Mémoire de projet de fin d'étude pour l'obtention du diplôme de licence sciences et techniques en Mathématiques Appliquées

La régression dans l'analyse des données massives

NAKHLA Khaoula
TINA Djara Olivier

Présenté devant le jury composé de:
Pr.Abddelghani Bentahar: Encadrant
Pr.Rachid Eljid: Examinateur
Pr.Mohamed Hanini: Examinateur

14 juillet 2021

Plan

كلية العلوم و التقنيات سطات
FACULTÉ DES SCIENCES ET TECHNIQUES SETTAT

Plan

كلية العلوم و التقنيات سطات
FACULTÉ DES SCIENCES ET TECHNIQUES SETTAT

Plan

كلية العلوم و التقنيات سطات
FACULTÉ DES SCIENCES ET TECHNIQUES SETTAT

Plan

كلية العلوم و التقنيات سطات
FACULTÉ DES SCIENCES ET TECHNIQUES SETTAT

Plan

كلية العلوم و التقنيات سطات
FACULTÉ DES SCIENCES ET TECHNIQUES SETTAT

Plan

كلية العلوم و التقنيات سطات
FACULTÉ DES SCIENCES ET TECHNIQUES SETTAT

Plan

Introduction

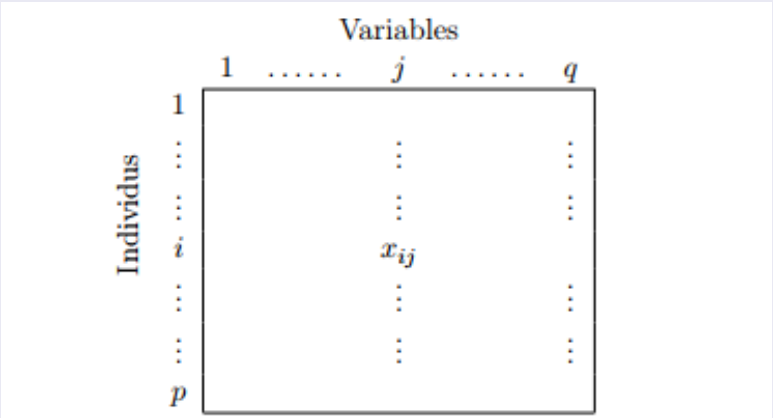
" Un danger prévu est à moitié évité "

Plan

Analyse en Composante Principale

Introduction

On dispose d'un tableau de données relatives à q variables quantitatives $x_{.1}; \ldots; x_{.q}$ portant sur p individus $x_{1.}; \ldots; x_{p.}$. Le tableau des données X a la forme suivante :



Dans toute la suite on interprète ce tableau comme une matrice X

Analyse en Composante Principale

A.C.P

ACP sur une matrice variance covariance

Dans ce cas on suppose que la métrique $M = I$ et les individus ont même poids $\omega_i = 1/p$, p étant le nombre d'individus. Avec ces conditions. la matrice VM est la matrice des variances covariances

$$(Cov(x_{.k}, x_{.l}))_{1 \leq k,l \leq q}$$

ACP normée

On réalise l'ACP sur la matrice des corrélations des variables $x_{.1}; ...; x_{.q}$ avec $M = I$ et $\omega_i = 1/p$, c-a-d

