

CONSERVATOIRE NATIONNALE DES ARTS ET METIERS

ANNEE 2011-2012

DEPARTEMENT D'INGENIERIE MATHEMATIQUE

LA PROCEDURE PLS DE SAS

« UNE APPLICATION A UN PORTEFEUILLE D'ASSURANCE AUTOMOBILE »

Professeur Gilbert SAPORTA

Présenté par :

Luc NGOMA MBAMBY

Table des matières	pages
Introduction.....	4
Partie I : La Régression PLS.....	5
1.1- Présentation-Définition.....	5
1.2 - Historique.....	6
1.3 - La Méthode PLS Linéaire.....	6
a- Le Modèle de base.....	6
b- La Méthode d'optimisation.....	7
c- Les algorithmes.....	7
Partie II-CADRE PRATIQUE.....	8
2.1-Présentation des données et notation des variables.....	8
2.2- Syntaxe de la procédure PLS de SAS.....	9
2.3 - Création de la Table et Stockage des données.....	9
2.4 - Applications et résultats des estimations.....	10
a. Estimation PLS sur les variables tarifaires d'un portefeuille automobile.....	10
b. Méthodes d'extraction des facteurs.....	20
b1. Méthode de régression par réduction des rangs.....	21
b.2 Méthode de la régression sur composantes principales.....	21
b.3 Méthode des moindres carrés partiels.....	22
C. Choix optimal du nombre de facteurs PLS.....	24
d. Construction des facteurs PLS.....	25
e. L'analyse du score des facteurs optimaux.....	28
f. Analyse descriptive.....	35
Partie III- CONCLUSION	40
ANNEXE n°1 : Code SAS de Transformation et de codage des variables.....	41

ANNEXE n°2 : Sortie des différents programmes.....	44
Bibliographie:.....	46

Introduction

Ce projet a pour but de mettre en évidence, la procédure PLS de SAS à travers une application d'un portefeuille d'assureur Automobile.

Il est constitué de trois parties. La première partie est suivie de trois points au quels nous avons abordé la base théorique de la régression PLS linéaire. Nous avons ensuite présenté un aperçu historique ainsi que la méthode de régression à travers le modèle de base, les contraintes d'optimisation et les algorithmes PLS.

La seconde partie, composée de quatre points est consacrée à la procédure PLS de SAS dans un cadre pratique. Nous avons dans un premier temps, présenté les données, défini les variables et spécifié la syntaxe de la procédure PLS de SAS. Dans un second temps, nous avons créé la Table, présenté les différentes applications de la régression PLS de SAS, les résultats des estimations et la construction des graphiques.

La troisième et la dernière partie est consacrée à la conclusion.

Partie I : La Régression PLS

1.1- Présentation-Définition

Dans bien des domaines scientifiques ou économiques, il est à présent possible de prédire, contrôler ou expliquer le comportement d'une ou des plusieurs variables dites expliquées par des indicateurs plus faciles à mesurer et donc à contrôler appelées variables explicatives.

La relation entre les différentes variables est souvent déterminée par les différentes techniques statistiques couramment utilisées entre autres techniques de régression linéaire simple dont l'estimation fait appel à plusieurs méthodes d'analyse des données de plus en plus sophistiquées comme la régression multiple¹, l'analyse en composante principale¹, l'analyse discriminante¹ et la corrélation canonique¹.

Seulement, la difficulté apparaît quand le nombre des variables expliquées est souvent plus important que le nombre d'observation. D'où l'apparition de la méthode dite de régression PLS, connue à ce jour pour résoudre ces genres de problèmes.

La régression PLS est une technique qui généralise et combine les caractéristiques de l'analyse en composante principale et la régression multiple. Elle est particulièrement utile quand on a besoin de prédire un ensemble de variables dépendantes à partir d'un ensemble de variables explicatives appelées prédicteurs qui peuvent être très fortement corrélées entre elles.

Quand les prédicteurs sont peu nombreux, non significatifs, colinéaires et ont une relation connue avec les réponses, alors la régression linéaire multiple est la meilleure méthode pour utiliser les données. Cependant, si l'une de ces trois conditions n'est pas vérifiée, la régression multiple peut être inefficace et inappropriée. Dans cette optique,

¹ Nous les citons à titre indicatif sans les définir. Nous renvoyons le lecteur qui souhaite approfondir la connaissance de ces mots à l'ouvrage du Professeur Gilbert SAPORTA : « Probabilités, analyse des données et statistique ; édition technip, août 2011 broché, 17 x 24 cm, 656 p. »

la régression PLS est définie comme une méthode pour construire des modèles de prédictions quand les facteurs sont nombreux et très colinéaires.

Enfin, la Régression PLS est une technique quantitative de décomposition spectrale étroitement liée à la régression en composante principale (PCR). Cependant, dans PLS, la décomposition est faite d'une manière légèrement différente. Au lieu de décomposer d'abord la matrice spectrale X en un ensemble de vecteurs propres et de scores, et de les régresser contre les Y dans une étape séparées, PLS utilise l'information de Y en même temps que le processus de décomposition. Ce qui implique que les variables expliquants le mieux Y seront plus fortement pondéré. De plus, les vecteurs propres et les scores calculés en utilisant PLS seront différents de ceux de PCR.

L'idée principale de PLS est de donner le plus d'information possible sur Y dans les premiers vecteurs construits.

1.2 - Historique

La régression PLS encore appelée Partial Least Squares Régression tire son origine des sciences sociales et devient très populaires en chimie grâce au fils d'HERMANN, SVANTE.

Elle est née de l'association de l'algorithme NIPALS (non lineare iterative Partial Least Squares) développées par H.WOLD en 1966 pour l'analyse en composante principale et de l'approche PLS pour l'estimation des modèles d'équations structurelles sur les variables latentes.

1.3 - La Méthode PLS Linéaire

a- Le Modèle de base

Le but principal de la régression PLS est comme nous venons de l'indiquer est de construire un modèle linéaire de type :

$$Y = XB_{pc} + E_{n,c}$$

Où B_{pc} : coefficient de régression ; Y : variable réponse et X : variables prédictives.

b- La Méthode- L'optimisation

PLS est une méthode de régression linéaire dont toutes les variables sont mesurées sur les mêmes individus. La méthode procède par différentes étapes successives permettant le calcul des composantes principales².

Dans la procédure PLS, on s'intéresse à maximiser la fonction $\varphi(w, q)$ sous la contrainte de Lagrangien L tel que :

$$R^p \times R^c \rightarrow R$$

$$(w, q) \rightarrow \varphi(w, q) = \text{cov}(t, u) = t^T D u = u^T D t.$$

Le Lagrangien associé à cette fonction est :

$$R^p \times R^c \times R \times R \rightarrow R$$

$$(w, q, \lambda, \mu) \rightarrow L(w, q, \lambda, \mu) = \varphi(w, q) + \frac{\lambda}{2}(1 - \|w\|^2) + \frac{\mu}{2}(1 - \|q\|^2).$$

Ainsi, Maximiser la fonction φ revient à maximiser L sans contrainte.

c- Les algorithmes

Nous avons entre autres, l'algorithme PLS1, PLS2 et SIMPLS.

L'algorithme PLS1 NIPALS possède plusieurs propriétés mathématiques qui permettent de le généraliser au cas de plusieurs Y c'est-à-dire régression multiple (PLS2).

Les calculs se font en différentes étapes selon la prise en compte des données complètes ou manquantes¹.

L'algorithme SIMPLS proposé par Jong (1993) permet de relier un groupe de réponse Y à un groupe de prédictors X . On utilise cet algorithme dans la régression PLS lorsque Y se limite à une seule variable et donne des résultats très proches dans le cas général.

C'est cet algorithme que nous allons utiliser dans la deuxième partie.

² Article : la Régression PLS par S.VANCOLEN, p.11-16, JUIN 2004, Université de Neuchâtel (SUISSE)

II-CADRE PRATIQUE

2.1-Présentation des données et notation des variables

Les données utilisées dans le cadre de ce projet concernent les variables de tarification d'un portefeuille d'assureur automobile. Elles sont stockées dans un fichier Excel et comporte vingt variables du type quantitatifs et qualitatifs. Il s'agit de :

CHARGETOT : Charge totale du sinistre ;

NBP : indicateur sur la personnalité du conducteur ;

CSP : catégorie socioprofessionnelle (1 à 100 : code CSP) ;

K8000 : option kilométrage limité (oui, non) ;

STATUT : le statut de l'assuré ;

USAGE : usage du véhicule (trajet privé, professionnel);

ENE : énergie (essence, diesel, GO)

SEXE : Sexe du conducteur (Homme, Femme, Personne morale) ;

AGECOND : âge du conducteur

PERMIS : catégorie du permis (noviste, moyen, ancien)

CRM : coefficient de réduction bonus-malus,

GARAGE : Type de garage (individuel clos, collectif clos, autres) ;

SEGM : segment du véhicule (0 non renseigné, segment plus bas, segment plus haut) ;

ALI : type d'alimentation du véhicule (INJ, CAR, INS, IDS) ;

VIT : vitesse maximale du véhicule ;

CAR : type du véhicule (berline, camionnette, coupé) ;

2.2 Syntaxe de la procédure PLS de SAS

La PROC PLS est une procédure statistique du module STAT du logiciel SAS version 9.2 permettant de construire des modèles de régression et de les estimer par la méthode des moindres carrés partiels. La spécification de cette procédure suit la syntaxe présentée ci-après :

PROC PLS < options > ;

CLASS variables nominales ;

MODEL réponse = facteurs < options > ;

OUTPUT OUT = tables_SAS < options > ;

2.3 Création de la Table et Stockage des données

- **Création de la bibliothèque permanente**

Il faut créer une bibliothèque permanente car les bibliothèques temporaires (work, Sasuser, Sashelp, Maps) de SAS se vident lorsqu'on quitte SAS.

- **Stockage des données**

Les données sont sous format Excel, il faudra les convertir sous le format SAS. Elles seront stockées dans la Bibliothèque créé nommée LUC.

- **Création de la Table de données**

Deux manières peuvent être utilisées pour créer la table de données :

-Soit par l'utilisation de l'étape DATA ;

-Soit par l'importation des données. Nous utilisons cette dernière possibilité pour la création du tableau de données et du codage des variables. La démarche SAS ayant servi à cette création est présentée en annexe n° 1(page41).

2.4 Applications et résultats des estimations

Ce premier programme est une mise en application de la régression PLS sur un portefeuille d'assurance automobile, dans le cadre de l'analyse de la sinistralité. Il traduit une analyse du coefficient de détermination et de la corrélation. Les résultats obtenus sont présentés ci-après :

a. Estimation PLS sur les variables tarifaires d'un portefeuille automobile

Programme n°1 : Estimation de la régression PLS

```
proc pls data=luc.basedonnee2;  
  
    model chargetot AGECOND CRM = NBAP CSP K8000 SEXE STATUT USAGE  
    ENE PERMIS GARAGE ALI VIT SEGM CAR;  
  
run;
```

Les résultats obtenus sont ci-après :

Figure n°1: résultats de la régression PLS

The SAS System

18:41 Wednesday, August 29, 2012 2

The PLS Procedure

Percent Variation Accounted for

by Partial Least Squares Factors

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	16.9006	16.9006	37.7543	37.7543
2	14.6929	31.5935	15.1094	52.8637
3	14.0355	45.6290	6.7835	59.6472
4	10.4076	56.0366	4.9451	64.5923
5	9.2874	65.3239	3.5734	68.1656
6	7.8545	73.1784	2.6246	70.7902
7	4.3013	77.4797	3.5848	74.3750

8	4.7580	82.2377	2.1826	76.5577
9	4.7432	86.9810	1.2198	77.7775
10	4.5508	91.5318	0.9693	78.7468
11	3.4286	94.9603	0.5657	79.3125
12	2.2032	97.1635	0.3107	79.6232
13	2.8365	100.0000	0.1091	79.7324

The SAS System

18:41 Wednesday, August 29, 2012 1

The PLS Procedure

Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Partial Least Squares
PLS Algorithm	NIPALS
Number of Response Variables	3
Number of Predictor Parameters	13
Missing Value Handling	Exclude
Number of Factors	13
Number of Observations Read	23
Number of Observations Used	23

