Introduction à Rcommander

Pauline Scherdel

Septembre 2014

Table des matières

1 Introduction à Rcmdr sous R

Ce document constitue une présentation succincte du package Rcommander (Rcmdr), une sur-couche du logiciel R. Il s'agit d'une interface graphique qui facilite l'interactivité avec le logiciel R. En particulier, on s'intéressera à l'importation et la manipulation des données quantitatives et qualitatives, à la description d'un jeu de données et à l'analyse statistique.

L'interface du package `Rcmdr` est assez rudimentaire. Elle est composée d'un menu avec des listes déroulantes afin de remplacer les fonctions R à taper dans un script, d'une fenêtre "script R" avec les commandes R, d'une fenêtre "Sortie" relative aux résultats et d'une fenêtre "Message" relative aux messages d'erreur. Les commandes exécutées par le menu sont traduites en script R dans la fenêtre script. Il est possible de taper des commandes R directement dans cette fenêtre.

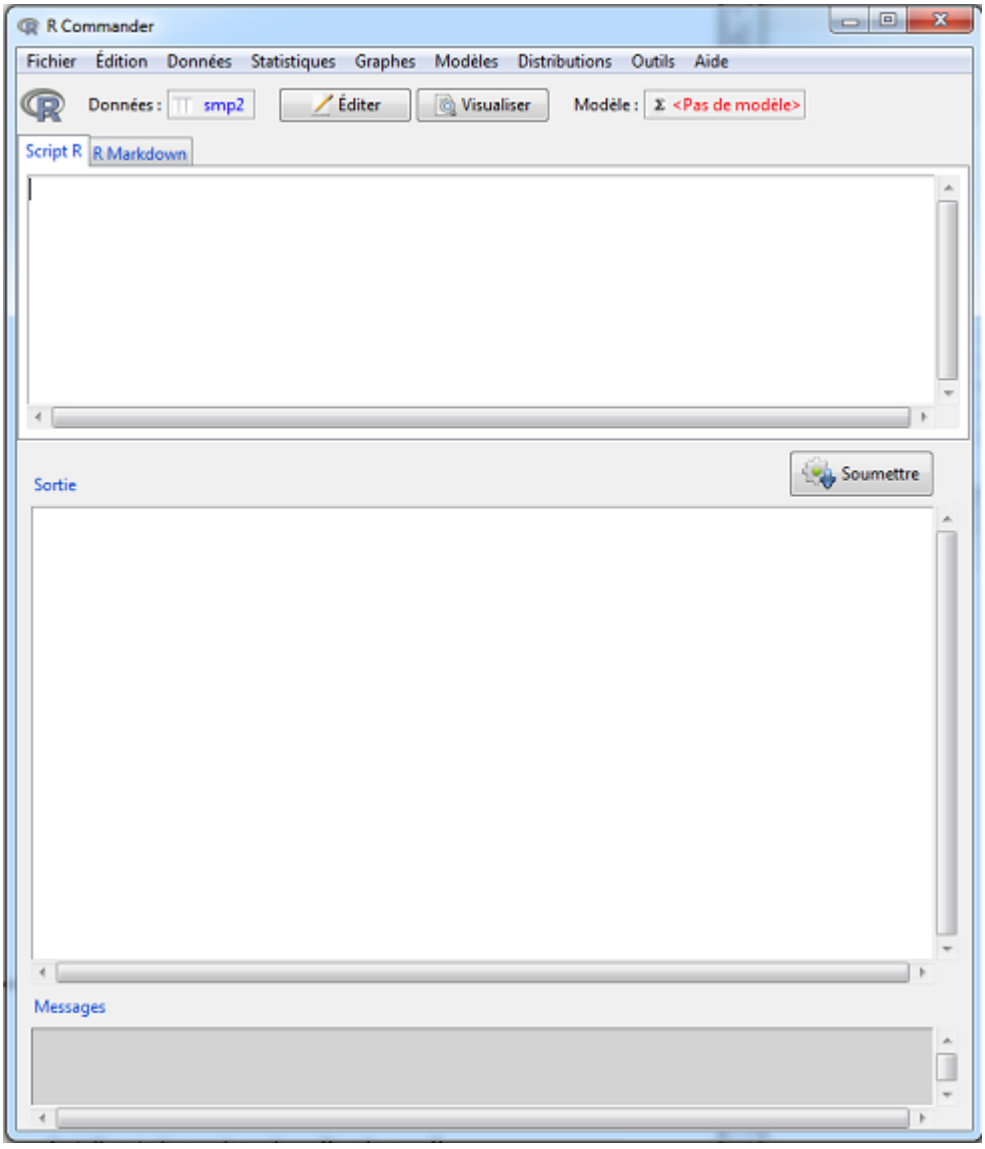


FIGURE 1 – Aperçu de l'interface de Rcmdr

2 Interagir avec R

Démarrer avec R. Quelque soit le système d'exploitation utilisé (Windows, Mac, Linux), R fonctionne comme tout autre logiciel : il suffit généralement de double-cliquer sur l'icône de l'application pour démarrer R. On dispose ensuite d'une console interactive dans laquelle on peut commencer à saisir des commandes après l'invite R `>`. Les résultats seront affichés aussitôt dans la console.