$$Z = VM = (R(x_{.k}, x_{.l}))_{1 \leq k,l \leq q}$$

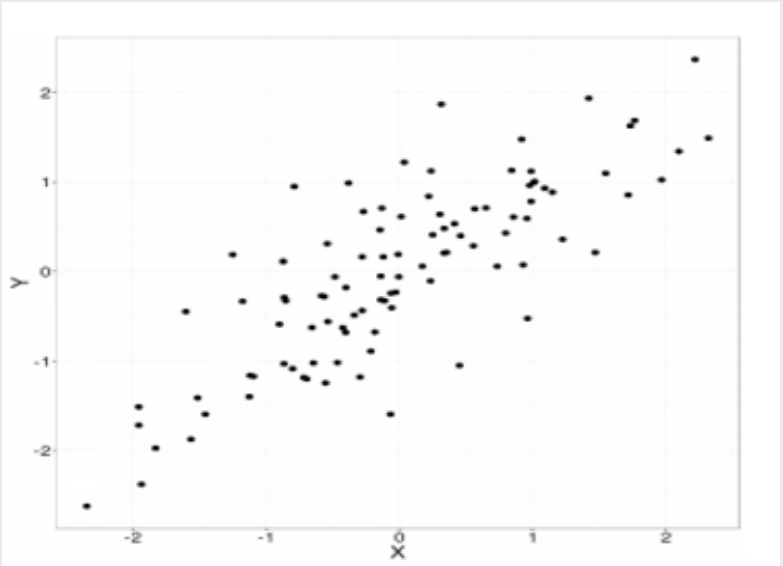
où

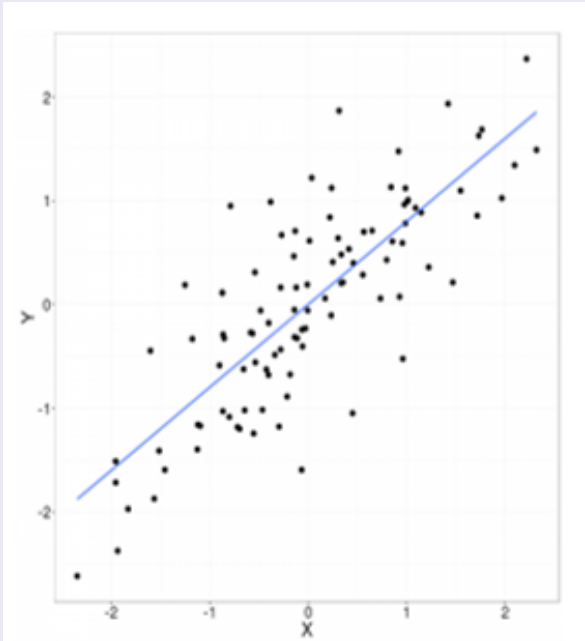
$$R(x_{.k}, x_{.l}) = \frac{1}{n} \sum_{i=1}^{p} \left(\frac{x_{ik} - \bar{x}_{.k}}{\sigma_k}\right)\left(\frac{x_{il} - \bar{x}_{.l}}{\sigma_l}\right)$$

Plan

La régression linéaire simple

Introduction





La régression linéaire simple

Modélisation

On cherche à modéliser la relation entre deux variables quantitatives continues.
Un modèle de régression linéaire simple est de la forme suivante :

$$y = \beta_1 + \beta_2 x + \varepsilon$$

où :
- y est la variable à expliquer (à valeurs dans $\mathbb{R}$) ;
- x est la variable explicative (à valeurs dans $\mathbb{R}$) ;
- ε est le terme d'erreur aléatoire du modèle ;
- β_1 et β_2 sont deux paramètres à estimer.

La régression linéaire simple

Estimateurs des Moindres Carrés Ordinaires

On appelle estimateurs des Moindres Carrés Ordinaires (en abrégé MCO) $\hat{\beta}_1$ et $\hat{\beta}_2$ les valeurs minimisant la quantité :

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

Autrement dit, la droite des moindres carrés minimise la somme des carrés des distances verticales des points (x_i, y_i) du nuage à la droite ajustée

$$y = \hat{\beta}_1 + \hat{\beta}_2 x$$

.

La régression linéaire simple

Estimateurs sans biais

Théorème
$\hat{\beta_1}$ et $\hat{\beta_2}$ sont des estimateurs sans biais de $\beta 1$ et $\beta 2$.

Gauss-Markov

Théorème
Parmi les estimateurs sans biais linéaires en y, les estimateurs $\hat{\beta}_j$ sont de variances minimales.

Estimateur non biaisé de σ^2

Théorème
La statistique $\hat{\sigma^2} = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{(n-2)}$ est un estimateur sans biais de σ^2.

La régression linéaire simple

Estimateurs sans biais

Théorème
$\hat{\beta}_1$ et $\hat{\beta}_2$ sont des estimateurs sans biais de $\beta 1$ et $\beta 2$.

Gauss-Markov

Théorème
Parmi les estimateurs sans biais linéaires en y, les estimateurs $\hat{\beta}_j$ sont de variances minimales.

Estimateur non biaisé de σ^2

Théorème
La statistique $\hat{\sigma^2} = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{(n-2)}$ est un estimateur sans biais de σ^2.

La régression linéaire simple

Estimateurs sans biais

Théorème
$\hat{\beta}_1$ et $\hat{\beta}_2$ sont des estimateurs sans biais de $\beta 1$ et $\beta 2$.

Gauss-Markov

Théorème
Parmi les estimateurs sans biais linéaires en y, les estimateurs $\hat{\beta}_j$ sont de variances minimales.

