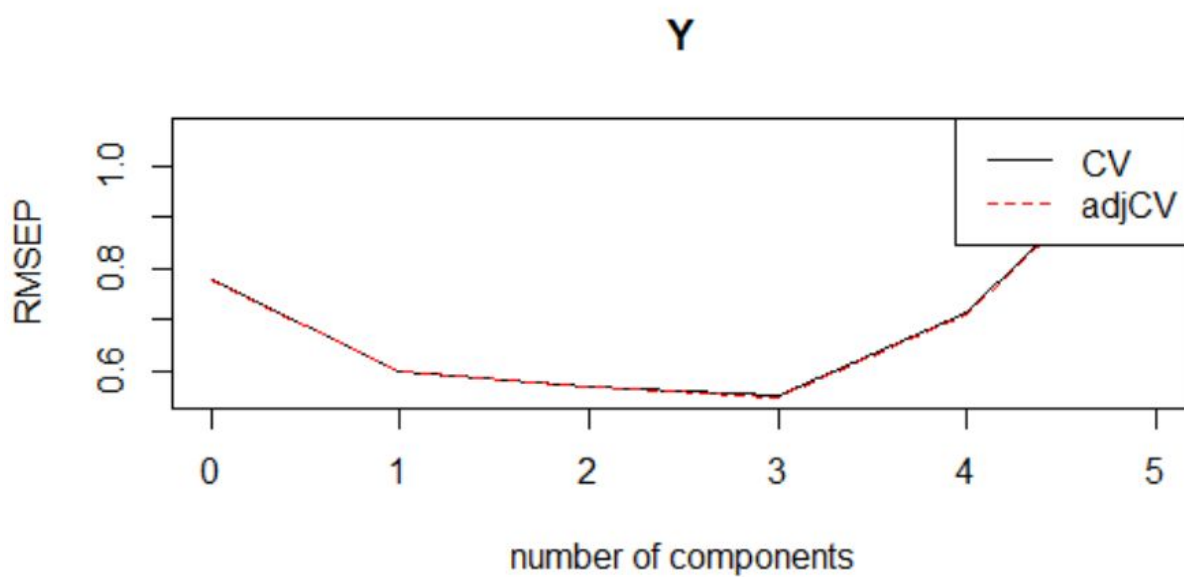
VALIDATION: RMSEP
Cross-validated using 10 random segments.
      (Intercept)  1 comps  2 comps  3 comps
CV          0.7761  0.6034  0.5739  0.5567
adjCV       0.7761  0.6017  0.5699  0.5525

TRAINING: % variance explained
      1 comps  2 comps  3 comps
X      59.55  75.73  86.22
Y      45.29  57.43  61.13
```

VIII.1.2-Méthode Graphique:

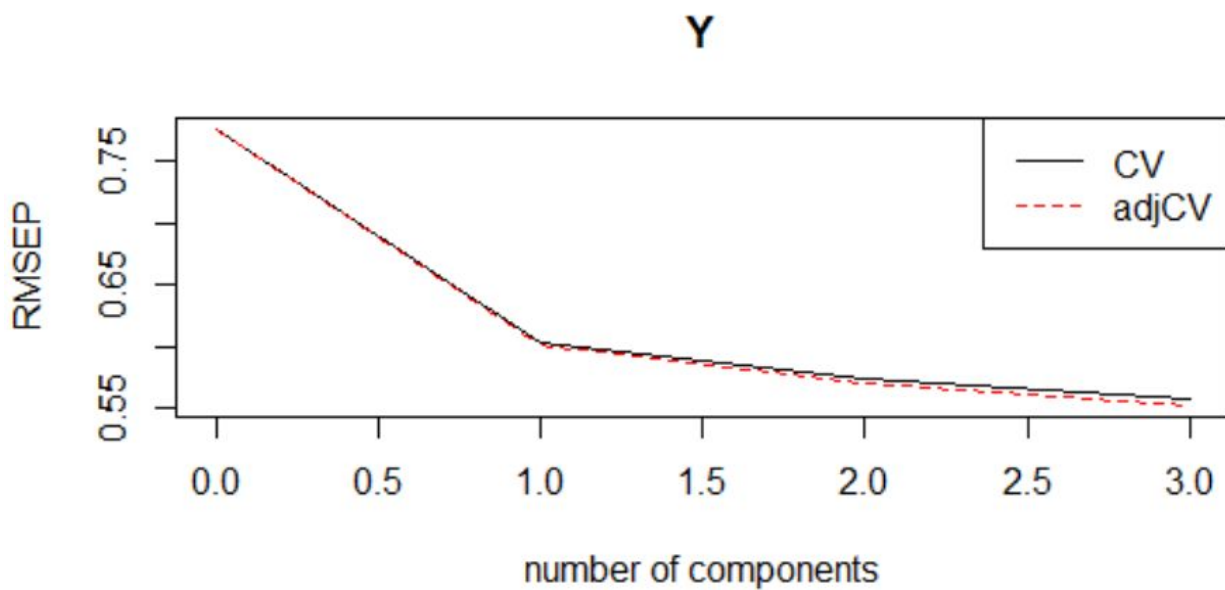
l'analyse graphique nous indique retenir 5 composantes ne permet pas de minimiser l'erreur de prévision

graphique PLS(a)



le graphique PLS(b) nous confirme bien que le nombre de composante optimal est égal à 3.

Graphique PLS(b)



VIII.2-Régression de Y sur les composantes (PLS1)

ici on considère le cas de la PLS à une seule variable à expliquer, c'est à dire la modélisation d'une variable dépendante, communément appelé PLS1.

les résultats de la première estimation nous indique que, la variable CL n'est pas significatif voir table PLS(a)

Table PLS(a)

