



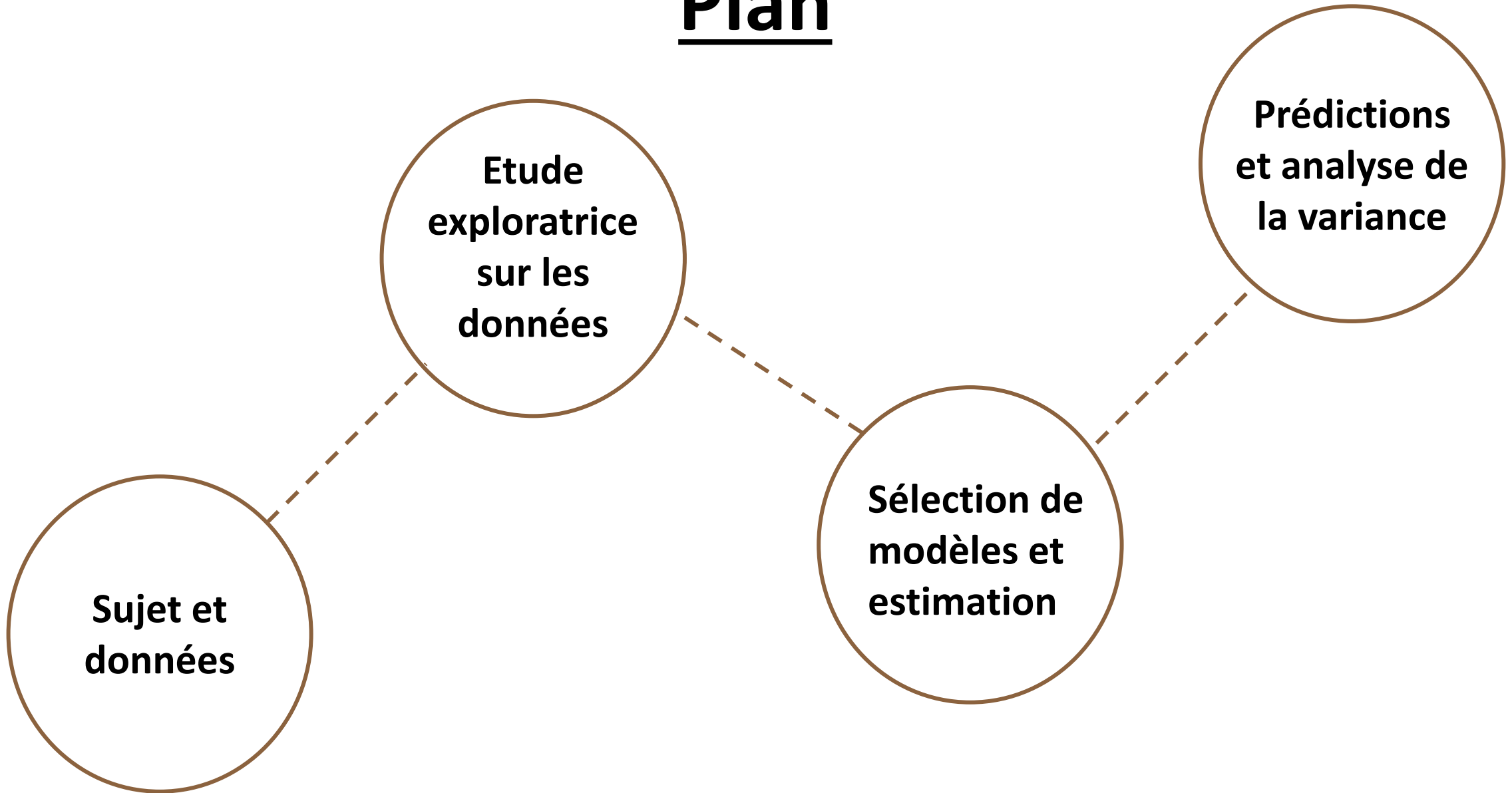
Rapport d'étude statistique sur l'efficacité des techniques utilisées en agriculture

Présenté par :

Michel Nassalang

Elève-Ingénieur à l'Institut polytechnique de Saint-Louis
Cycle Informatique

Plan



Sujets et données

Notre sujet porte sur des données provenant d'une enquête sur l'agriculture dans la région de Podor. Cette enquête avait pour but d'évaluer l'efficacité des techniques et méthodes utilisées dans l'agriculture ou acquises au cours d'une formation. Avec le soutien de l'ONG Enda Pronal, les données ont été récoltées sur un territoire de 43 villages dans la commune de Guédé Village. Elles ont été collectées par un étudiant de l'université Gaston Berger du nom de Félix Charlemagne Sagna. L'étude se base sur les cultures et les terres de chaque famille logeant dans ces villages.

Etude exploratrice sur les données

1. Description des variables

```
Rcmdr> culture <- readXL("C:/Users/Miki_Biboy/Documents/files_R/mydata.xlsx",  
Rcmdr+   rownames=FALSE, header=TRUE, na="", sheet="Feuil5", stringsAsFactors=TRUE)  
RcmdrMsg: [2] NOTE: Le jeu de données culture a 93 lignes et 9 colonnes.
```

On a d'abord importé les données et on l'a nommé culture.

Les variables de ce jeu de données sont au nombre de 9 avec :

Sexe =====> Genre du représentant de la famille,

Nombre.d.années =====> Nombre d'années de cultures,

Formation =====> Réalisation de la formation,

Association =====> Présence dans une association

Rendement..Kg.ha =====> Rendement en Kg par Hectare,

Charge.moy.ha =====> Charge moyenne par Hectare,

Revenu.moy.ha =====> Revenu moyenne par Hectare,

Couverture alimentaire =====> Existence d'une couverture alimentaire,

Superficie..ha =====> Superficie par Hectare.