Estimateur non biaisé de σ^2

Théorème
La statistique $\hat{\sigma^2} = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{(n-2)}$ est un estimateur sans biais de σ^2.

Plan

La régression linéaire multiple

Modélisation

Étant donné un échantillon $(Y_i, X_{i1}, ..., X_{ip})$ $i \in \{1, \ldots, n\}$, on cherche à expliquer, avec le plus de précision possible, les valeurs prises par Y_i, dite variable endogène, à partir d'une série de variables explicatives $X_{i1}, ..., X_{ip}$. Le modèle théorique, formulé en termes de variables aléatoires, prend la forme :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

où ε_i est l'erreur du modèle qui exprime, ou résume, l'information manquante dans l'explication linéaire des valeurs de Y_i à partir des $X_{i1}, ..., X_{ip}$ (problème de spécifications, variables non prises en compte, etc.). Les coefficients $\beta_0, \beta_1, ..., \beta_p$ sont les paramètres à estimer.

Modélisation

Ecriture matricielle :

$$Y = X\beta + \varepsilon,$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} ; \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} ; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{ij} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{in} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

La régression linéaire multiple

Estimation par la méthodes des moindres carrés

Théorème
L'estimateur $\hat{\beta}$ des moindres carrés a pour expression :

$$\hat{\beta} = (X'X)^{-1}(X'Y).$$

Estimation par la méthodes des moindres carrés

Théorème L'estimateur β des moindres carrés est de variance minimale parmi les estimateurs linéaires sans biais de β

La régression linéaire multiple

Estimation par la méthodes des moindres carrés

Théorème
L'estimateur $\hat{\beta}$ des moindres carrés a pour expression :

$$\hat{\beta} = (X'X)^{-1}(X'Y).$$

Estimation par la méthodes des moindres carrés

Théorème L'estimateur β des moindres carrés est de variance minimale parmi les estimateurs linéaires sans biais de β

La régression linéaire multiple

Sommes des carrés

Théorème La somme des carrées totales des écarts à la moyenne

$$SCT = (y - \bar{y})'(y - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

se décompose en une somme de deux termes :

$$SCE = (\hat{y} - \bar{y})'(\hat{y} - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SCR = \varepsilon'\varepsilon = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2.$$

La régression linéaire multiple

Le coefficient de Détermination R^2

R^2 est un indicateur spécifique permet de traduire la variance expliquée par le modèle, il s'agit du coefficient de détermination. Sa formule est la suivante :

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

Enfin, si le R^2 est certes un indicateur pertinent, il présente un défaut parfois ennuyeux, il a tendance à mécaniquement augmenter à mesure que l'on ajoute des variables dans le modèle. De ce fait, il est inopérant si l'on veut comparer des modèles comportant un nombre différent de variables. Il est conseillé dans ce cas d'utiliser **le coefficient de détermination ajusté** qui est corrigé des degrés de libertés. Le R^2 ajusté est toujours inférieur au R^2

La régression linéaire multiple

Test de signification globale du modèle

L'objectif du test global de Fisher est d'étudier la liaison globale entre Y et les variables explicatives $X_j, (j = 1, ..., p)$(significative ou pas) . On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0 \\ H_1 : \exists! \beta_j \neq 0, \qquad (j = 1, ..., p) \end{cases}$$

Il est posssible d'utiliser la statistique de Fisher F pour tester cette hypothèse :

$$F = \frac{SCE/p}{\frac{SCR}{n-p-1}}$$

où F suit une loi de Fisher avec p et $(n - p - 1)$ dégeé de liberté.
Si

$$F \geq f^{p-\alpha}_{1,n-p-1},$$

(avec $f^{p-\alpha}_{1,n-p-1}$, est le quantile d'ordre $1 - \alpha$ d'une loi de Fisher à $(p, n - p - 1)$ degrés de liberté) alors on rejette H_0

La régression linéaire multiple

Test de Student de signification du paramètre du modèle

L'objectif du test de Student est d'évaluer l'influence de la variable X_j sur $Y (j = 1, ..., p)$.
On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0, \qquad (j = 1, ..., p) \end{cases}$$

où β_j est le paramètre associé à la variable explicative X_j.
L'hypothèse H_0 de nullité d'un paramètre du modèle peut être testée au moyen de statistique de Student :

$$T(\hat{\beta}_j) = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$$

comparer à $t^{1-\frac{\alpha}{2}}_{n-p-1}$, où $t^{1-\frac{\alpha}{2}}_{n-p-1}$ est le quantile d'ordre $(1 - \frac{\alpha}{2})$ d'une loi de Student à $(n - p - 1)$ degrés de liberté.