3 Installer et charger le package Rcmdr sous R

Il faut installer le package Rcmdr grâce à la commande `install.packages()` :

```
install.packages("Rcmdr")
```

On obtient la fenêtre suivante :

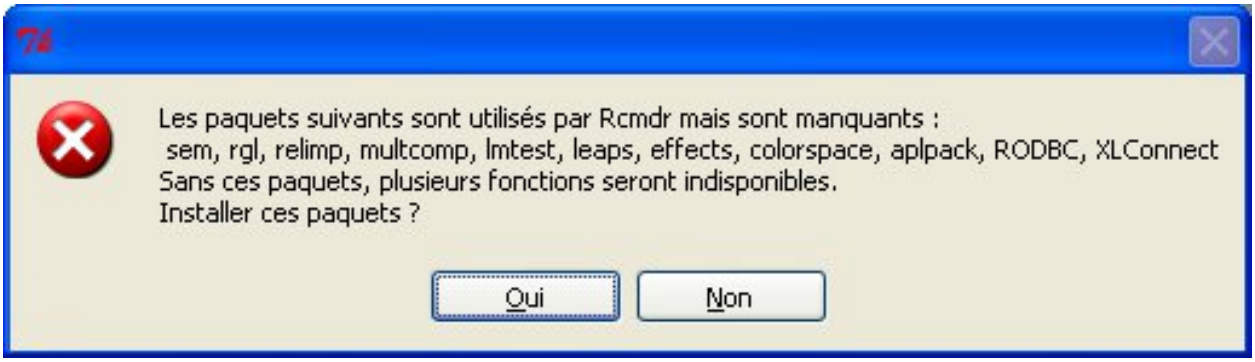


FIGURE 2 – Message lors de l'installation du package Rcmdr

Cliquer sur Oui pour installer les dépendances.

On installera aussi les packages `epicalc`, `epitools` et `prettyR` :

```
install.packages("epicalc")
install.packages("epitools")
install.packages("prettyR")
```

Les packages sont installés définitivement (tant qu'on ne les désinstalle pas).

Ensuite, il faut "charger les packages" à chaque session de R pour avoir accès aux fonctions qui les composent. On utilise pour cela la commande `library()` :

```
library(Rcmdr)
library(epicalc)
library(epitools)
library(prettyR)
```

4 Importation des données

Il est possible d'importer un jeu de données à partir d'un fichier Excel, Access, dBase ou texte mais également à partir d'autres formats comme SAS, SPSS ou STATA. Sous MAC, il est impossible d'importer un jeu de données à partir d'un fichier Excel, Access ou dBase.