	Sexe	Nombre d'années	Formation	Association	Rendement (Kg/ha)	Charge moy/ha	Revenu moy/ha	Couverture alimentaire	Superficie (ha)
1	Masculin	17	0	0	12517	375000	1125000	0	1.00
2	Masculin	15	0	1	12517	375000	1125000	1	0.70
3	Masculin	16	0	1	32718	350000	2453800	0	0.90
4	Masculin	12	0	0	1487	800000	3717000	1	1.00
5	Masculin	12	1	0	12517	375000	1125000	0	0.80
6	Masculin	14	0	1	12517	375000	1125000	0	1.00
7	Masculin	17	0	0	12517	375000	1125000	0	1.20
8	Masculin	24	0	0	32718	350000	2453800	0	0.60
9	Masculin	7	0	1	12517	375000	1125000	1	0.80
10	Masculin	12	0	0	18660	420000	1772700	1	0.70
11	Masculin	28	0	0	12517	350000	1125000	1	0.67
12	Masculin	16	1	0	5230	155000	392250	1	2.00
13	Masculin	16	0	1	12517	350000	1125000	0	0.75
14	Masculin	10	0	0	12517	350000	1125000	0	0.90
15	Masculin	26	0	0	12517	350000	1125000	0	1.25
16	Masculin	21	0	0	1487	800000	3717000	1	0.75
17	Masculin	38	1	0	12517	350000	1125000	1	0.80
18	Masculin	12	0	1	32718	350000	2453800	0	1.00
19	Masculin	12	1	0	12517	350000	1125000	1	1.50
20	Masculin	16	0	0	12517	350000	1125000	1	0.82
21	Masculin	17	0	0	12517	350000	1125000	0	1.75

Dans l'ensemble des données, on a 4 variables catégorisées qui sont :

- Le sexe des représentants dans les familles
- La formation sur les techniques
- L'existence d'une association
- La couverture alimentaire

Et on a 5 variables quantitatives à savoir :

- Le nombre d'années
- Le rendement
- Les charges moyennes
- Les revenus moyens
- La superficie des terres

```
Rcmdr> summary(culture)
```

Sexe	Nombre.d.années	Formation	Association
Féminin :50	Min. : 2.000	Min. :0.0000	Min. :0.0000
Masculin:43	1st Qu.: 3.000	1st Qu.:0.0000	1st Qu.:1.0000
	Median : 5.000	Median :1.0000	Median :1.0000
	Mean : 8.495	Mean :0.6344	Mean :0.7634
	3rd Qu.:12.000	3rd Qu.:1.0000	3rd Qu.:1.0000
	Max. :38.000	Max. :1.0000	Max. :1.0000
Rendement..Kg.ha.	Charge.moy.ha	Revenu.moy.ha	Couverture.alimentaire
Min. : 1258	Min. :130000	Min. : 392250	Min. :0.0000
1st Qu.: 9250	1st Qu.:130000	1st Qu.: 855000	1st Qu.:0.0000
Median : 9520	Median :280000	Median :1125000	Median :1.0000
Mean :10754	Mean :260484	Mean :1509698	Mean :0.6129
3rd Qu.:12517	3rd Qu.:350000	3rd Qu.:1917900	3rd Qu.:1.0000
Max. :32718	Max. :800000	Max. :3717000	Max. :1.0000
Superficie..ha.			
Min. :0.0300			
1st Qu.:0.0500			
Median :0.3800			
Mean :0.4768			
3rd Qu.:0.8000			
Max. :2.0000			