- Si $|T(\hat{\beta}_j)| \geq t^{1-\frac{\alpha}{2}}_{n-p-1}$, on rejette H_0
- Si $|T(\hat{\beta}_j)| < t^{1-\frac{\alpha}{2}}_{n-p-1}$, on ne peut pas rejeter H_0

Plan

Application

Introduction

Application

Présentation des données

On souhaite expliquer la nocivité des cigarettes (teneur en monoxyde de carbone – CO) (y) à partir de leur composition : TAR (goudron) (x_1), NICOTINE (x_2) et WEIGHT (poids) (x_3), soit p = 3 variables explicatives.
Nous disposons de n = 24 observations. Les observations sont sont présentés sous forme d'un tableau des données (voir figure).

	TAR	NICOTINE	POIDS	CO
Alpine	14.1	0.86	985.3	13.6
Benson&Hedges	16.0	1.06	1093.8	16.6
CamelLights	8.0	0.67	928.0	10.2
Carlton	4.1	0.40	946.2	5.4
Chesterfield	15.0	1.04	888.5	15.0
GoldenLights	8.8	0.76	1026.7	9.0
Kent	12.4	0.95	922.5	12.3
Kool	16.6	1.12	937.2	16.3
L&M	14.9	1.02	885.8	15.4
LarkLight	13.7	1.01	964.3	13.0
Marlboro	15.1	0.90	931.6	14.4
Merit	7.8	0.57	970.5	10.0

	TAR	NICOTINE	POIDS	CO
MultiFilter	11.4	0.78	1124.0	10.2
NewportLight	9.0	0.74	851.7	9.5
Now	1.0	0.13	785.1	1.5
OldGold	17.0	1.26	918.6	18.5
PallMallLight	12.8	1.08	1039.5	12.6
Raleigh	15.8	0.96	957.3	17.5
SalemUltra	4.5	0.42	910.6	4.9
Tareyton	14.5	1.01	1007.0	15.9
TrueLight	7.3	0.61	980.6	8.5
ViceroyRichLight	8.6	0.69	969.3	10.6
VirginiaSlims	15.2	1.02	949.6	13.9
WinstonLights	12.0	0.82	1118.4	14.9

Application

Analyse univariée

On effectue l'analyse univariée qui consiste à déterminer la moyenne et l'ecart type des variables

TAR	NICOTINE	POIDS
11.4833333	0.8283333	962.1708333

TAR	NICOTINE	POIDS
4.4151583	0.2655866	79.4512072

Application

Analyse univariée

On calculer ensuite la matrice des corrélations entre les variables à l'aide de la commande :

	TAR	NICOTINE	POIDS
TAR	1.00	0.96	0.28
NICOTINE	0.96	1.00	0.29
POIDS	0.28	0.29	1.00

Remarque : la matrice des corrélations est à diagonale unité ce qui explique la parfaite corrélation d'une variable avec elle même.

Application

A.C.P

Pour réalisé l'ACP : on aura besoin des packages à installer tel que factomineR, factoextra, ggplot2,... Puis on ecrit la commande :
Cette analyse réalisée sur nos 3 variables nous permet de voir plusieurs résultats

	name	description
1	"$eig"	"valeurs propres"
2	"$var"	"resultat pour les variables"
3	"varcoord"	"coordonnées des variables"
4	"varcor"	"correlations entres variables – dimensions"
5	"varcos2"	"qualité des variables"
6	"varcontrib"	"contributions des variables"
7	"$ind"	"resultat pour les individus"
8	"indcoord"	"coordonnées des individus"
9	"indcos2"	"cos2 for the individuals"
10	"indcontrib"	"contributions des individus"
11	"$call"	"résumés statistiques"
12	"$call$centre"	"la moyenne des variables"
13	"$call$ecart.type"	"l'erreur standard des variables"
14	"$call$row.w"	"le poids des individus"
15	"$call$col.w"	"le poids des variables"

Application

A.C.P

On peut par exemple décider d'afficher les valeurs propres de notre matrice :

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.10651349	70.217116	70.21712
comp 2	0.85338559	28.446186	98.66330
comp 3	0.04010092	1.336697	100.00000

$comp1, ..., comp3$ sont les composantes principales de l'ACP, les valeurs de la 1^{ere} colonne sont les valeurs propres associées aux vecteurs propres $comp1, ..., comp3$. Chaque valeur propre λ_s étant la variance de la sème composante principale.
La 2^{eme} colonne représente le pourcentage de la variance c-a-d pour chaque composante s le pourcentage de variance est

$$(\lambda_s / \sum \lambda_i) \times 100$$

Application

A.C.P

Qualités et Contributions des variables

Les tableaux suivants résument les qualités et contributions des variables de notre exemple.