FIGURE n°2: graphique des corrélations

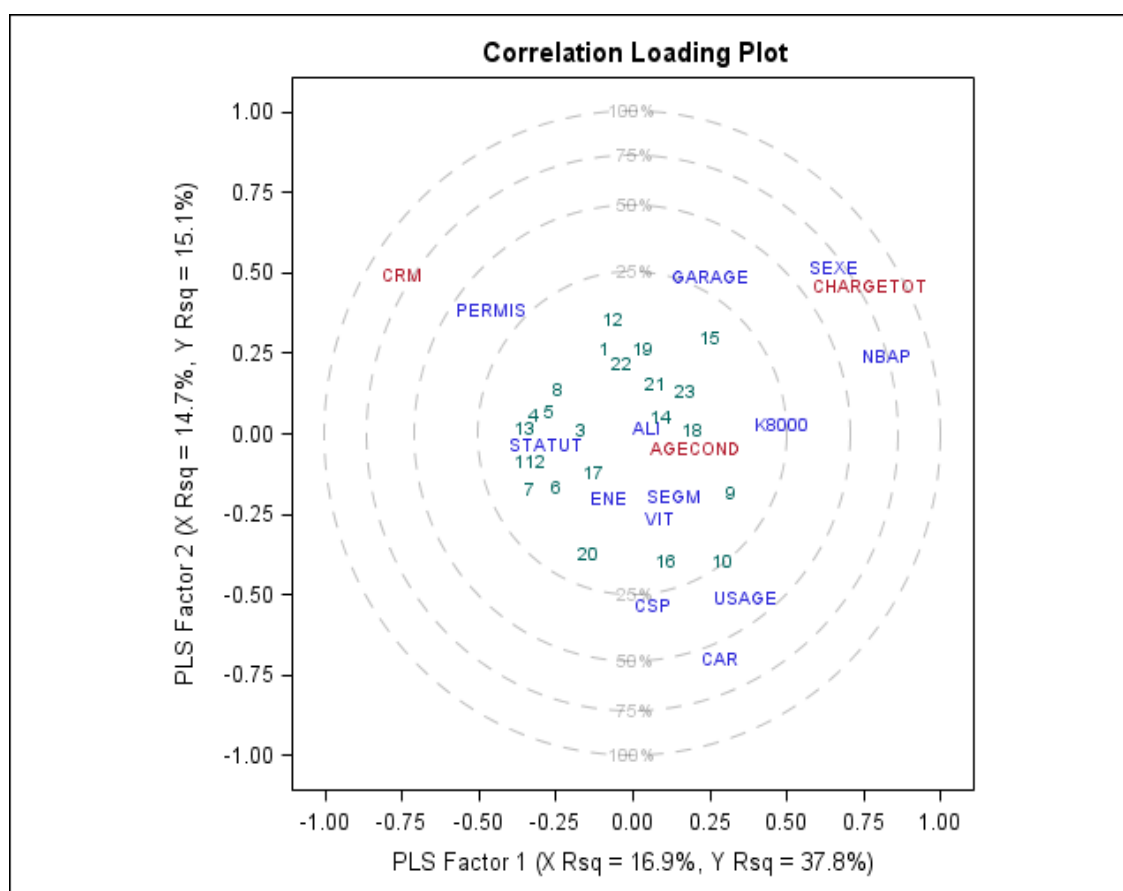
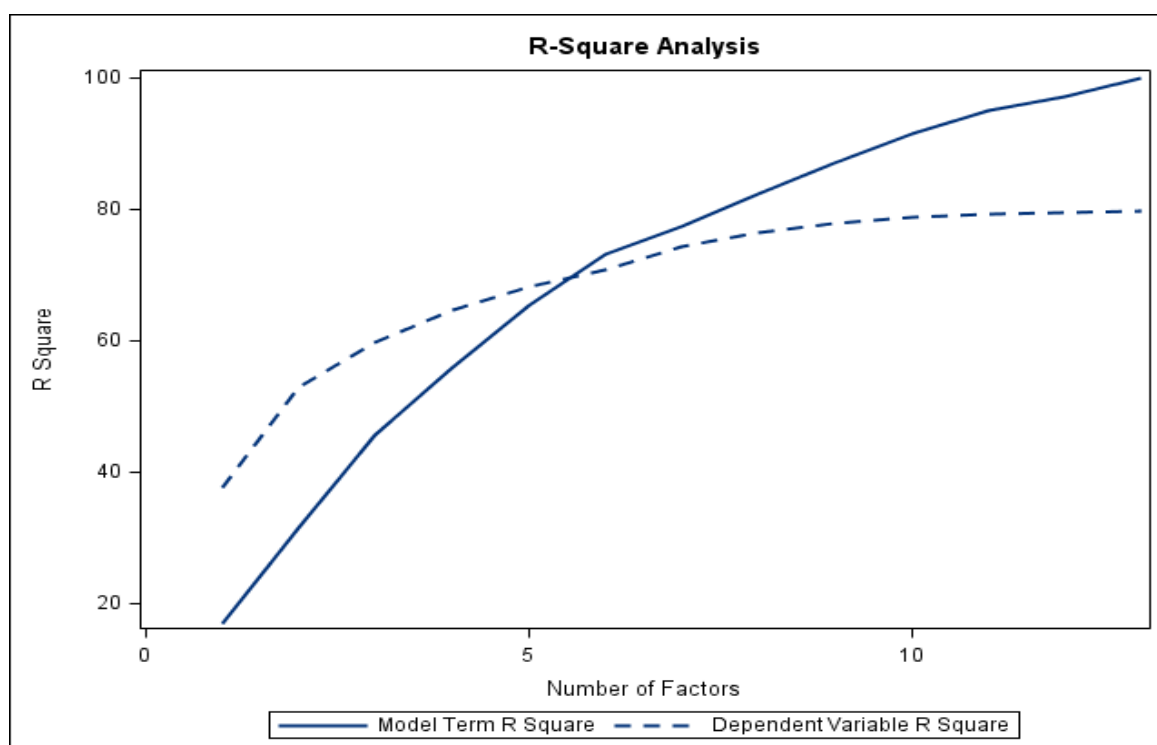


FIGURE n° 3: graphique du coefficient de détermination



Programme n°2: Estimation de la régression PLS et construction des splines

```
Proc pls data=luc.basedonnee2 cv=split;

    model chargetot AGECOND CRM = NBAP CSP K8000 SEXE STATUT USAGE
    ENE PERMIS GARAGE ALI VIT SEGM CAR;

run;
```

Les résultants de ce Code SAS sont résumés dans la Figure ci-après:

Figure n°4 : Resultants des estimations

The SAS System

18:41 Wednesday, August 29, 2012 2

The PLS Procedure				
Percent Variation Accounted for				
by Partial Least Squares Factors				
Number of				
Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	16.9006	16.9006	37.7543	37.7543
2	14.6929	31.5935	15.1094	52.8637
3	14.0355	45.6290	6.7835	59.6472
4	10.4076	56.0366	4.9451	64.5923
5	9.2874	65.3239	3.5734	68.1656
6	7.8545	73.1784	2.6246	70.7902
7	4.3013	77.4797	3.5848	74.3750
8	4.7580	82.2377	2.1826	76.5577
9	4.7432	86.9810	1.2198	77.7775
10	4.5508	91.5318	0.9693	78.7468
11	3.4286	94.9603	0.5657	79.3125
12	2.2032	97.1635	0.3107	79.6232
13	2.8365	100.0000	0.1091	79.7324

The SAS System

18:41 Wednesday, August 29, 2012 3

```
The PLS Procedure

Data Set                LUC.BASEDONNEE2

Factor Extraction Method  Partial Least Squares

PLS Algorithm            NIPALS
```

Number of Response Variables	3
Number of Predictor Parameters	13
Missing Value Handling	Exclude
Number of Factors	13

Number of Observations Read	23
-----------------------------	----

Number of Observations Used	23
-----------------------------	----

The SAS System	18:41 Wednesday, August 29, 2012	4
----------------	----------------------------------	---

The PLS Procedure

Percent Variation Accounted for
by Partial Least Squares Factors

Number of

Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total

1	16.9006	16.9006	37.7543	37.7543
2	14.6929	31.5935	15.1094	52.8637
3	14.0355	45.6290	6.7835	59.6472
4	10.4076	56.0366	4.9451	64.5923
5	9.2874	65.3239	3.5734	68.1656
6	7.8545	73.1784	2.6246	70.7902
7	4.3013	77.4797	3.5848	74.3750
8	4.7580	82.2377	2.1826	76.5577
9	4.7432	86.9810	1.2198	77.7775
10	4.5508	91.5318	0.9693	78.7468
11	3.4286	94.9603	0.5657	79.3125
12	2.2032	97.1635	0.3107	79.6232
13	2.8365	100.0000	0.1091	79.7324

The PLS Procedure

Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Partial Least Squares
PLS Algorithm	NIPALS
Number of Response Variables	3
Number of Predictor Parameters	13
Missing Value Handling	Exclude
Maximum Number of Factors	13
Validation Method	7-fold Split-sample Validation
Number of Observations Read	23
Number of Observations Used	23

Split-sample Validation for the Number of Extracted Factors

Number of Extracted Factors	Root Mean PRESS
0	1.09233
1	0.980405
2	0.962026
3	1.093625
4	1.203187
5	1.274464
6	1.356129
7	1.372021
8	1.385233
9	1.411795
10	1.405046
11	1.453273
12	1.591623
13	1.683531

Minimum root mean PRESS	0.9620
-------------------------	--------

Minimizing number of factors	2
------------------------------	---

The PLS Procedure

Percent Variation Accounted for
by Partial Least Squares Factors

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	16.9006	16.9006	37.7543	37.7543
2	14.6929	31.5935	15.1094	52.8637