2. Description des données

Pour une description des données, on a réalisé les statistiques descriptives pour une meilleure approche dans l'analyse portée sur ces données.

```
Rcmdr> numSummary(culture[,c("Association", "Charge.moy.ha",  
Rcmdr+ "Couverture.alimentaire", "Formation", "Nombre.d.années",  
Rcmdr+ "Rendement..Kg.ha.", "Revenu.moy.ha", "Superficie..ha."), drop=FALSE],  
Rcmdr+ statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

	mean	sd	IQR	0%
Association	0.7634409	0.4272727	0.00	0.00
Charge.moy.ha	260483.8709677	143091.0027329	220000.00	130000.00
Couverture.alimentaire	0.6129032	0.4897261	1.00	0.00
Formation	0.6344086	0.4842059	1.00	0.00
Nombre.d.années	8.4946237	7.6195168	9.00	2.00
Rendement..Kg.ha.	10754.1182796	7128.1158234	3267.00	1258.00
Revenu.moy.ha	1509698.3870968	918143.8622075	1062900.00	392250.00
superficie..ha.	0.4767742	0.4420657	0.75	0.03

	25%	50%	75%	100%	n
Association	1.00	1.00	1.0	1	93
Charge.moy.ha	130000.00	280000.00	350000.0	800000	93
Couverture.alimentaire	0.00	1.00	1.0	1	93
Formation	0.00	1.00	1.0	1	93
Nombre.d.années	3.00	5.00	12.0	38	93
Rendement..Kg.ha.	9250.00	9520.00	12517.0	32718	93
Revenu.moy.ha	855000.00	1125000.00	1917900.0	3717000	93
superficie..ha.	0.05	0.38	0.8	2	93

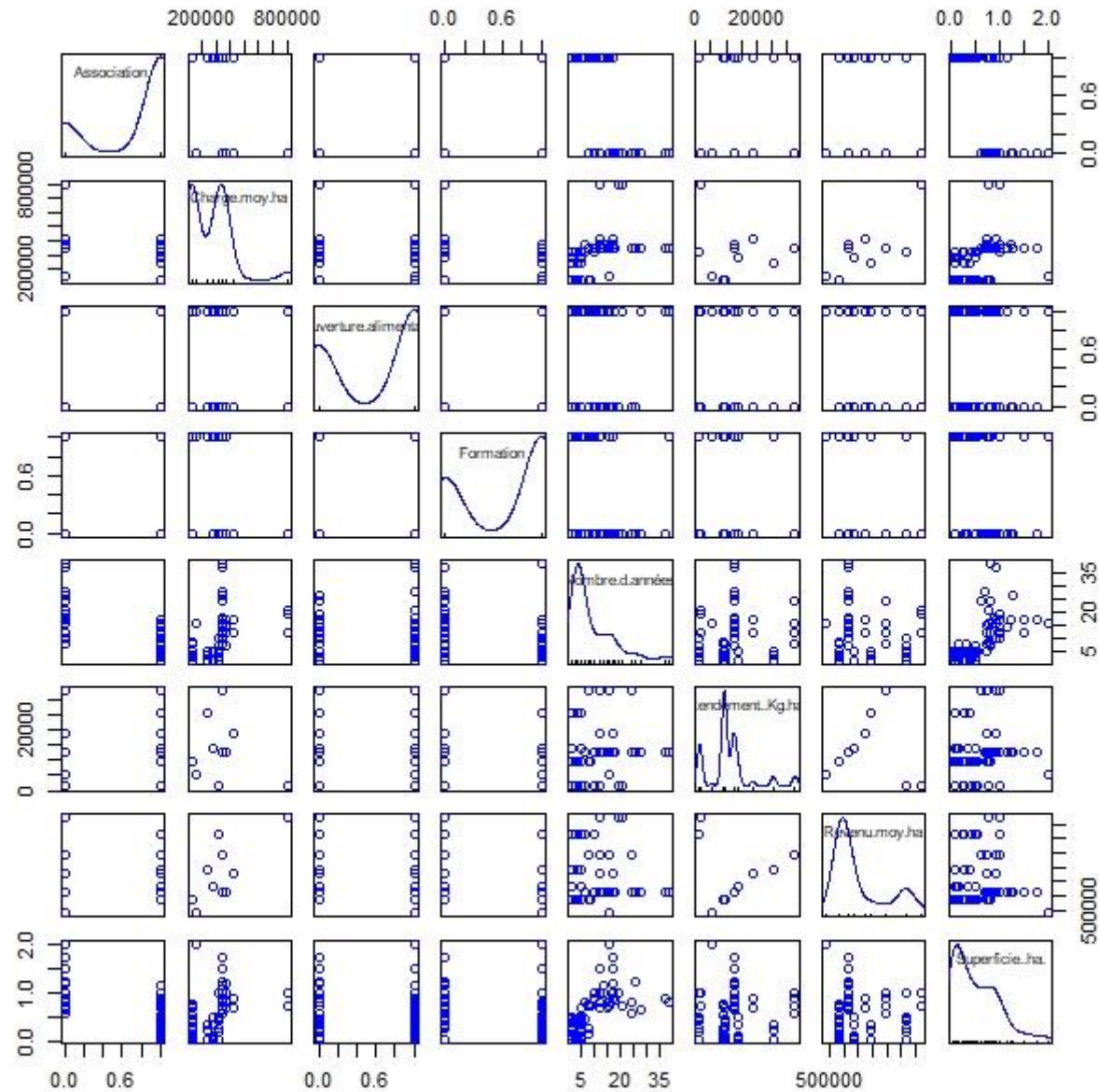
Par la suite, dans la description nous avons déterminé la matrice de corrélation qui donne les coefficients de corrélation de chaque variable avec une autre variable.

```
Rcmdr> cor(culture[,c("Charge.moy.ha", "Formation", "Nombre.d.années",  
Rcmdr+ "Rendement..Kg.ha.", "Revenu.moy.ha", "Superficie..ha.")], use="complete")
```

	Charge.moy.ha	Formation	Nombre.d.années	Rendement..Kg.ha.
Charge.moy.ha	1.00000000	-0.4939468	0.53773113	-0.03263265
Formation	-0.49394676	1.0000000	-0.45719160	-0.25749533
Nombre.d.années	0.53773113	-0.4571916	1.00000000	0.21511856
Rendement..Kg.ha.	-0.03263265	-0.2574953	0.21511856	1.00000000
Revenu.moy.ha	0.67245786	-0.1009121	0.02363724	-0.24073221
Superficie..ha.	0.49050375	-0.3981012	0.66830488	0.18323422

	Revenu.moy.ha	Superficie..ha.
Charge.moy.ha	0.67245786	0.49050375
Formation	-0.10091208	-0.39810118
Nombre.d.années	0.02363724	0.66830488
Rendement..Kg.ha.	-0.24073221	0.18323422
Revenu.moy.ha	1.00000000	-0.01423204
Superficie..ha.	-0.01423204	1.00000000

Nous avons par la suite le graphe de liaison qui montre la dépendance de chaque variable avec une autre variable.



3. Analyse des sorties

Pour l'analyse des sorties, on détermine les types de variables à savoir la variable réponse et les variables explicatives. Pour ce faire, on cherche d'abord la variable réponse en choisissant la variable ayant la valeur de corrélation la plus élevée par rapport aux autres variables. Les autres variables seront considérées par la suite comme des variables explicatives. D'après la matrice de corrélation qu'on a obtenu :

- ✓ Variable réponse : Revenu moyen par hectare
- ✓ Variables explicatives : Charge moyenne par hectare
Nombre d'années
Formation
Rendement
Superficie des terres

Sélection des modèles et estimation

1. Définition des modèles

On crée des modèles évolutifs en fonction de leur significativité .

Nous allons définir les modèles pour faire une sélection du meilleur modèle.

Pour sélectionner le « meilleur modèle » (si tant est qu'il existe), on va s'appuyer sur une mesure qui permet de comparer les modèles entre eux (par exemple, le coefficient de détermination ajusté R^2_a , le critère d'information bayésien (**SIC** ou **BIC**), ou encore le critère d'information d'Akaike (**AIC**)). Il existe deux méthodes souvent employées pour effectuer de la sélection :

- méthode exhaustive ;
- méthode pas-à-pas.

On va donc partir du modèle où la consommation d'électricité est modélisée à l'aide de toutes les variables de notre data frame, et se fier au **critère AIC** pour décider de retirer ou non une variable, à chaque itération.

Le critère **AIC** est défini par :

$$AIC = n \times \log \left(\frac{SCR}{n} \right) + 2(l + 1),$$

avec n le nombre d'observations, SCR la somme des carrés des résidus du modèle estimé et l le nombre de variables explicatives.

Intuitivement, si le modèle s'ajuste mieux aux données, la somme des carrés des résidus va diminuer et donc le terme $n \times \log(SCRn)$ va également diminuer. Le critère AIC devrait donc potentiellement baisser, en fonction de la valeur du terme restant $2(l+1)$, qui vient pénaliser les modèles avec plus de variables. En effet, la somme des carrés des résidus augmente ou reste stable avec l'ajout de variables dans le modèle. Aussi, il est nécessaire de pénaliser l'ajout de variables supplémentaires.

Maintenant, on procède à la définition des modèles :

- **Modèle A**

Ce modèle est basé sur toutes les variables explicatives désignées au départ.