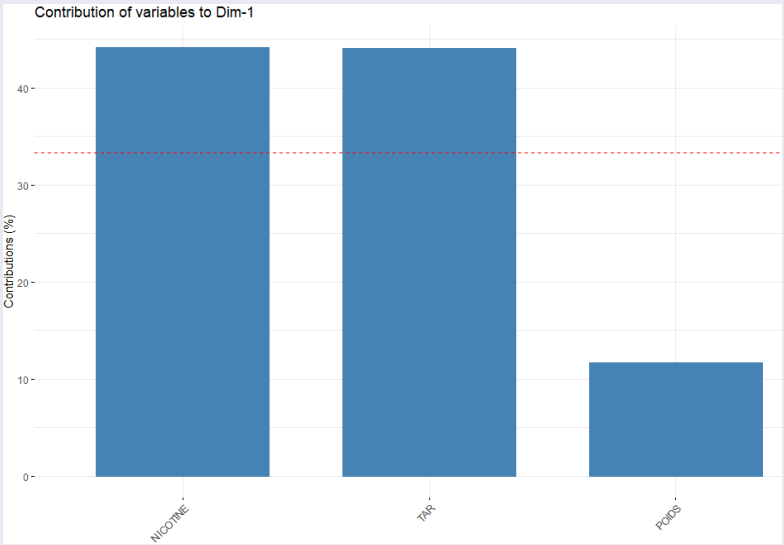
	Dim.1	Dim.2
TAR	0.9294242	0.05054306
NICOTINE	0.9306239	0.04930815
POIDS	0.2464654	0.75353438

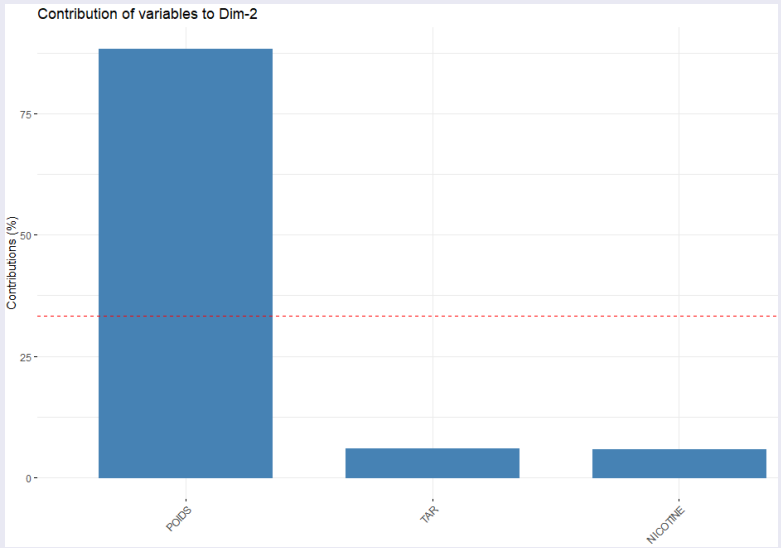
	Dim.1	Dim.2
TAR	44.12144	5.922653
NICOTINE	44.17840	5.777945
POIDS	11.70016	88.299403

Remarque Plus la contribution est grande, plus la représentation est meilleure.