Figure n°5 : graphique de l'analyse du coefficient de détermination

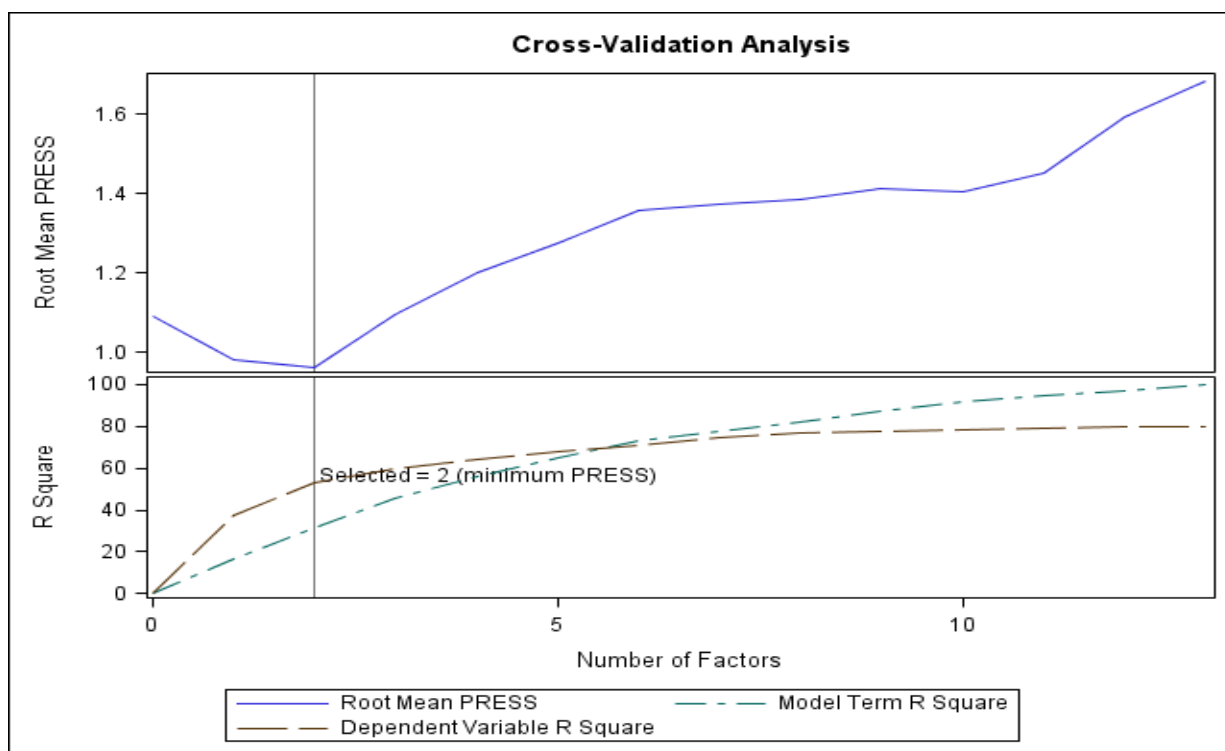
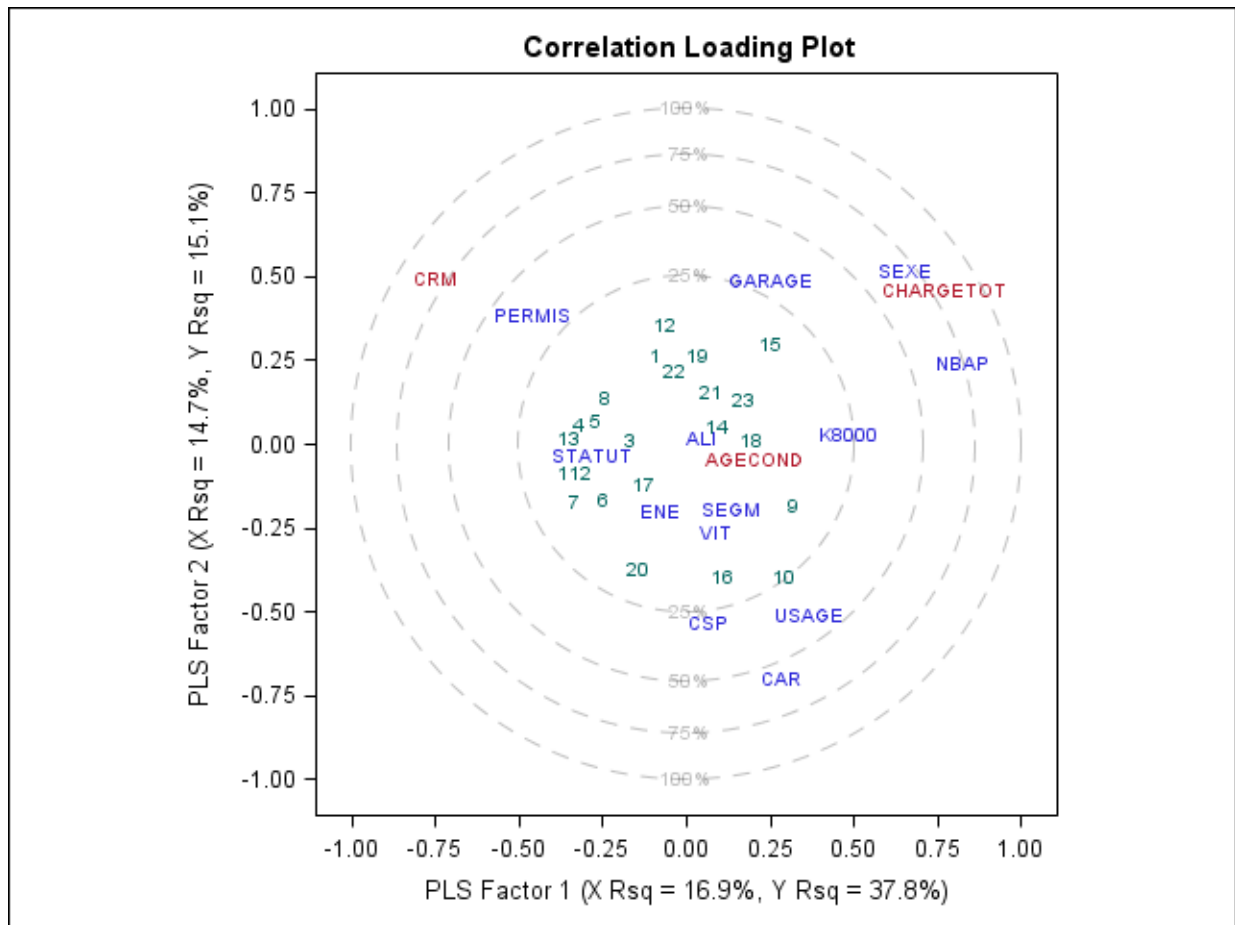


Figure n° 6 : graphique des corrélations



Programme n° 3 : Méthode de validation croisée

Cette première exécution demande un nombre de facteurs important ainsi qu'une estimation par la méthode de validation croisée de l'erreur de prédiction. Le Code SAS utilisé est le suivant :

```
Ods html;

Proc pls data=luc.basedonnee2 cv=one nfac=10;

    model chargetot= AGECOND CRM  NBAP CSP K8000 SEXE STATUT USAGE
    ENE PERMIS GARAGE ALI VIT SEGM CAR;

run;

Ods html close ;
```

Les résultats obtenus sont présentés dans la figure n°7 suivant :

The PLS Procedure

Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Partial Least Squares
PLS Algorithm	NIPALS
Number of Response Variables	1
Number of Predictor Parameters	15
Missing Value Handling	Exclude
Maximum Number of Factors	10
Validation Method	Leave-one-out Cross Validation

Number of Observations Read	23
Number of Observations Used	23

Cross Validation for the Number of Extracted Factors	
Number of Extracted Factors	Root Mean PRESS
0	1.045455
1	0.706411
2	0.643191
3	0.643862
4	0.672538
5	0.646289
6	0.641452

Cross Validation for the Number of Extracted Factors	
Number of Extracted Factors	Root Mean PRESS
7	0.619294
8	0.618508
9	0.604879
10	0.581677

Minimum root mean PRESS	0.5817
Minimizing number of factors	10

The SAS System

The PLS Procedure

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	16.3083	16.3083	77.0764	77.0764
2	14.4910	30.7994	9.2291	86.3055
3	7.1477	37.9471	5.3775	91.6830
4	11.0574	49.0045	2.3323	94.0153
5	9.1332	58.1377	1.6600	95.6753
6	7.7713	65.9090	0.5921	96.2674
7	2.6984	68.6074	0.9749	97.2423
8	6.5631	75.1705	0.2291	97.4714

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
9	6.0684	81.2389	0.1562	97.6276
10	4.8607	86.0996	0.1354	97.7631

b. Méthodes d'extraction des facteurs

Le nombre de facteurs est sélectionné comme étant celui qui minimise l'estimation de l'erreur par validation croisée. Le modèle est ensuite réestimé en fixant le nombre de facteurs et en introduisant le 2^{ème} groupe d'observation pour calculer l'erreur de prédiction sur ce groupe. Pour l'extraction des facteurs ou composante, l'option METHOD est utilisée. Pour cela, l'une des méthodes suivantes est spécifiée :

b1.Méthode de régression par réduction des rangs

Le code SAS qui permet de mettre en pratique cette méthode est le suivant :

```
Ods html;
```

```
Proc pls data=luc.basedonnee2 nfac=1 method= r r r;
```

```
    model chargetot= AGECOND CRM  NBAP CSP K8000 SEXE STATUT USAGE
ENE PERMIS GARAGE ALI VIT SEGM CAR;
```

```
run;
```

```
Ods html close;
```

Les résultants obtenus sont présentés dans la figure n°8 :

The SAS System

The PLS Procedure

Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Reduced Rank Regression
Number of Response Variables	1

Number of Predictor Parameters	15
Missing Value Handling	Exclude
Number of Factors	1

Number of Observations Read	23
Number of Observations Used	23

The SAS System

The PLS Procedure

Percent Variation Accounted for by Reduced Rank Regression Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	11.6986	11.6986	98.1419	98.1419

b.2 Méthode de la régression sur composantes principales

L'option PCR permet de sélectionner la régression sur composantes principales. Le code SAS suivant fait état de cette application.

```
Ods html;
```

```
Proc pls data=luc.basedonnee2 nfac=2 method= pcr;
```

```
    model chargetot= AGECOND CRM  NBAP CSP K8000 SEXE STATUT USAGE
    ENE PERMIS GARAGE ALI VIT SEGM CAR;
```

```
run;
```

```
Ods html close;
```

Les résultats obtenus sont présentés dans la figure n°9 :

The SAS System

The PLS Procedure

Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Principle Components Regression
Number of Response Variables	1
Number of Predictor Parameters	15
Missing Value Handling	Exclude
Number of Factors	2

Number of Observations Read	23
Number of Observations Used	23

The SAS System

The PLS Procedure

Percent Variation Accounted for by Principal Components				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	20.1046	20.1046	20.6871	20.6871
2	17.5138	37.6184	3.7314	24.4185

b.3 Méthode des moindres carrés partiels

Cette méthode permet d'utiliser par défaut de l'algorithme standard des moindres carrés partiels. Le code SAS suivant permet de mettre en œuvre cette dernière.

```
Ods html;
```

```
Proc pls data=luc.basedonnee2 nfac=2 method = pls;
```

```
    model chargetot= AGECOND CRM  NBAP CSP K8000 SEXE STATUT USAGE  
    ENE PERMIS GARAGE ALI VIT SEGM CAR;
```

```
run;
```

```
Ods html close;
```

Les résultants obtenus sont présentés dans la figure n°10 :

The SAS System

The PLS Procedure

Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Partial Least Squares
PLS Algorithm	NIPALS
Number of Response Variables	1
Number of Predictor Parameters	15
Missing Value Handling	Exclude
Number of Factors	2

Number of Observations Read	23
Number of Observations Used	23

The SAS System

The PLS Procedure

Percent Variation Accounted for by Partial Least Squares Factors		
Number of Extracted	Model Effects	Dependent Variables

Factors	Current	Total	Current	Total
1	16.3083	16.3083	77.0764	77.0764
2	14.4910	30.7994	9.2291	86.3055

C. Choix optimal du nombre de facteurs PLS

Le choix du nombre optimal du nombre de facteur PLS s'effectue pour l'utilisation de l'option CV=TESTSET. Le code suivant permet d'effectuer ce test de validation du nombre de facteurs. Les résultats obtenus sont présentés dans la Figure n° 11.

Programme n° 3: Code SAS permettant d'effectuer le test par validation croisée

Cette option permet de spécifier une table SAS contenant les observations utilisées comme échantillon test dans la procédure de validation croisée. Le code suivant met en évidence cette procédure :

```

Ods html;

Ods graphics on;

Proc pls data=luc.basedonnee2 nfac=10 cv=testset(ftest)
                                cvtest(stat=press seed=12345);
    model chargetot AGECOND CRM = NBAP CSP K8000 SEXE STATUT USAGE
    ENE PERMIS GARAGE ALI VIT SEGM CAR;

run;

Ods graphics off;

Ods html close;

```

The SAS System

The PLS Procedure

Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Partial Least Squares

PLS Algorithm	NIPALS
Number of Response Variables	3
Number of Predictor Parameters	13
Missing Value Handling	Exclude
Maximum Number of Factors	10
Validation Method	Test Set Validation (WORK_NULL_)
Validation Testing Criterion	Prob PRESS > 0.1
Number of Random Permutations	1000
Random Permutation Seed	12345

Number of Observations Read	23
Number of Observations Used	23

D. Construction des facteurs PLS

La construction de ces facteurs PLS s'effectue par le code SAS suivant.