```
call:
lm(formula = Revenu.moy.ha ~ Association + Charge.moy.ha + Couverture.alimentaire +
  Formation + Nombre.d.années + Rendement..Kg.ha. + Superficie..ha.,
  data = culture)
```

Residuals:

Min	1Q	Median	3Q	Max
-965345	-268020	-124040	223325	1411871

Coefficients:

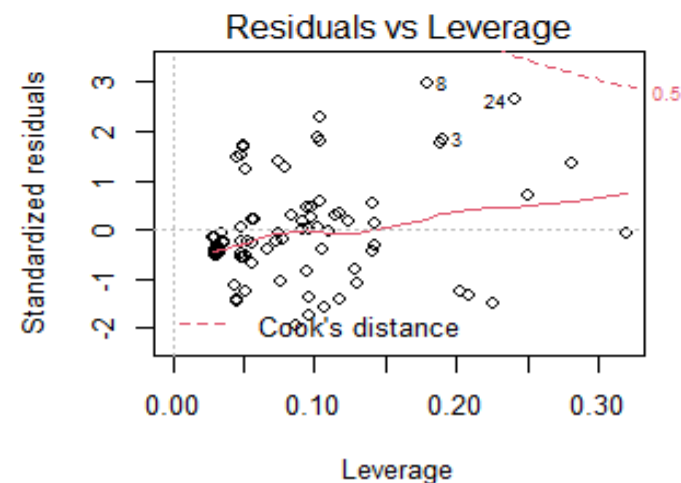
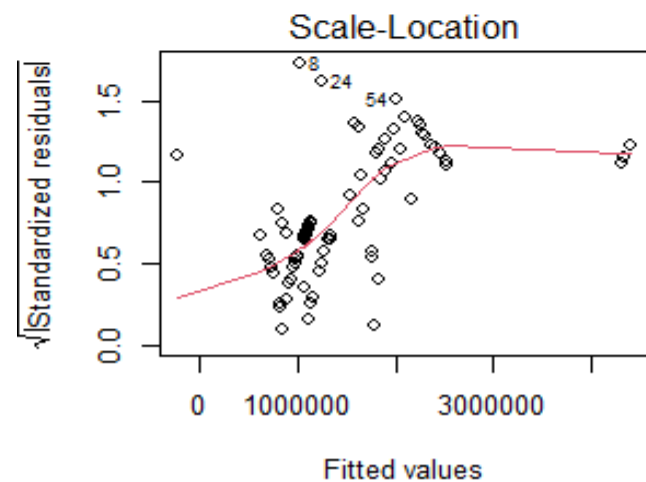
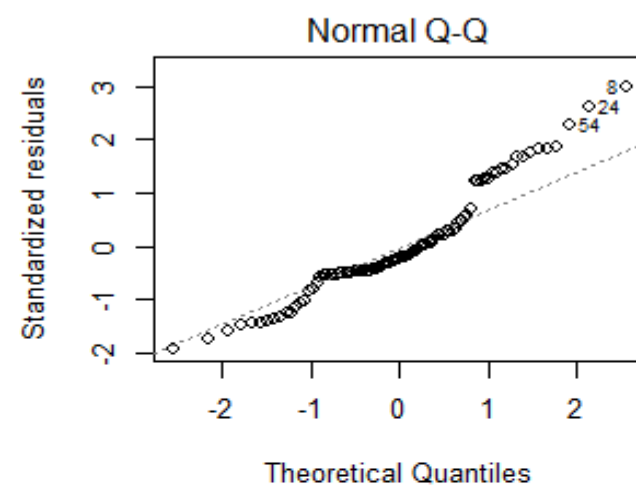
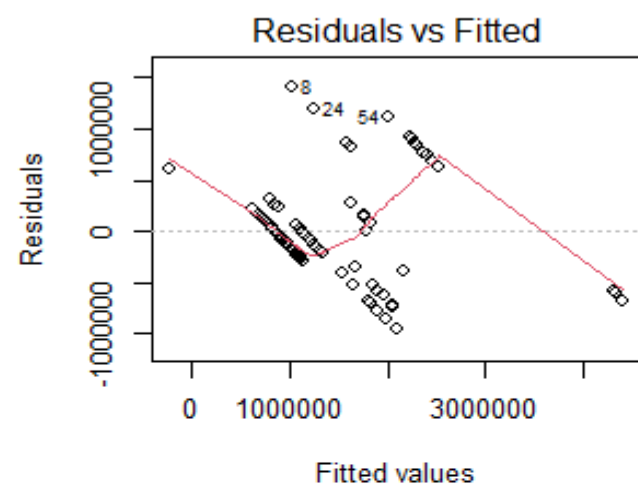
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-438373.0379	364270.7570	-1.203	0.2322	
Association	522961.1667	217725.8031	2.402	0.0185	*
Charge.moy.ha	6.8650	0.5191	13.224	<2e-16	***
Couverture.alimentaire	-27245.6159	116357.5325	-0.234	0.8154	
Formation	320494.9297	141954.0828	2.258	0.0265	*
Nombre.d.années	-18185.2969	11769.6225	-1.545	0.1260	
Rendement..Kg.ha.	-7.5201	8.3192	-0.904	0.3686	
Superficie..ha.	-399942.6017	180765.6287	-2.212	0.0296	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 520700 on 85 degrees of freedom
 Multiple R-squared: 0.7029, Adjusted R-squared: 0.6784
 F-statistic: 28.73 on 7 and 85 DF, p-value: < 2.2e-16

Diagnostics graphiques du modèle A

evenu.moy.ha ~ Association + Charge.moy.ha + Couverture.alimentaire +



Meilleure sélection en backward avec le critère d'information d'Akaike (AIC) sur le modèle A :

Step: AIC=2452.83

Revenu.moy.ha ~ Association + Charge.moy.ha + Formation + Nombre.d.années +
Superficie..ha.

	Df	Sum of Sq	RSS	AIC
<none>			23266530257113	2452.8
- Nombre.d.années	1	741117582800	24007647839913	2453.7
- Superficie..ha.	1	1410721914102	24677252171215	2456.3
- Association	1	1602852419908	24869382677021	2457.0
- Formation	1	1793564947176	25060095204289	2457.7
- Charge.moy.ha	1	52880388068788	76146918325901	2561.1

Call:

```
lm(formula = Revenu.moy.ha ~ Association + Charge.moy.ha + Formation +  
    Nombre.d.années + Superficie..ha., data = culture)
```

Coefficients:

(Intercept)	Association	Charge.moy.ha	Formation
-572328.979	529163.580	6.986	344598.267
Nombre.d.années	Superficie..ha.		
-19338.825	-411352.910		

Variables explicatives retenues avec la sélection : Association, Charge moyenne, Formation, Nombre d'années, Superficie des terres

Modèle B

Ce modèle est basé sur la selection avec le critère d'information d'Akaike (**AIC**) sur le modèle A

call:

```
lm(formula = Revenu.moy.ha ~ Association + Charge.moy.ha + Formation +  
    Nombre.d.années + Superficie..ha., data = culture)
```

Residuals:

Min	1Q	Median	3Q	Max
-987252	-263730	-101154	201816	1291859

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-572328.9790	330012.5950	-1.734	0.0864	.
Association	529163.5800	216147.0745	2.448	0.0164	*
Charge.moy.ha	6.9863	0.4968	14.062	<2e-16	***
Formation	344598.2667	133064.0444	2.590	0.0113	*
Nombre.d.années	-19338.8254	11616.9660	-1.665	0.0996	.
Superficie..ha.	-411352.9104	179101.8420	-2.297	0.0240	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

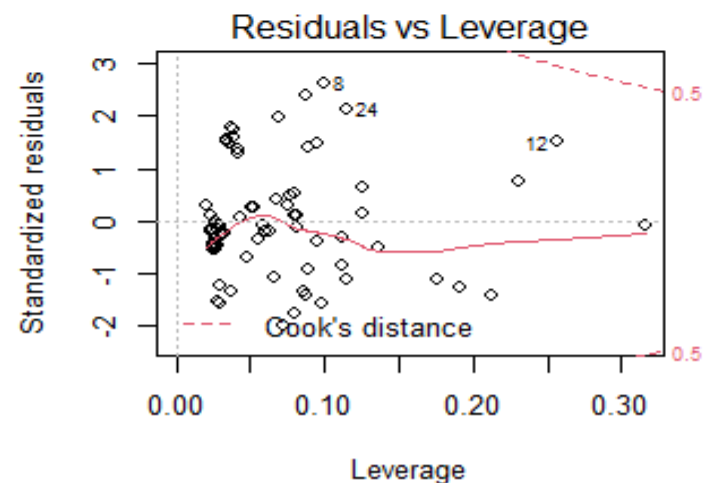
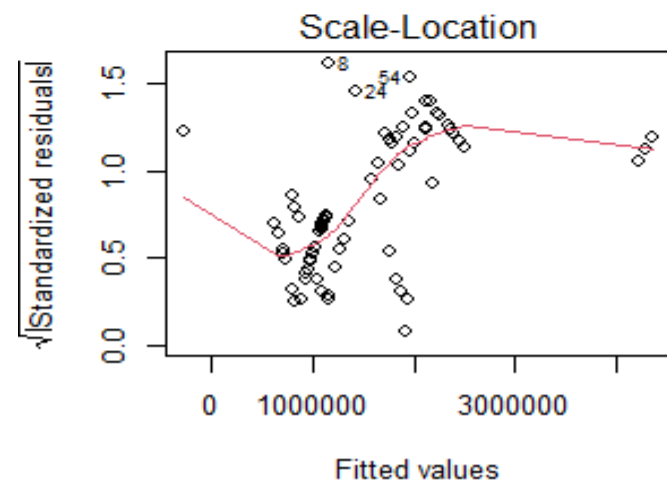
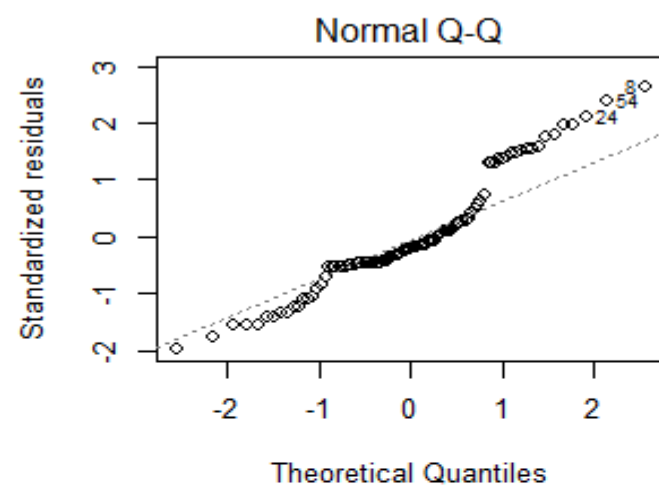
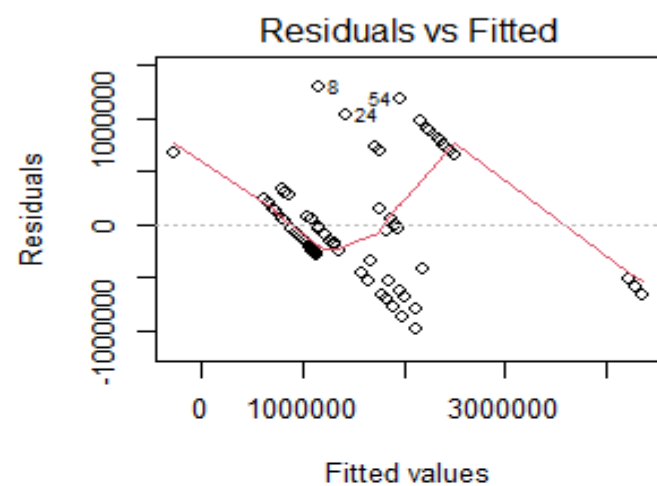
Residual standard error: 517100 on 87 degrees of freedom

Multiple R-squared: 0.7, Adjusted R-squared: 0.6828

F-statistic: 40.6 on 5 and 87 DF, p-value: < 2.2e-16

Diagnostics graphiques du modèle B

$\text{venu.moy.ha} \sim \text{Association} + \text{Charge.moy.ha} + \text{Formation} + \text{Nombre.d.anr}$



Meilleure sélection en backward avec le critère d'information d'Akaike (AIC) sur le modèle B :