Application

A.C.P



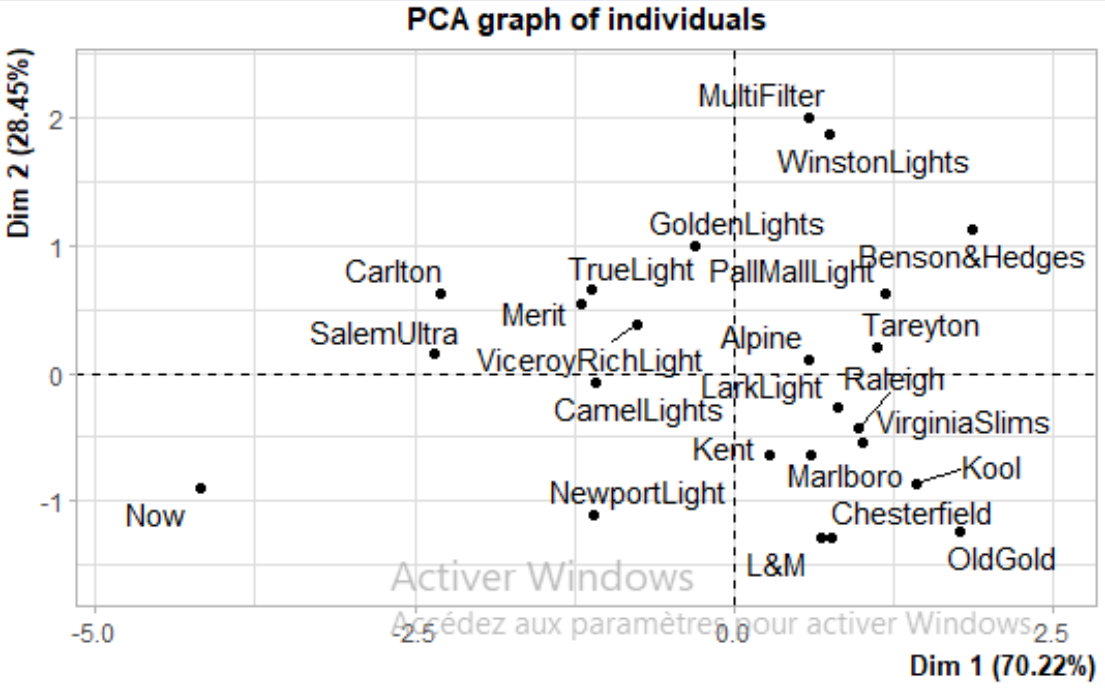


Contributions des variables dans la formation des axes

Application

A.C.P

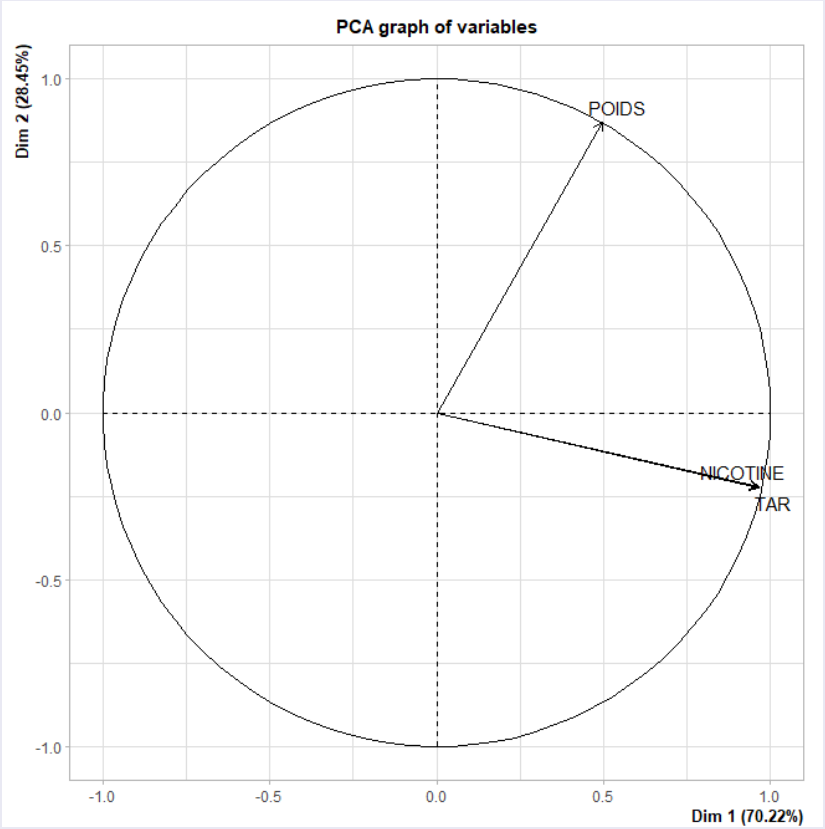
Le graphe des individus :
C'est la représentation des individus sur les 2 dimensions les plus importantes



Application

A.C.P

Le cercle de corrélation des variables



Application

A.C.P Synthèse

Ce qu'on retient de cette étude est que les variables TAR(goudron) et NICOTINE sont fortement corrélées donc on peut les répresenter par la variable qui contribue le plus à la formation de l'axe 1 avec une meilleure qualité de representation : c'est à dire la variable NICOTINE.