Programme n°4 : Code SAS permettant de construire les facteurs

```

Ods html;

Ods graphics on;

proc pls data=luc.basedonnee2 nfac=4 plot=XLoadingProfiles;

    model chargetot AGECOND CRM = NBAP CSP K8000 SEXE STATUT USAGE
ENE PERMIS GARAGE ALI VIT SEGM CAR;

run;

Ods graphics off;

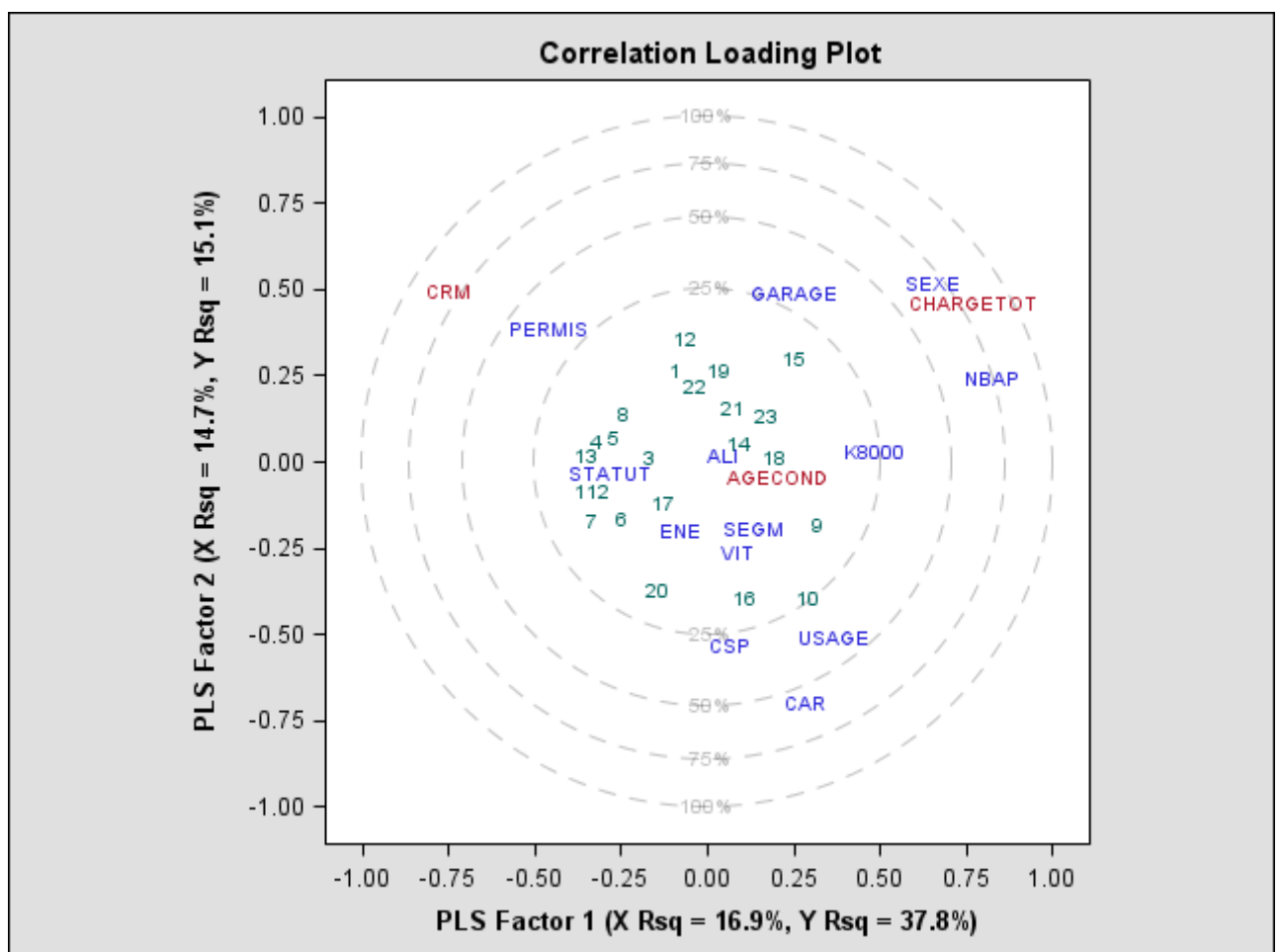
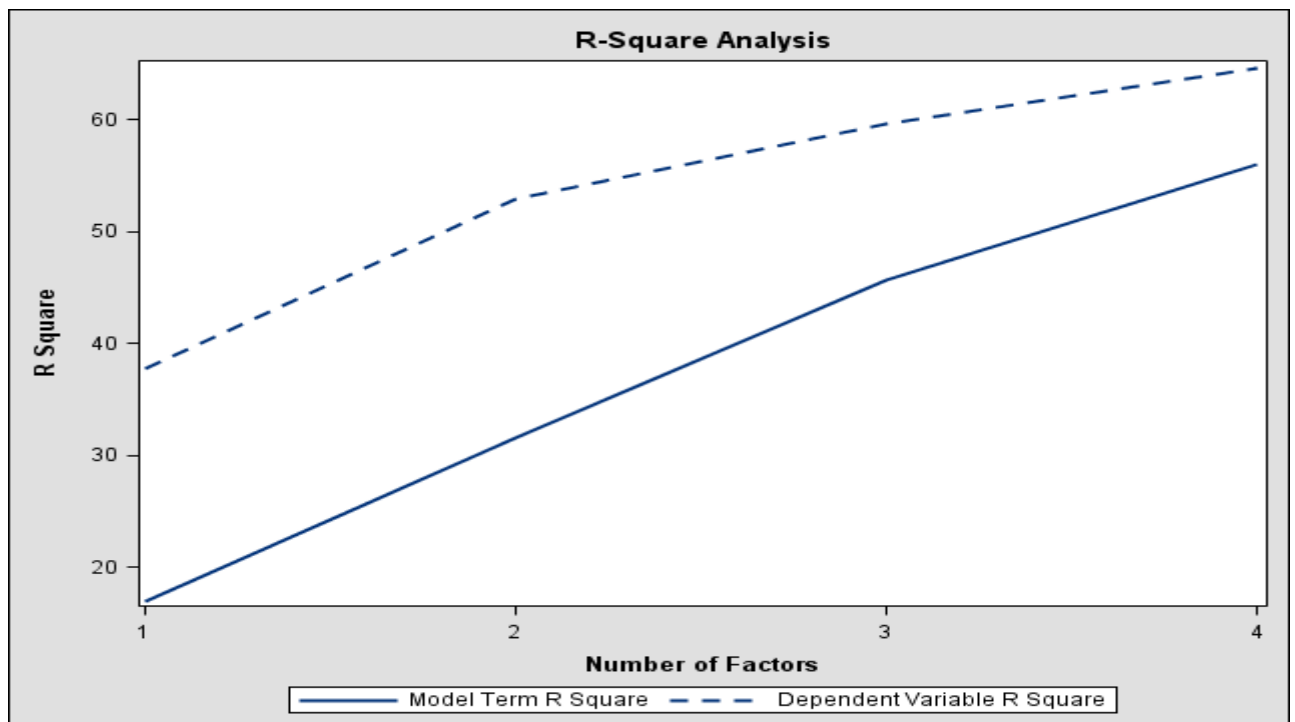
Ods html close;
```

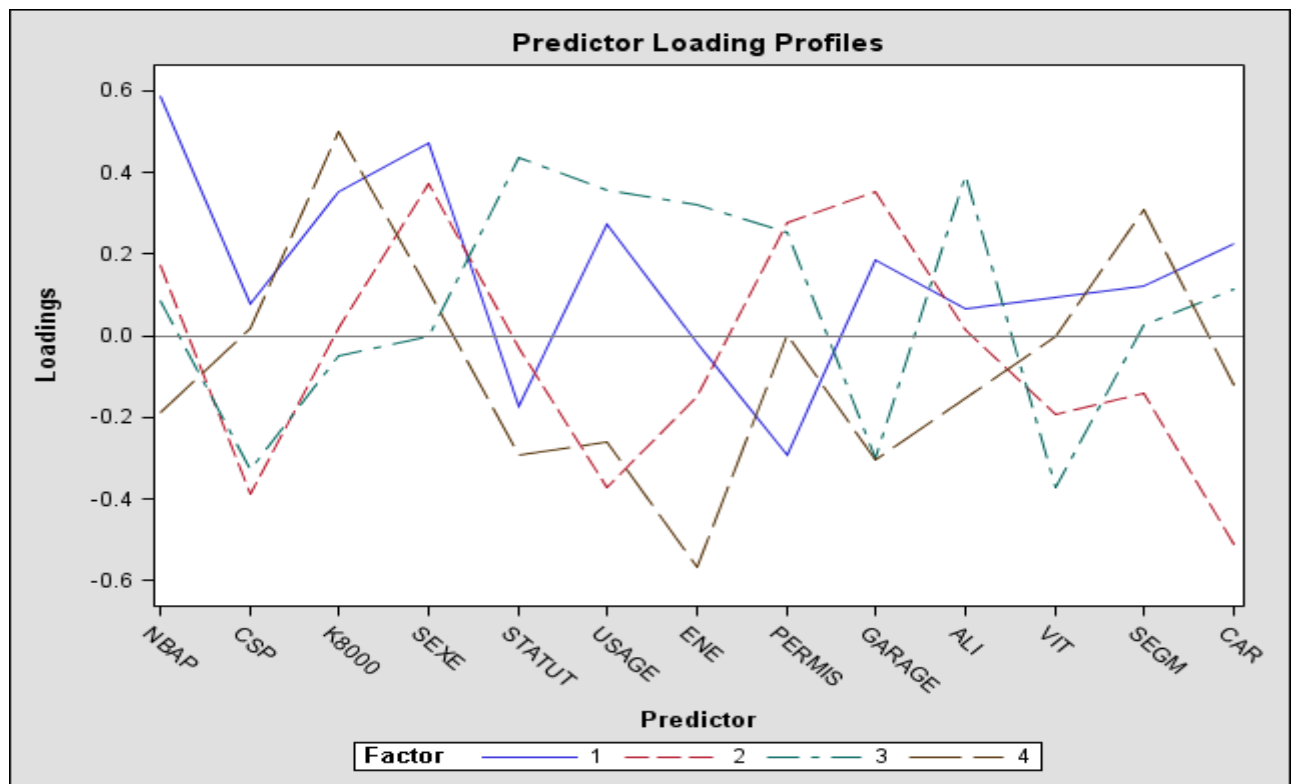
Les résultats obtenus sont ci-après dans la Figure n°12:

Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Partial Least Squares
PLS Algorithm	NIPALS
Number of Response Variables	3
Number of Predictor Parameters	13
Missing Value Handling	Exclude
Number of Factors	4

Number of Observations Read	23
Number of Observations Used	23

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	16.9006	16.9006	37.7543	37.7543
2	14.6929	31.5935	15.1094	52.8637
3	14.0355	45.6290	6.7835	59.6472
4	10.4076	56.0366	4.9451	64.5923





Nous constatons que ce graphique montre les quatre facteurs ou composantes.

e. L'analyse du score des facteurs optimaux

e.1 Programme n° : Score des facteurs optimaux

La construction de ces facteurs PLS s'effectue par le code SAS suivant.

```
Ods html;
```

```
Ods graphics on;
```

```
proc pls data=luc.basedonnee2 nfac=4 plot=XYScores;
```

```
    model chargetot= AGECOND CRM NBAP CSP K8000 SEXE STATUT USAGE ENE  
    PERMIS GARAGE ALI VIT SEGM CAR;
```

```
run;
```

```
Ods graphics off;
```

```
Ods html close;
```

Les résultats obtenus sont ci-après dans la Figure n°13:

The SAS System

The PLS Procedure

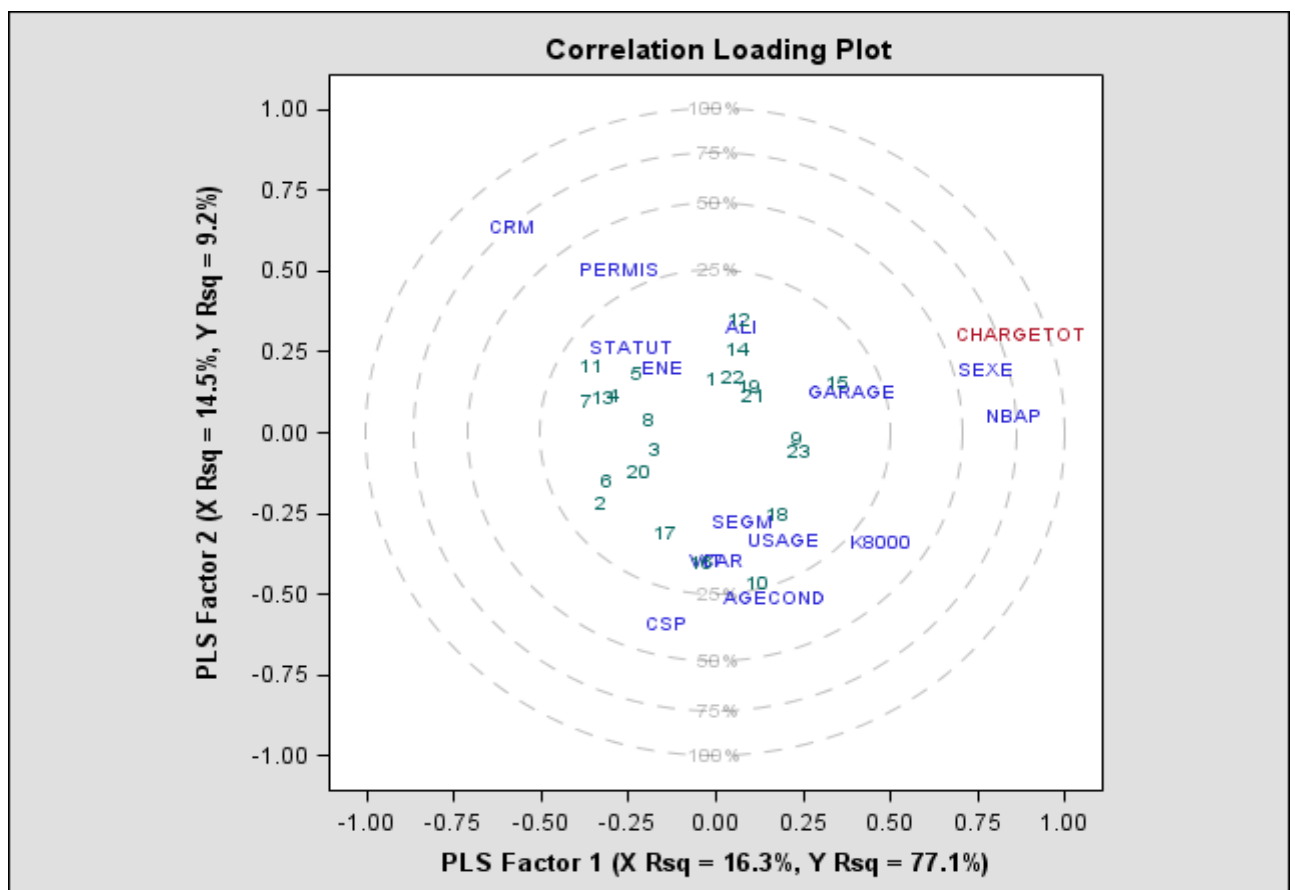
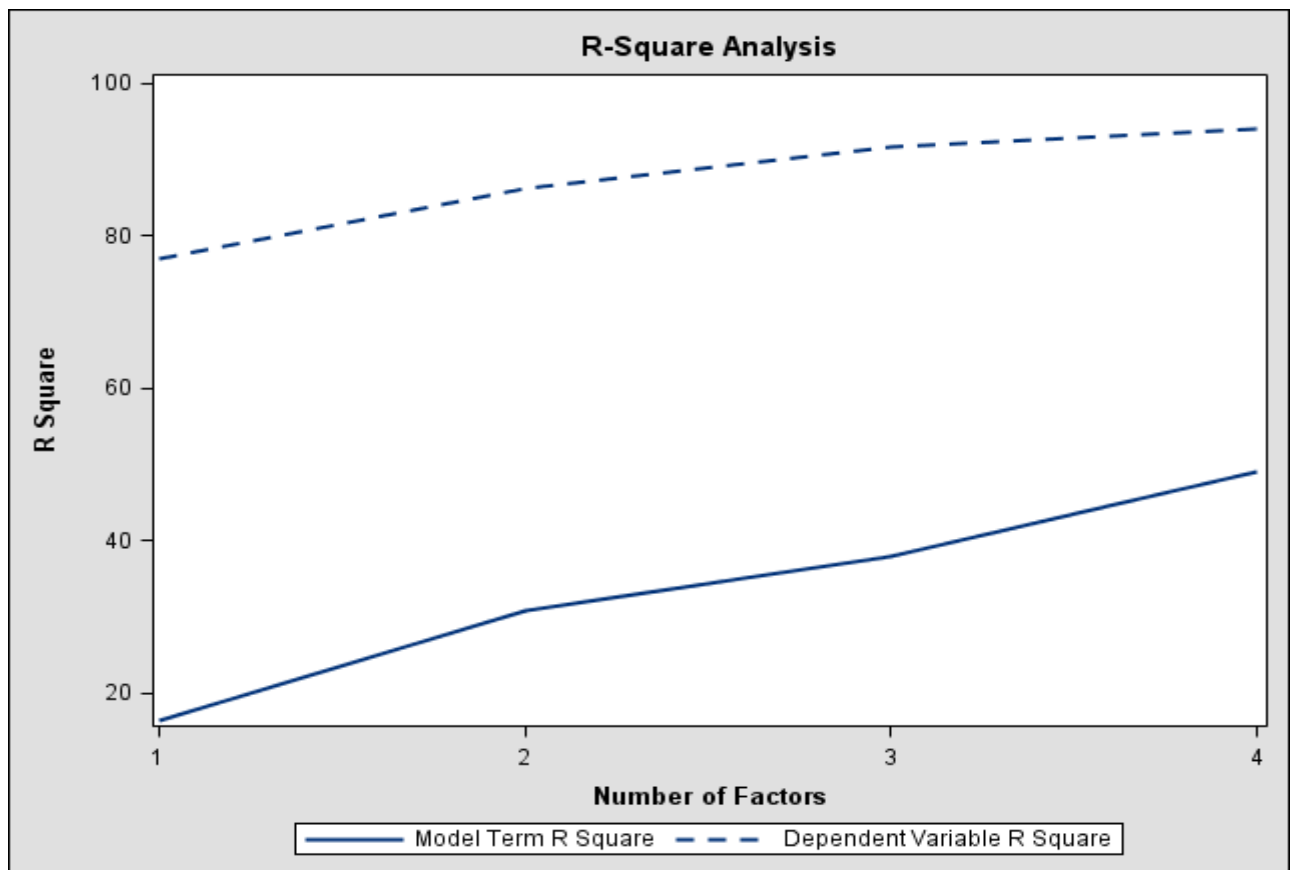
Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Partial Least Squares
PLS Algorithm	NIPALS
Number of Response Variables	1
Number of Predictor Parameters	15
Missing Value Handling	Exclude
Number of Factors	4

Number of Observations Read	23
Number of Observations Used	23

The SAS System

The PLS Procedure

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	16.3083	16.3083	77.0764	77.0764
2	14.4910	30.7994	9.2291	86.3055
3	7.1477	37.9471	5.3775	91.6830
4	11.0574	49.0045	2.3323	94.0153





e.4 Programme n° : présentation des variables d'importance et profils des paramètres estimés.

```
Ods html;
```

```
Ods graphics on;
```

```
proc pls data=luc.basedonnee2 nfac=4 plot=(ParmProfiles VIP);
```

```
    model charge_tot= AGECOND CRM NBAP CSP K8000 SEXE STATUT USAGE ENE  
    PERMIS GARAGE ALI VIT SEGM CAR;
```

```
run;
```

```
Ods graphics off;
```

```
Ods html close;
```

Les résultats obtenus sont ci-après dans la Figure n°14:

The SAS System

The PLS Procedure

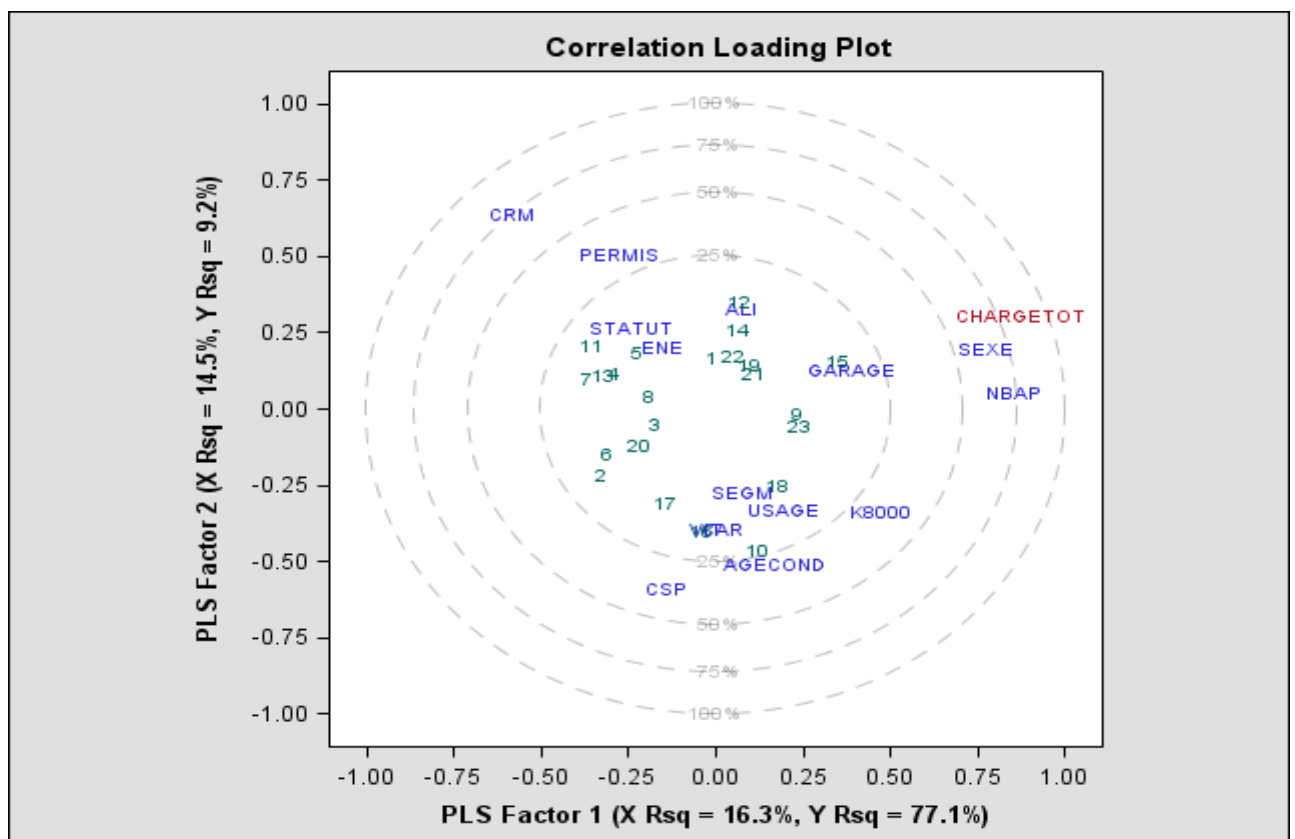
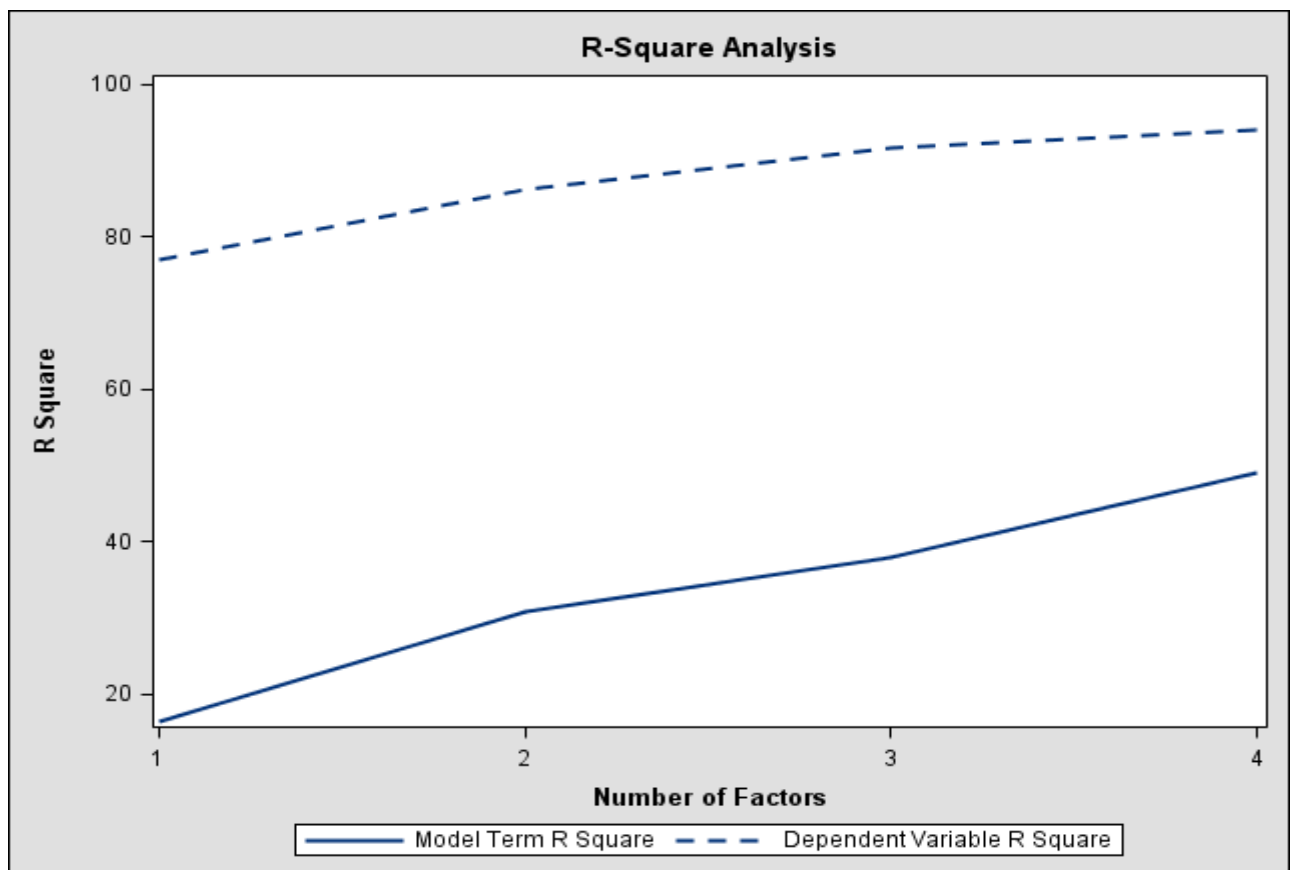
Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Partial Least Squares
PLS Algorithm	NIPALS
Number of Response Variables	1
Number of Predictor Parameters	15
Missing Value Handling	Exclude
Number of Factors	4