```
Rcmdr> stepwise(modele_B, direction='backward', criterion='AIC')
```

```
Direction: backward
```

```
Criterion: AIC
```

```
Start: AIC=2452.83
```

```
Revenu.moy.ha ~ Association + Charge.moy.ha + Formation + Nombre.d.années +  
superficie..ha.
```

	Df	Sum of Sq	RSS	AIC
<none>			23266530257113	2452.8
- Nombre.d.années	1	741117582800	24007647839913	2453.7
- superficie..ha.	1	1410721914102	24677252171215	2456.3
- Association	1	1602852419908	24869382677021	2457.0
- Formation	1	1793564947176	25060095204289	2457.7
- Charge.moy.ha	1	52880388068788	76146918325901	2561.1

```
Call:
```

```
lm(formula = Revenu.moy.ha ~ Association + Charge.moy.ha + Formation +  
Nombre.d.années + superficie..ha., data = culture)
```

```
Coefficients:
```

(Intercept)	Association	Charge.moy.ha	Formation
-572328.979	529163.580	6.986	344598.267
Nombre.d.années	superficie..ha.		
-19338.825	-411352.910		

Variables explicatives retenues avec la sélection : Association, Charge moyenne, Formation, Nombre d'années, Superficie des terres

2. Estimation des paramètres des modèles par la méthode des moindres carrés ordinaires

D'après la sélection en Backward en se fiant au **critère d'Akaike (AIC)** pour chaque modèle, nous avons obtenu une estimation des paramètres du modèle par la **méthode des moindres carrés ordinaires** :

✓ Pour le **modèle A** :

Revenu.moy.ha ~ Association + Charge.moy.ha + Couverture.alimentaire +
Formation + Nombre.d.années + Rendement..Kg.ha. + Superficie..ha.

(Intercept)	Association	Charge.moy.ha
-438373.037870	522961.166734	6.865022
Couverture.alimentaire	Formation	Nombre.d.années
-27245.615905	320494.929725	-18185.296882
Rendement..Kg.ha.	Superficie..ha.	
-7.520125	-399942.601692	

✓ Pour le **modèle B** :