```
Call:
lm(formula = SEV ~ CL + TCLD + AS + SSQMSN + TH, data = datamining)

Residuals:
    Min       1Q   Median       3Q      Max
-1.05130 -0.25826  0.07433  0.33815  0.72419

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.865800    1.317721   0.657  0.515710
CL           0.055907    0.034366   1.627  0.113287
TCLD        -0.085130    0.023546  -3.615  0.000988 ***
AS           0.138434    0.068339   2.026  0.050946 .
SSQMSN       0.032080    0.009515   3.371  0.001920 **
TH           0.053992    0.020725   2.605  0.013669 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4871 on 33 degrees of freedom
Multiple R-squared:  0.6488,    Adjusted R-squared:  0.5956
F-statistic: 12.2 on 5 and 33 DF,  p-value: 9.749e-07
```

on retirant la variable CL et en estimant de nouveau, on obtient les résultats dans la table PLS(b)

Table PLS(b)

```
Call:
lm(formula = SEV ~ TCLD + AS + SSQMSN + TH, data = datamining,
    method = "simpls", validation = "CV")

Residuals:
    Min       1Q   Median       3Q      Max
-1.1681 -0.2233  0.1425  0.3095  0.7331

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.747109   1.229954   1.420  0.16458
TCLD         -0.078483   0.023744  -3.305  0.00224 **
AS           0.147108   0.069761   2.109  0.04241 *
SSQMSN       0.033776   0.009684   3.488  0.00137 **
TH           0.053137   0.021214   2.505  0.01721 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4988 on 34 degrees of freedom
Multiple R-squared:  0.6207,    Adjusted R-squared:  0.5761
F-statistic: 13.91 on 4 and 34 DF,  p-value: 8.048e-07
```

tous les coefficients associés aux variables explicatives sont significatives car leurs probabilités critiques sont inférieures à 5%.

le R^2 (ajusté) explique 58% de la variation de $\ln Y^*$ et est expliquée par 4 régresseurs.

VIII.2.1-Variance expliquée en fonction du nombre de composantes:

les 3 composantes expliquent 69,60% de la variance de X (régresseurs), et 64,85% de la variance de Y

Table PLS(c)

```
> datamining.pls <- mvr(Y ~ X, ncomp = 3,
+                       method = "oscorespls", scale = TRUE)
> summary(datamining.pls)
Data:   X dimension: 39 12
        Y dimension: 39 1
Fit method: oscorespls
Number of components considered: 3
TRAINING: % variance explained
      1 comps  2 comps  3 comps
X      47.63   58.96   69.60
Y      55.80   63.58   64.85
```

VIII.2.2-Poids des variables dans l'explication des composantes:

la variable **TCLD** participe à l'explication des composantes 1 et 2, la variable **AS** participe à l'explication des trois composantes, la variable **SSQMSN** explique les trois composantes et enfin la variable **TH** explique la composante 2 et 3,

Tableau PLS(d)

```
> loading.weights(datamining.pls)
```

Loadings:	Comp 1	Comp 2	Comp 3
LSESB	-0.298	0.177	
CL	0.113		0.281
IMT	-0.330		0.495
TE	0.389		-0.261
TCLD	-0.331	-0.297	
NI	0.183	-0.377	-0.156
AS	0.276	0.125	0.339
PA	-0.318	-0.372	0.295
QE	0.345		
EV	0.296	-0.207	0.173
SSQMSN	0.324	-0.122	0.572
TH		0.719	0.131

	Comp 1	Comp 2	Comp 3
ss loadings	1.000	1.000	1.000
Proportion Var	0.083	0.083	0.083
Cumulative Var	0.083	0.167	0.250

VIII.2.3-Analyse des coefficients de chaque variables:

quatre coefficients des variables expliquent le bien-être (SEV) de façon pertinente:

- une réduction du taux de chômage à longue durée d'une unité augmente toute chose égale par ailleurs une amélioration du bien-être de 7,8%
- une augmentation d'une année d'étude engendre une amélioration du bien-être de 15%

- une augmentation d'une unité de sécurité des individus lorsqu'ils sans seul la nuit apporte une contribution au bien être de 3,4%
- Le taux d'Homicide lorsqu'il augmente de 1% s'accompagne d'une augmentation du bien être de 5,3%

Table PLS(e)

```
> coef(RegModel.1)
(Intercept)      TCLD          AS      SSQMSN      TH
1.74710911 -0.07848343  0.14710769  0.03377566  0.05313659
```

Conclusion:

ce travail nous a permis de réaliser les régressions PCR et PLS1, mais aussi de pouvoir comprendre les facteurs facteurs qui peuvent contribuer de façon pertinente à l'explication du bien être dans la zone de l'OCDE.

Bibliographie:


Sites

https://odr.inra.fr/intranet/carto/cartowiki/index.php/Regression_lin%C3%A9aire_avec_R

<https://www.xlstat.com/fr/solutions/fonctionnalites/regression-sur-les-composantes-principales>

<https://egallic.fr/l3-eco-gestion-regression-lineaire-avec-r-probleme-de-multicolinearite/>

Livres



Lise Bellanger et **Richard Tomasse (2014)** “EXploration de données et méthodes statistique, Data analysis et Data mining avec le logiciel R”, *ellipses*.

Pierre-André Cornillon et **Eric Matzner-Lober** (2011) “Régression avec R”, *Springer*.