Number of Observations Read	23
Number of Observations Used	23

The SAS System

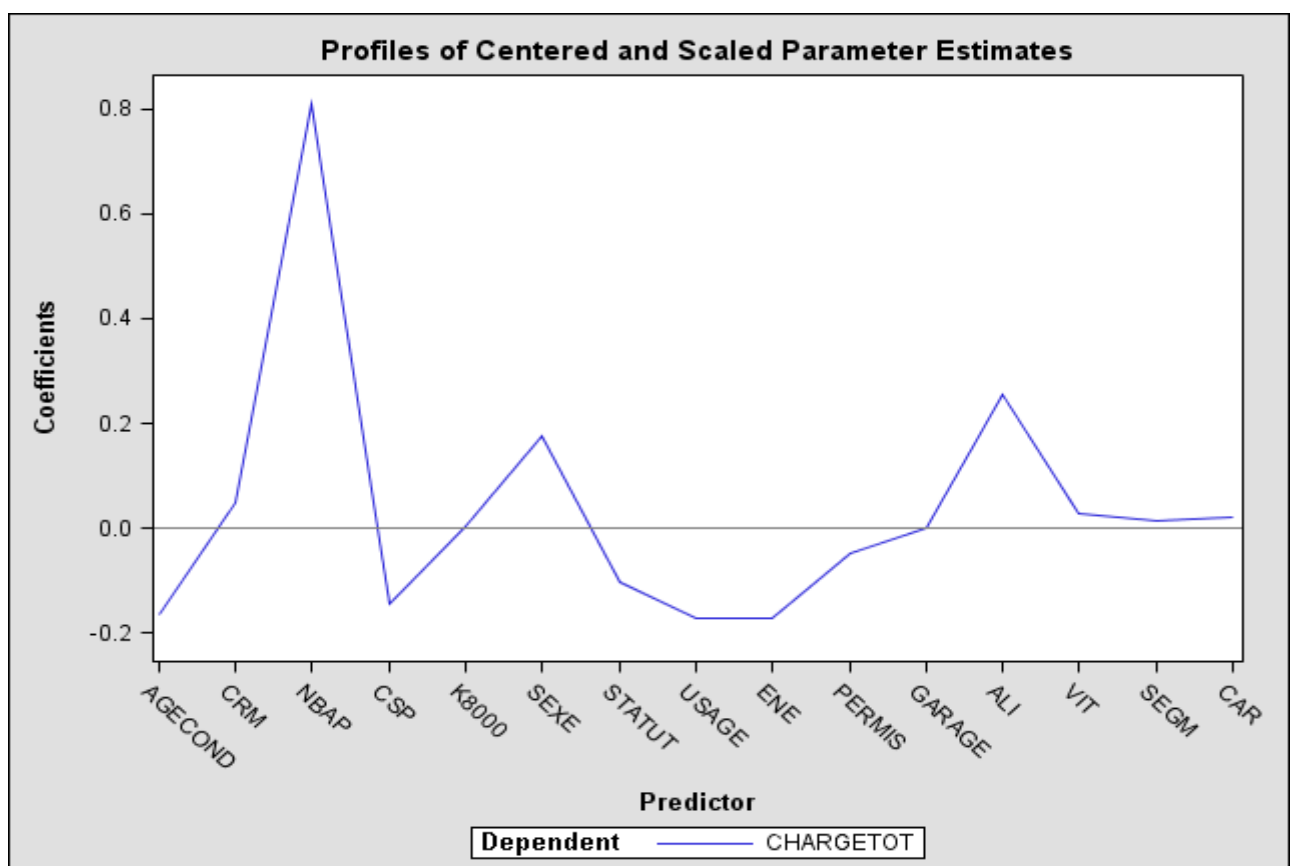
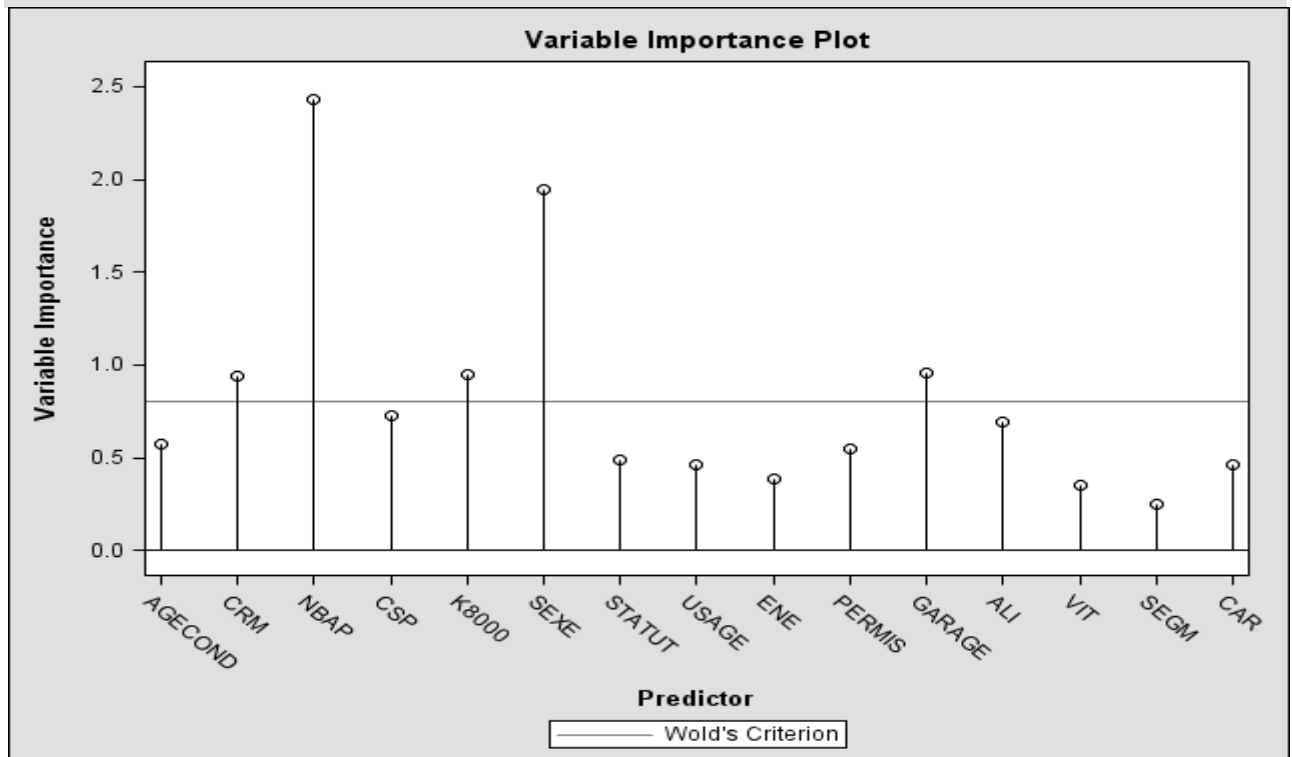
The PLS Procedure

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	16.3083	16.3083	77.0764	77.0764
2	14.4910	30.7994	9.2291	86.3055
3	7.1477	37.9471	5.3775	91.6830
4	11.0574	49.0045	2.3323	94.0153



The PLS Procedure

The SAS System



f. Analyse descriptive

Le Code SAS utiliser pour effectuer cette analyse est le suivant:

```
Ods html;
```

```
Ods graphics on;
```

```
proc pls data=luc.basedonnee2 nfac=2 plot=(diagnostics dmod);
```

```
    model chargetot=  AGECOND  CRM  NBAP
```

```
                  CSP      K8000
```

```
                  SEXE STATUT USAGE ENE PERMIS GARAGE ALI
```

```
                  VIT SEGM CAR;
```

```
run;
```

```
Ods graphics off;
```

```
Ods html close;
```

Les résultats obtenus sont ci-après dans la Figure n°15:

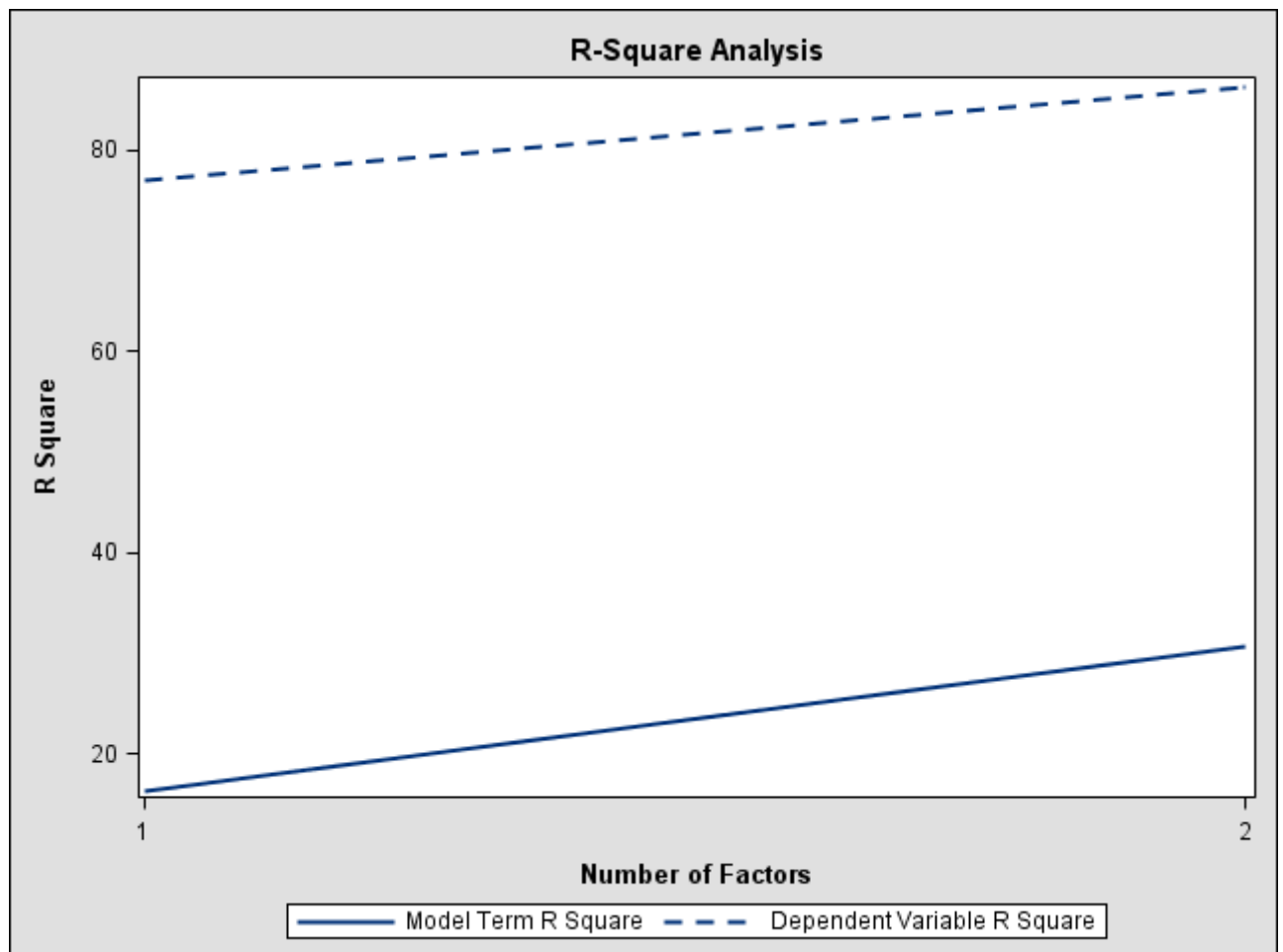
The SAS System

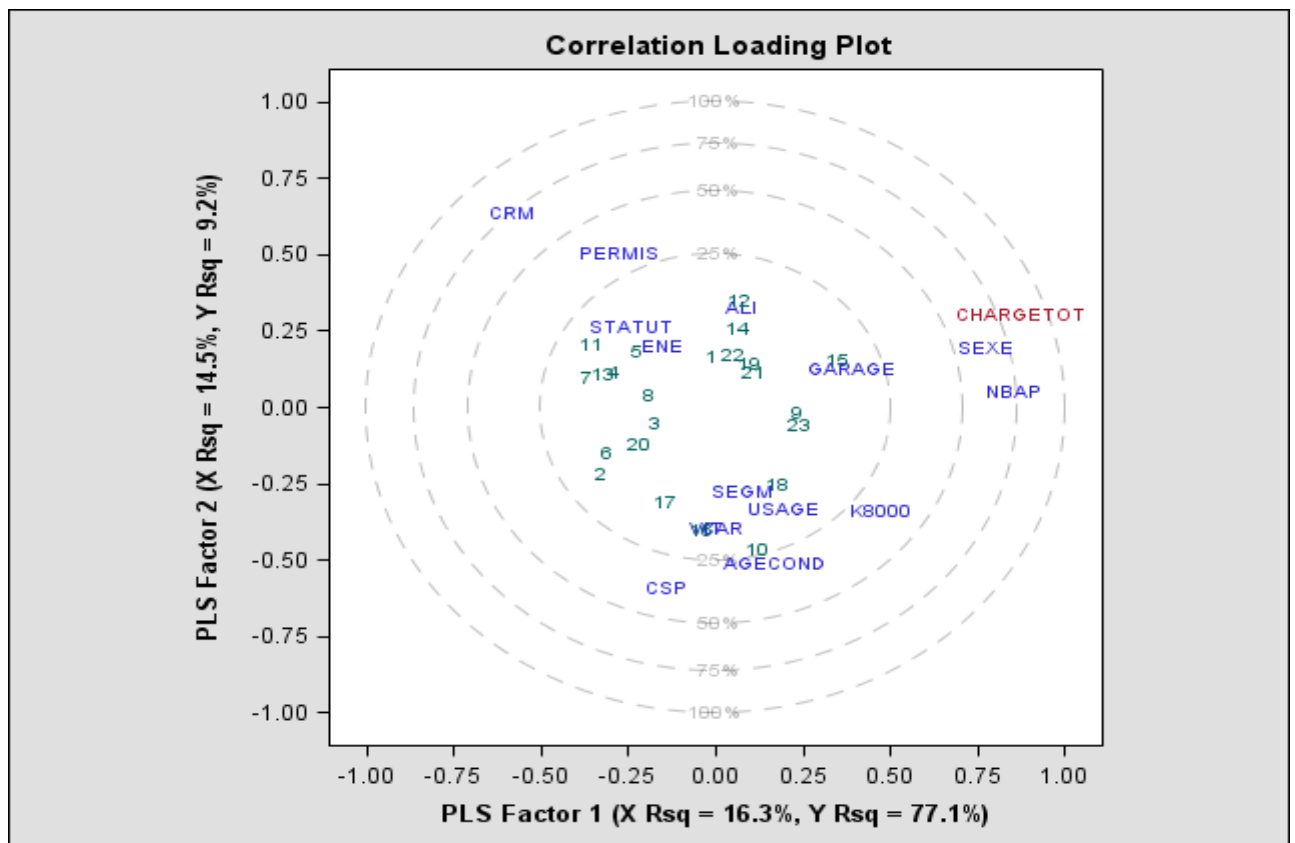
The PLS Procedure

Data Set	LUC.BASEDONNEE2
Factor Extraction Method	Partial Least Squares
PLS Algorithm	NIPALS
Number of Response Variables	1
Number of Predictor Parameters	15
Missing Value Handling	Exclude
Number of Factors	2

Number of Observations Read	23
Number of Observations Used	23

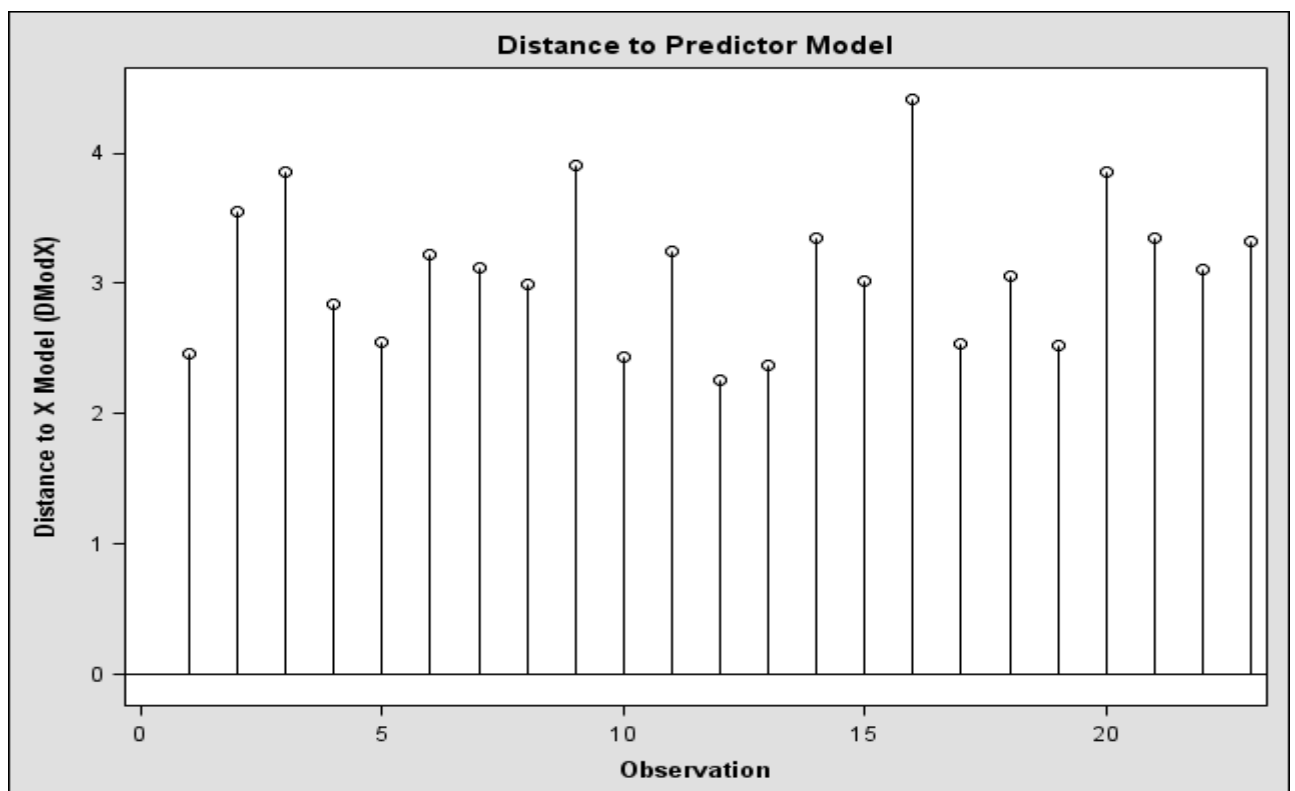
Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	16.3083	16.3083	77.0764	77.0764
2	14.4910	30.7994	9.2291	86.3055

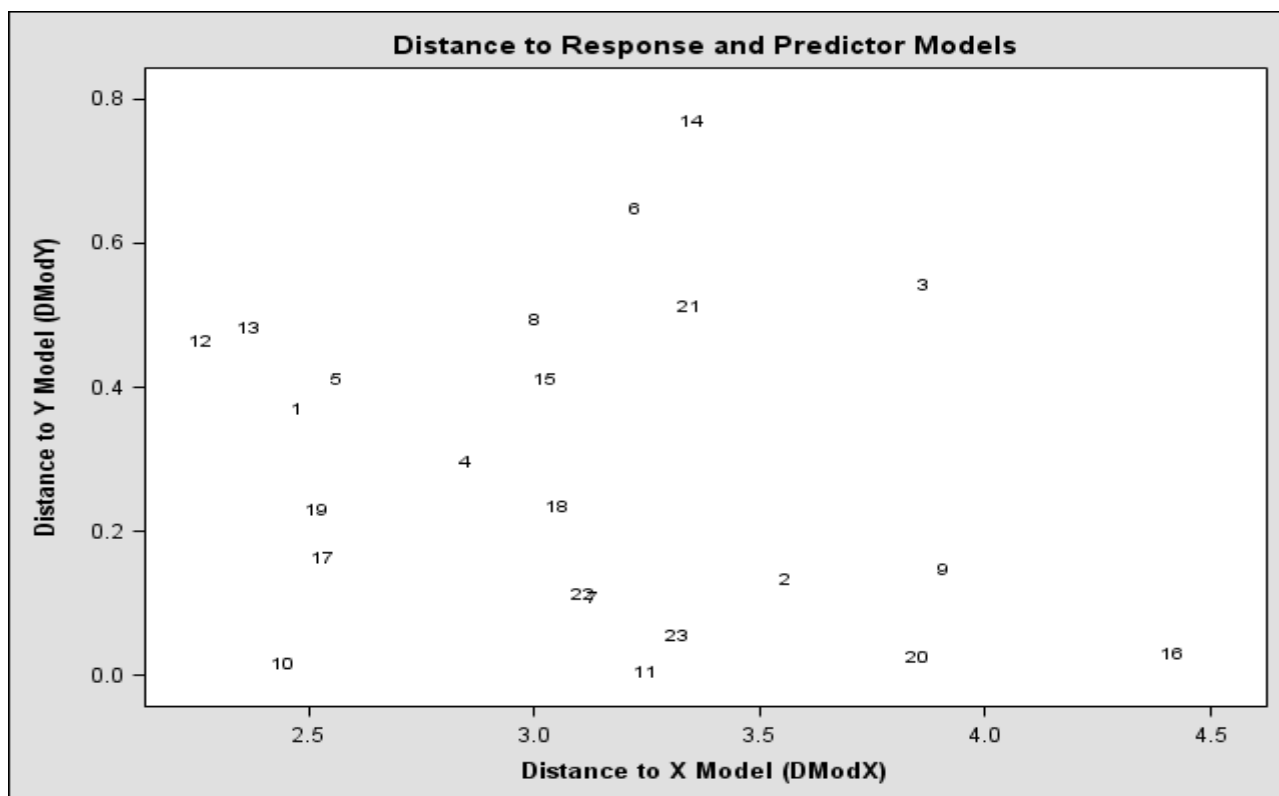
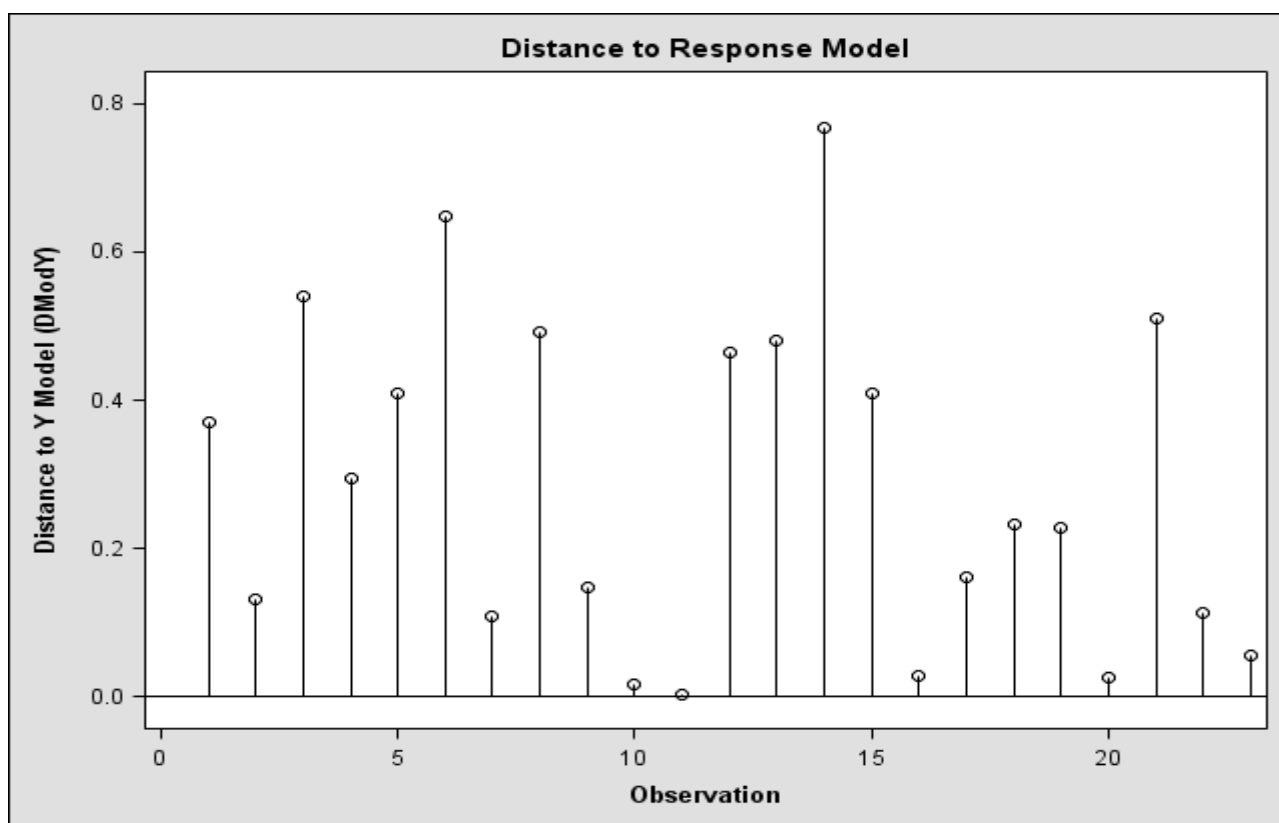