```
call:
lm(formula = Revenu.moy.ha ~ Association + Charge.moy.ha + Formation +
  Nombre.d.années + Superficie..ha., data = culture)
```

Coefficients:			
(Intercept)	Association	Charge.moy.ha	Formation
-572328.979	529163.580	6.986	344598.267
Nombre.d.années	Superficie..ha.		
-19338.825	-411352.910		

✓ Pour le modèle A :

$$\begin{aligned}\text{Revenu.moy.ha} = & - 438373.037870 + 522961.166734 \times \text{Association} + 6.865022 \times \text{Charge.moy.ha} \\ & - 27245.615905 \times \text{Couverture alimentaire} + 320494.929725 \times \text{Formation} \\ & - 18185.296882 \times \text{Nombre.d.années} - 7.520125 \times \text{Rendement..Kg.ha.} \\ & - 399942.601692 \times \text{Superficie..ha.}\end{aligned}$$

✓ Pour le modèle B:

$$\begin{aligned}\text{Revenu.moy.ha} = & - 438373.037870 + 522961.166734 \times \text{Association} + 6.865022 \times \text{Charge.moy.ha} \\ & + 320494.929725 \times \text{Formation} - 18185.296882 \times \text{Nombre.d.années} \\ & - 399942.601692 \times \text{Superficie..ha.}\end{aligned}$$

3. Signification des paramètres

Avant d'interpréter un coefficient (sens, magnitude de l'effet), il convient de s'assurer que celui-ci est significatif, autrement dit, qu'il est significativement différent de zéro (**H₀**, soit une absence d'effet). Pour cela on utilise un test de Student. On calcule la statistique t pour chaque variable assumée suivre une loi de Student ; que l'on compare ensuite avec la valeur théorique issue d'une table de Student (déterminée par le niveau du test, et la nombre d'observations).

On utilise souvent un niveau de 5% (soit un intervalle de confiance de 95%). Si la t -value calculée est supérieure (en valeur absolue) à la valeur théorique déterminée, alors on rejette **H₀** : Le coefficient est bien significativement différent de zéro, et on peut l'interpréter (signe, magnitude,..). Un autre moyen de réaliser le test est de regarder la p -value associée au coefficient, soit la probabilité pour que la valeur t -calculée soit supérieur en valeur absolue à la valeur théorique. Si cette probabilité est inférieur au seuil utilisé (ici 5%), alors le coefficient est significatif.

✓ Pour le modèle A

Avec le test de Student, on interprète la signification de chaque variable sur l'image suivante.

Call:

```
lm(formula = Revenu.moy.ha ~ Association + Charge.moy.ha + Couverture.alimentaire +  
  Formation + Nombre.d.années + Rendement..Kg.ha. + Superficie..ha.,  
  data = culture)
```

Residuals:

Min	1Q	Median	3Q	Max
-965345	-268020	-124040	223325	1411871

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-438373.0379	364270.7570	-1.203	0.2322	
Association	522961.1667	217725.8031	2.402	0.0185	*
Charge.moy.ha	6.8650	0.5191	13.224	<2e-16	***
Couverture.alimentaire	-27245.6159	116357.5825	-0.234	0.8154	
Formation	320494.9297	141954.0328	2.258	0.0265	*
Nombre.d.années	-18185.2969	11769.6225	-1.545	0.1260	
Rendement..Kg.ha.	-7.5201	8.3192	-0.904	0.3686	
Superficie..ha.	-399942.6017	180765.6287	-2.212	0.0296	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 520700 on 85 degrees of freedom

Multiple R-squared: 0.7029, Adjusted R-squared: 0.6784

F-statistic: 28.73 on 7 and 85 DF, p-value: < 2.2e-16

Avec le test on constate que les coefficients Association, Charge.moy.ha, Formation, Superficie..ha sont significatifs

✓ Pour le modèle B

Avec le test de Student, on interprète la signification de chaque variable sur l'image suivante.

```
call:
lm(formula = Revenu.moy.ha ~ Association + Charge.moy.ha + Formation +
    Nombre.d.années + Superficie..ha., data = culture)
```

Residuals:

Min	1Q	Median	3Q	Max
-987252	-263730	-101154	201816	1291859

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-572328.9790	330012.5950	-1.734	0.0864	.
Association	529163.5800	216147.0745	2.448	0.0164	*
Charge.moy.ha	6.9863	0.4968	14.062	<2e-16	***
Formation	344598.2667	133064.0444	2.590	0.0113	*
Nombre.d.années	-19338.8254	11616.9660	-1.665	0.0996	.
Superficie..ha.	-411352.9104	179101.8420	-2.297	0.0240	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 517100 on 87 degrees of freedom
Multiple R-squared: 0.7, Adjusted R-squared: 0.6828
F-statistic: 40.6 on 5 and 87 DF, p-value: < 2.2e-16

Avec le test on constate que les coefficients Association, Charge.moy.ha, Formation, Nombre.d.années, Superficie..ha, sont significatifs

4. Validation du modèle

Enfin, on peut regarder la qualité de la régression (au regard des données), mesurée par le coefficient de détermination (R-Squared ou R^2), qui se définit comme la part de variation dans la variable y qui est expliquée par des variations dans les variables explicatives (souvent exprimé en %).

Plus sa valeur est proche de 1, et plus l'adéquation entre le modèle et les données observées va être forte. Cependant, cette valeur est fortement influencée, entre autres, par le nombre de variables explicatives incluses dans la régression. Le R^2 ajusté (Adjusted R-Squared) va alors tenir compte de ce nombre et sera donc plus correct.

Attention également, même si il n'existe aucune règle, ni aucune échelle précise qui indiquerait pour quelles valeurs du R^2 , la qualité doit être considérée "mauvaise" ou au contraire comme "excellente", une valeur trop élevée (R^2 ou R^2 ajusté supérieurs à 85%) peut cacher un grave problème (notamment d'endogénéité), et donc des résultats totalement erronés. De plus, il n'est bien souvent pas possible d'atteindre des valeurs jugées "satisfaisantes", en raison des données à disposition pour l'analyse ; et il n'est donc pas rare que l'économètre doive se contenter d'un R^2 de "seulement" 40% par exemple (voire 30%) ! Ici notre 70% est donc plus que convenable, et la recherche d'un R^2 élevé ne doit pas être un but en soi..

✓ Pour le modèle A

R-squared = 0.7029

R-squared ajusté = 0.6784

call:

```
lm(formula = Revenu.moy.ha ~ Association + Charge.moy.ha + Couverture.alimentaire +  
  Formation + Nombre.d.années + Rendement..Kg.ha. + Superficie..ha.,  
  data = culture)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-965345	-268020	-124040	223325	1411871