Application

Modélisation

Soit le modèle de la régression linéaire multiple(RLM) avec 2 variables explicatives :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

où β_0, β_1 et, β_2 sont des paramètres inconnus(les coefficients du modèle).
-x_{i1} :la i-ème valeur de la variable X1(représente la NICOTINE).
-x_{i2} :la i-ème valeur de la variable X2(correpond aux poids).
-y_i :la i-ème valeur de la variable Y(représente la teneur de monoxyde de carbone).
Le dernier terme ε_i représente la déviation entre ce que le modèle prédit et la réalité (l'erreur du modèle).
On peut récrire ce modèle sous la forme matricielle suivante :

$$Y = X\beta + \varepsilon$$

Application

Estimation

On estime les paramètres et on obtient :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

Il s'agit de calculer le vecteurs des estimateurs$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ définit par l'égalité suivante :

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

Les estimateurs $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ sont donnés directement, sous R, on obtient :

```
Call:
lm(formula = y ~ x2 + x3)

Coefficients:
(Intercept)           x2           x3
  -2.516865    14.641152     0.002557
```

Application

Estimation

Alors on calcule,

$$X^tX = \begin{pmatrix} 18.0896 & 19266.81 & 19.88 \\ 19266.8110 & 22363732.47 & 23092.10 \\ 19.8800 & 23092.10 & 24.00 \end{pmatrix}.$$

Donc

$$(X^tX)^{-1} = \begin{pmatrix} 0.6713505296 & -6.420687e-04 & 0.061677756 \\ -0.0006420687 & 7.501716e-06 & -0.006686085 \\ 0.0616777558 & -6.686085e-03 & 6.423732861 \end{pmatrix}.$$

Application

Estimation

Et

$$X^t Y = \begin{pmatrix} 264.076 \\ 281145.080 \\ 289.700 \end{pmatrix}$$

Ainsi,on obtient :

$$\hat{\beta} = \begin{pmatrix} -2.516865419 \\ 14.641151792 \\ 0.002556661 \end{pmatrix}$$

Donc l'equation de l'hyperplan des moindres carrés est donc donnée par :

$$\hat{y_i} = -2.516865419 + 14.641151792 x_{i2} + 0.002556661 x_{i3}$$

Application

Interprétation des résultats

Le signe du coefficient nous indique le sens de la relation alors on va étudier l'impact où l'influence de chaque variables explicatives X_i sur la variable expliquée y(CO).

Pour la variable X_2 représente - NICOTINE : On remarque que le coefficient de régression estimée associé à la variable NICOTINE est positif cela signifie que l'augmentation de la NICOTINE implique une augmentation de la teneur de monoxyde de carbone
Pour la variable x_3 qui représente le -POIDS, le coéfficient est positif avec une valeur faible, ce qui montre que le POIDS n'a pas beaucoup d'influence sur la teneur en monoxyde de carbone

Application

Evaluation

Résidus Pour tout $i \in 1.....24$, on appelle i-ème résidu la réalisation e_i de :
$e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i$
Sous R :
On obtient,le résultat suivant(voir Figure) :

1	2	3	4	5	6	7
1.00639675	0.80076867	0.53471227	-0.35870798	0.01847422	-2.23533383	-1.45074859
8	9	10	11	12	13	14
0.02267268	0.71820024	-1.73608613	1.35804338	1.69016936	-1.57691999	-0.99509512
15	16	17	18	19	20	21
0.10628110	0.22046533	-3.35322767	3.51386808	-1.06051388	1.05474444	-0.42129899
22	23	24				
0.53629914	-0.94491473	2.55175124				

Application

Estimation de la matrice de variance-covariance de $\hat{\beta}$

La matrice de variance-covariance, notée par $varcov(\hat{\beta})$ ou par $var(\hat{\beta})$,des coefficients est importante car elle renseigne sur la variable de chaque coefficient estimé et permet de faire les tests des hypothèses, notamment de voir si chaque coefficient est significativement différent de zéro.Elle est définie par :

$$varcov(\hat{\beta}) = \hat{\sigma^2}(X^t)(X)^{-1}$$

où $\hat{\sigma^2}$:est l'estimateur sans biais de la variance des résidus donné par :

$$\hat{\sigma^2} = \frac{\sum \hat{\varepsilon_i^2}}{n-m-1} = \frac{SCR}{n-m-1}$$

(où m est le nombre des variables explicatives).