4.1 Importation à partir d'un fichier Excel

Pour importer un jeu de données à partir d'un fichier Excel, Access ou dBase : **Données > Importer des données > Depuis un fichier Excel, Access ou dBase**

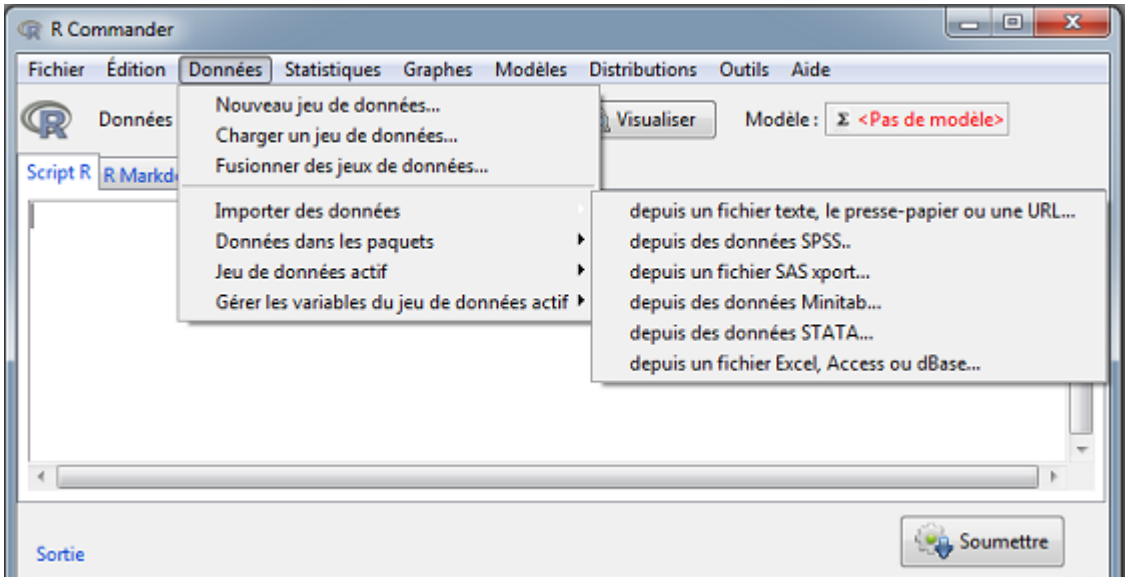


FIGURE 3 – Importation d'un fichier Excel - Etape 1

Il faut nommer le fichier qui vient d'être importé, par exemple en `smp2`, et parcourir vos documents pour chercher le jeu de données `smp2`.

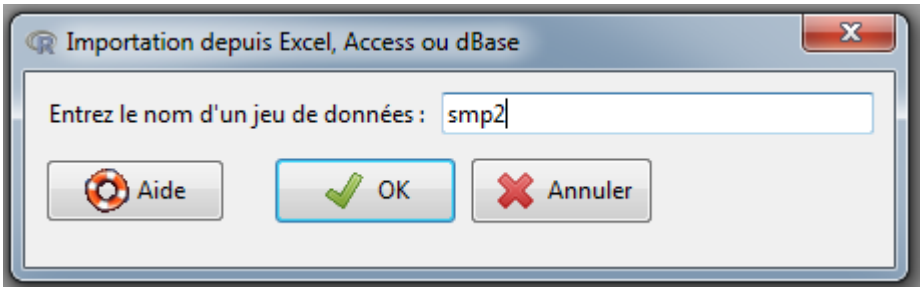


FIGURE 4 – Importation d'un fichier Excel - Etape 2

4.2 Importation à partir d'un fichier CSV

Pour importer un jeu de données à partir d'un fichier csv : **Données > Importer des données > Depuis un fichier texte, le presse-papiers ou URL**

Il faut nommer le fichier qui vient d'être importé, par exemple en `smp2`, et parcourir vos documents pour chercher le jeu de données `smp2`.

Après l'importation du jeu de données dans Rcmdr, il est important de vérifier le nombre d'observations et de variables, afin de savoir s'il est bien adéquat avec le fichier initial.