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-438373.0379	364270.7570	-1.203	0.2322	
Association	522961.1667	217725.8031	2.402	0.0185	*
Charge.moy.ha	6.8650	0.5191	13.224	<2e-16	***
Couverture.alimentaire	-27245.6159	116357.5325	-0.234	0.8154	
Formation	320494.9297	141954.0828	2.258	0.0265	*
Nombre.d.années	-18185.2969	11769.6225	-1.545	0.1260	
Rendement..Kg.ha.	-7.5201	8.3192	-0.904	0.3686	
superficie..ha.	-399942.6017	180765.6287	-2.212	0.0296	*

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 520700 on 85 degrees of freedom

Multiple R-squared: 0.7029, Adjusted R-squared: 0.6784

F-statistic: 28.73 on 7 and 85 DF, p-value: < 2.2e-16

✓ Pour le modèle B

R-squared = 0.7

R-squared ajusté = 0.6828

Call:

```
lm(formula = Revenu.moy.ha ~ Association + Charge.moy.ha + Formation +  
    Nombre.d.années + Superficie..ha., data = culture)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-987252	-263730	-101154	201816	1291859

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-572328.9790	330012.5950	-1.734	0.0864	.
Association	529163.5800	216147.0745	2.448	0.0164	*
Charge.moy.ha	6.9863	0.4968	14.062	<2e-16	***
Formation	344598.2667	133064.0444	2.590	0.0113	*
Nombre.d.années	-19338.8254	11616.9660	-1.665	0.0996	.
Superficie..ha.	-411352.9104	179101.8420	-2.297	0.0240	*

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 517100 on 87 degrees of freedom

Multiple R-squared: 0.7, Adjusted R-squared: 0.6828

F-statistic: 40.6 on 5 and 87 DF, p-value: < 2.2e-16

Validation du modèle : choix du modèle B

On valide le choix du modèle B car la qualité de ce dernier est plus bonne.

Prédictions et analyse de la variance

1. Predictions

La fonction `predict()` permet de prédire la valeur de y (variable réponse) pour de nouvelles données (des variables explicatives). Seul les deux premiers arguments sont requis : `se.fit` permet d'afficher l'écart-type de la valeur prédite, et `interval` et `level` permettent afficher ici les valeurs de l'intervalle de confiance fixé à 99%.

Nous avons prédit le revenu moyen de quatres cas suivant les images suivantes :

1^{ère} prédiction :

```
> predict(modele_B, newdata=data.frame(Association=1,Charge.moy.ha=465000, Formation=1, Nomb  
re.d.années=15, Superficie..ha.=1.6), se.fit=TRUE, interval = "prediction", level = 0.99)  
$fit  
      fit      lwr      upr  
1 2601826 1070259 4133394  
  
$se.fit  
[1] 266057.5  
  
$df  
[1] 87  
  
$residual.scale  
[1] 517137.7
```

2ème prédiction :

```
> predict(modele_B, newdata=data.frame(Association=0,Charge.moy.ha=450000, Formation=1, Nomb  
re.d.années=19, Superficie..ha.=2.3), se.fit=TRUE, interval = "prediction", level = 0.99)  
$fit  
      fit      lwr      upr  
1 1602566 55476.62 3149655  
  
$se.fit  
[1] 278705.2  
  
$df  
[1] 87  
  
$residual.scale  
[1] 517137.7
```

3ème prédiction :

```
> predict(modele_B, newdata=data.frame(Association=1,Charge.moy.ha=234300, Formation=0, Nomb  
re.d.années=11, Superficie..ha.=0.7), se.fit=TRUE, interval = "prediction", level = 0.99)  
$fit  
      fit      lwr      upr  
1 1093056 -310395 2496507  
  
$se.fit  
[1] 128720.8  
  
$df  
[1] 87  
  
$residual.scale  
[1] 517137.7
```

4ème prédiction :

```
> predict(modele_B, newdata=data.frame(Association=1,Charge.moy.ha=144300, Formation=1, Nomb  
re.d.années=13, Superficie..ha.=0.6), se.fit=TRUE, interval = "prediction", level = 0.99)  
$fit  
      fit      lwr      upr  
1 811342.9 -589857 2212543  
  
$se.fit  
[1] 125134.6  
  
$df  
[1] 87  
  
$residual.scale  
[1] 517137.7
```

2. Analyse de la variance

Avec l'analyse de la variance, nous avons obtenu :

```
Rcmdr> Anova(modele_B, type="II")  
Anova Table (Type II tests)
```

Response: Revenu.moy.ha

	Sum Sq	Df	F value	Pr(>F)	
Association	1602852419908	1	5.9935	0.01637	*
Charge.moy.ha	52880388068788	1	197.7344	< 2e-16	***
Formation	1793564947176	1	6.7066	0.01126	*
Nombre.d.années	741117582800	1	2.7712	0.09957	.
superficie..ha.	1410721914102	1	5.2751	0.02403	*
Residuals	23266530257113	87			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Conclusion et recommandations

Pour conclure nous pouvons dire que les méthodes et techniques utilisées dans l'agriculture s'avèrent importants d'après notre modèle. Nous voyons que le fait d'avoir une expérience avec la formation et les fonds déployées pour les cultures sont très importants pour avoir des revenus conséquents dans l'agriculture pratiquée. Le fait d'avoir des collaborateurs aussi à travers les associations est d'une grande aide pour augmenter ses bénéfices. Nous voyons aussi que le nombre d'années de pratiques et les superficies des terres n'ont pas une grande influence sur les revenus apportés aux villageois.

Pour les recommandations, les bienfaits à signaler se localisent autour de la formation à faire, les associations pour d'éventuels collaborateurs et les fonds déployés pour les cultures. Il est conseillé aux villageois de Guédé Village de prioriser les formations qui se font pour acquérir de nouvelles techniques et méthodes dans l'agriculture ce qui optimiserait la qualité de leurs cultures. Il est aussi bien de dépenser le maximum de fonds que l'on peut déployer pour augmenter ces revenus car plus on dépense plus on gagne.