Application

Estimation de la matrice de variance-covariance de $\hat{\beta}$

Donc, on obtient le résultat suivant :

$$\hat{\sigma^2} = \frac{SCR}{n-m-1} = \frac{54.63644}{21} = 2.601735$$

où n=24 et m=2
On peut alors calculer la matrice de variance-covariance sous R comme suit :

(où varcov : est la matrice de variance covariance), ou on peut calculer la matrice de variance-covariance,sous R, directement par la commande :
on obtient le résultat suivant : Les écarts-types $\hat{\sigma}(\hat{\beta}_j)$ des estimateurs $\hat{\beta}_j (j = 0, 1, 2)$ sont alors

	(Intercept)	x2	x3
(Intercept)	16.71285162	0.160469186	-1.739542e-02
x2	0.16046919	1.746676275	-1.670493e-03
x3	-0.01739542	-0.001670493	1.951748e-05

Application

Evaluation

donnés par les racines carrées des éléments diagonaux de cette matrice de variance-covariance :
Sous R, on a :

$$\hat{\sigma}(\hat{\beta_0}) = 4.088135470$$

$$\hat{\sigma}(\hat{\beta_1}) = 1.321618809$$

$$\hat{\sigma}(\hat{\beta_2}) = 0.004417859$$

Application

Evaluation globale de la régression

Coefficient de détermination Un des usages de la RLM consiste à prédire la valeur d'un Y pour un ensemble des valeurs $x_1, x_2, ..., x_p$ donné.La mesure de l'ajustement du modèle aux données est donc importante.
La proportion de variabilité expliqué par les 2 régresseurs est calculé par le coefficient de détermination R^2 qui eqt donné par la relation suivante :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Donc ce coefficient est un indicateur spécifique permet de traduire la variance expliquée par le modèle.
On obtient $R^2 = 86.79\%$ ce qui montre un bon ajustement.

Application

Test globale de Fisher

Ce test permet de répondre à la question suivante :
"Est ce que la liaison globale entre Y et les X_j est-elle significative ? "
On veut tester l'hypothèse nulle :

$$H_0 : \beta_1 = \beta_2 = \beta_3,$$

Contre l'hypothèse alternative :
$H_i : \exists\, j \in 1, 2, 3$ tel que $\beta_i \neq 0$
C'est-à-dire que Y dépend d'au moins une variable X_j :
On calcule la statistique de test :

$$F_{obs} = \frac{n-m-1}{m} \frac{R^2}{1-R^2} = 69.01$$

Application

Interprétation du test

Comme le $F_{obs} = 69.01$ est de loin supérieur à la valeur critique(valeur théorique 3.467). On conclut que le test est significatif, on rejette l'hypothèse nulle H_0 au seuil de significativité $\alpha = 5\%$. Ce resultat montre qu'au moins une de nos deux variables contribue à expliquer la nocivité du tabac. Il n'explique ne nous explique pas si toutes les variables y contribuent car il est global.

Application

Test paramétrique

```
Call:
lm(formula = y ~ x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3532 -1.0114  0.0645  0.8522  3.5139

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.516865   4.088135  -0.616    0.545
x2          14.641152   1.321619  11.078 3.13e-10 ***
x3           0.002557   0.004418   0.579    0.569
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.613 on 21 degrees of freedom
Multiple R-squared:  0.8679,    Adjusted R-squared:  0.8554
F-statistic: 69.01 on 2 and 21 DF,  p-value: 5.859e-10
```

Application

Interprétation du test

Comme la valeur critique est de 0.545, on décide de rejeter l'hypothèse H_0 pour $j = 1, 2$. Ainsi on comprend que la variable POIDS n'est pas très influente sur le modèle en présence de la variable NICOTINE. Par contre la variable NICOTINE à une grande influence dans le modèle même en présence de la variable POIDS.

Plan

Conclusion

En gros on retient que la nocivité du tabac vient essentiellement de la NICOTINE car elle est source d'addiction et dépend légèrement de la masse du tabac(POIDS). Plus une personne est addicte au tabac,plus elle en consomme et plus le CO issu de la combustion intervient dans sa circulation sanguine ce qui peut conduire à des maladies et à la mort.

MERCI DE VOTRE ATTENTION