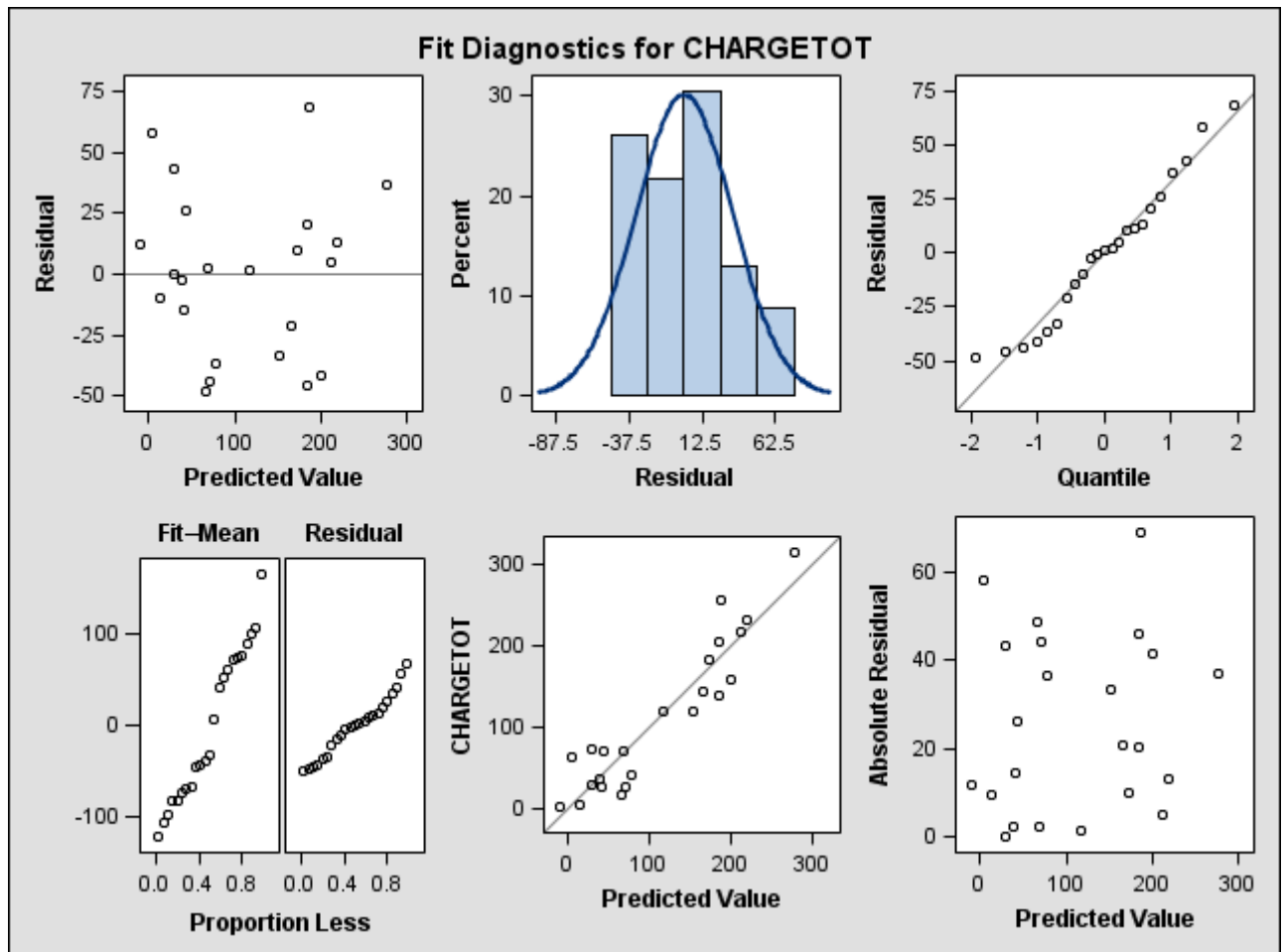
The SAS System

The PLS Procedure





The PLS Procedure



II. CONCLUSION

Dans ce projet, on s'est proposé d'analyser les variables tarifaires d'un portefeuille d'assurance automobile. Notre objectif consistait à présenter quelques éléments permettant de mieux lier les aspects théoriques servant à la mise en pratique de la régression PLS du logiciel SAS. Pour mener ce travail, deux axes nous ont paru intéressants. Le premier axe nous a permis d'examiner le cadre théorique de la procédure PLS et de mettre l'accent sur les méthodes d'estimation du jeu de données.

En abordant le deuxième axe de ce travail, nous avons mis en pratique cette procédure en analysant un portefeuille d'assurance automobile.

ANNEXE n°:1 Code SAS de Transformation et de codage des variables Le code SAS est le suivant :

```
/*****Transformation des variables tarifaires *****/
```

```
Data LUC.basedonnee2;
```

```
set LUC.basedonnee1;
```

```
If ALI=CAR then ALI1=0;
```

```
If ALI=INJ then ALI1=1;
```

```
If ALI=INS then ALI1=2;
```

```
If ALI=IDS then ALI1=3;
```

```
DROP ALI;
```

```
RENAME ALI1=ALI;
```

```
If STATUT=CELIBATAIRE then STATUT1=0;
```

```
If STATUT=MARIE then STATUT1=1;
```

```
If STATUT=DIVORCE then STATUT1=2;
```

```
If STATUT=CONCUBIN then STATUT1=3;
```

```
DROP STATUT;
```

```
RENAME STATUT1=STATUT;
```

```
If SEXE= HOMME then SEXE1=0;
```

```
If SEXE=FEMME then SEXE1=1;
```

```
If SEXE= PERSONNE MORALE then SEXE1=2;
```

```
Drop SEXE;
```

```
RENAME SEXE1=SEXE;
```

If K8000=N then K80001=0; else K80001=1;

Drop K8000;

RENAME K80001=K8000;

If USAGE = PRIVE TRAJET then USAGE1=0;

If USAGE = PROFESSIONNEL then USAGE1=1;

DROP USAGE;

RENAME USAGE1=USAGE;

If ENE=ESSENCE then ENE1=0;

If ENE=DIESEL then ENE1=1;

If ENE=GO then ENE1=2;

DROP ENE;

RENAME ENE1=ENE;

If CAR=BERLINE then CAR1=0;

If CAR=CAMIONETTE then CAR1=1;

If CAR=COUPÊ then CAR1=2;

DROP CAR;

RENAME CAR1=CAR;

If VIT="00-100" then VIT1=0;

If VIT="100-150" then VIT1=1;

If VIT="150-200" then VIT1=2;

If VIT="200-220" then VIT1=3;

DROP VIT;

RENAME VIT1=VIT;

If GARAGE =INDIVIDUEL CLOS then GARAGE1=0;

If GARAGE = CLOS COLLECTIF then GARAGE1=1;

If GARAGE =AUTRES then GARAGE1=2;

Drop GARAGE;

RENAME GARAGE1=GARAGE;

If SEGM= M then SEGM1=0; else SEGM1=1.

Drop SEGM;

RENAME SEGM1=SEGM;

If PERMIS=NOVISTE then PERMIS1=0;

If PERMIS=MOYEN then PERMIS1=1;

If PERMIS=ANCIEN then PERMIS1=2;

Drop PERMIS;

RENAME PERMIS1=PERMIS;

run;

ANNEXE n°2 Sortie des différents programmes

1. Sortie porgramme 1 :

Figure 1:

The SAS System

18:41 Wednesday, August 29, 2012 2

The PLS Procedure

Percent Variation Accounted for
by Partial Least Squares Factors

Number of		Model Effects		Dependent Variables	
Extracted					
Factors	Current	Total	Current	Total	
1	16.9006	16.9006	37.7543	37.7543	
2	14.6929	31.5935	15.1094	52.8637	
3	14.0355	45.6290	6.7835	59.6472	
4	10.4076	56.0366	4.9451	64.5923	
5	9.2874	65.3239	3.5734	68.1656	
6	7.8545	73.1784	2.6246	70.7902	
7	4.3013	77.4797	3.5848	74.3750	
8	4.7580	82.2377	2.1826	76.5577	
9	4.7432	86.9810	1.2198	77.7775	
10	4.5508	91.5318	0.9693	78.7468	
11	3.4286	94.9603	0.5657	79.3125	
12	2.2032	97.1635	0.3107	79.6232	
13	2.8365	100.0000	0.1091	79.7324	

The SAS System

18:41 Wednesday, August 29,

2012 3

The PLS Procedure

Data Set LUC.BASEDONNEE2
Factor Extraction Method Partial Least Squares

PLS Algorithm	NIPALS
Number of Response Variables	3
Number of Predictor Parameters	13
Missing Value Handling	Exclude
Number of Factors	13

Number of Observations Read	23
Number of Observations Used	23

2012 4

The SAS System

18:41 Wednesday, August 29,

The PLS Procedure

Percent Variation Accounted for
by Partial Least Squares Factors

Number of

Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	16.9006	16.9006	37.7543	37.7543
2	14.6929	31.5935	15.1094	52.8637
3	14.0355	45.6290	6.7835	59.6472
4	10.4076	56.0366	4.9451	64.5923
5	9.2874	65.3239	3.5734	68.1656
6	7.8545	73.1784	2.6246	70.7902

Bibliographie

[1] Help SAS 9.2

[2]. **Tenenhaus M., Gauchi J.,-P., Ménardo C.** *Régression PLS et applications. Revue de statistique appliquée*, 1995, 43(1) : 7-63.

[3]. **Tenenhaus M.** (1998), *La régression PLS. Théorie et pratique*, Editions Technip, Paris.

[4] **Gauchi J.-P.** Utilisation de la régression PLS pour l'analyse des plans d'expériences en chimie de formulation, *Revue de statistique appliquée*, 1995 43(1):65-89.

[5] Article : la Régression PLS par **S.VANCOLEN**, p.11-16, JUIN 2004, Université de Neuchâtel (SUISSE)

[6] Professeur Gilbert SAPORTA (août 2011): « Probabilités, analyse des données et statistique ; 3^e édition technip,