En cliquant sur `Visualiser`, il est possible d'apercevoir le jeu de données `smp2` :

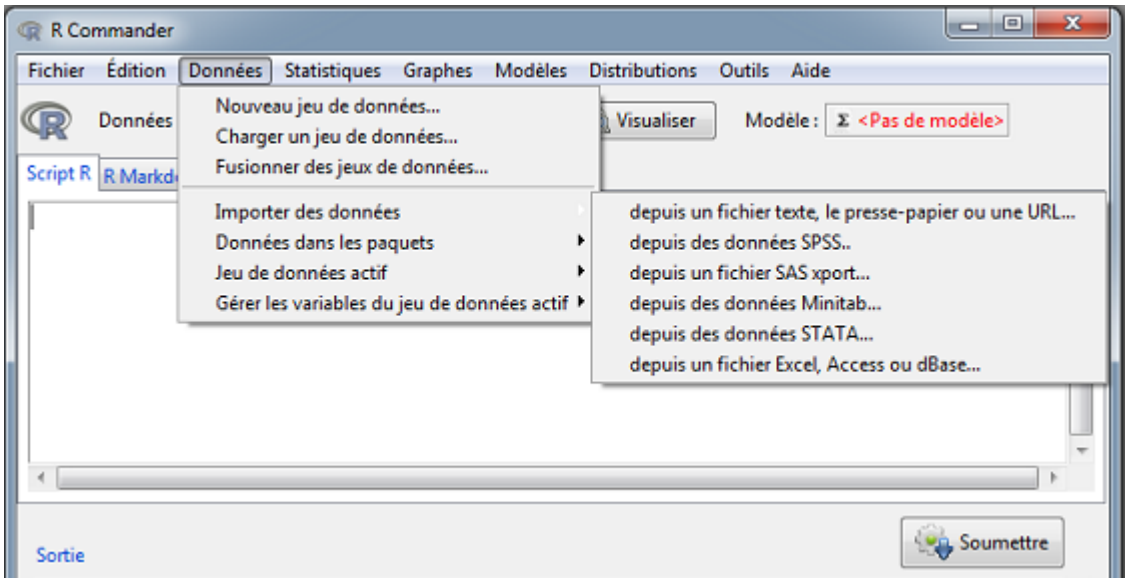


FIGURE 5 – Importation d'un fichier CSV - Etape 1

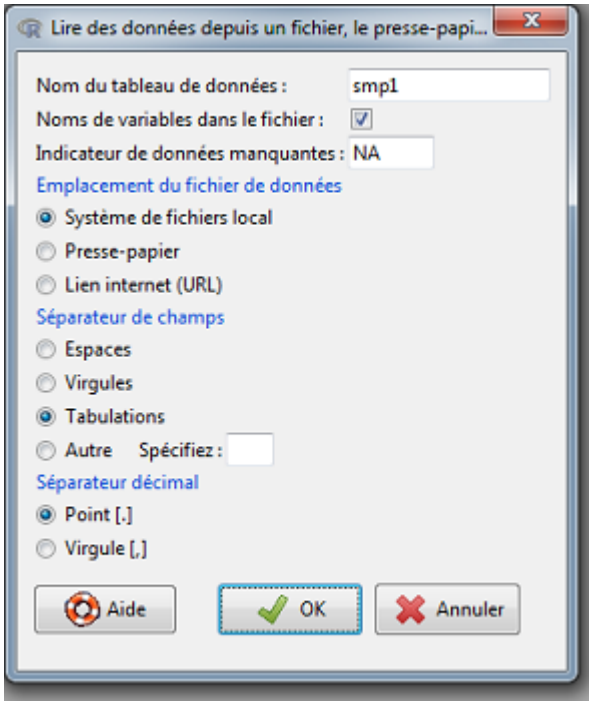


FIGURE 6 – Importation d'un fichier CSV - Etape 2

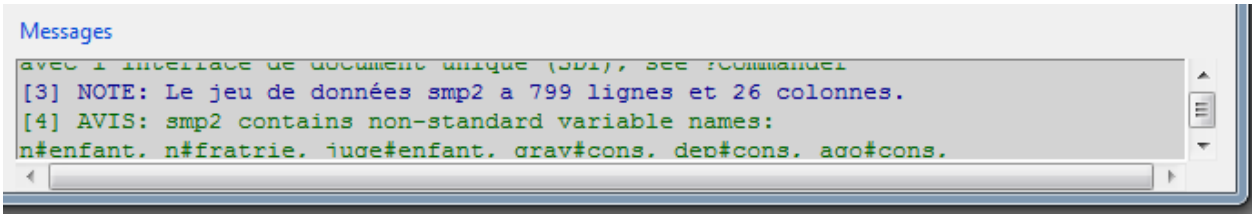
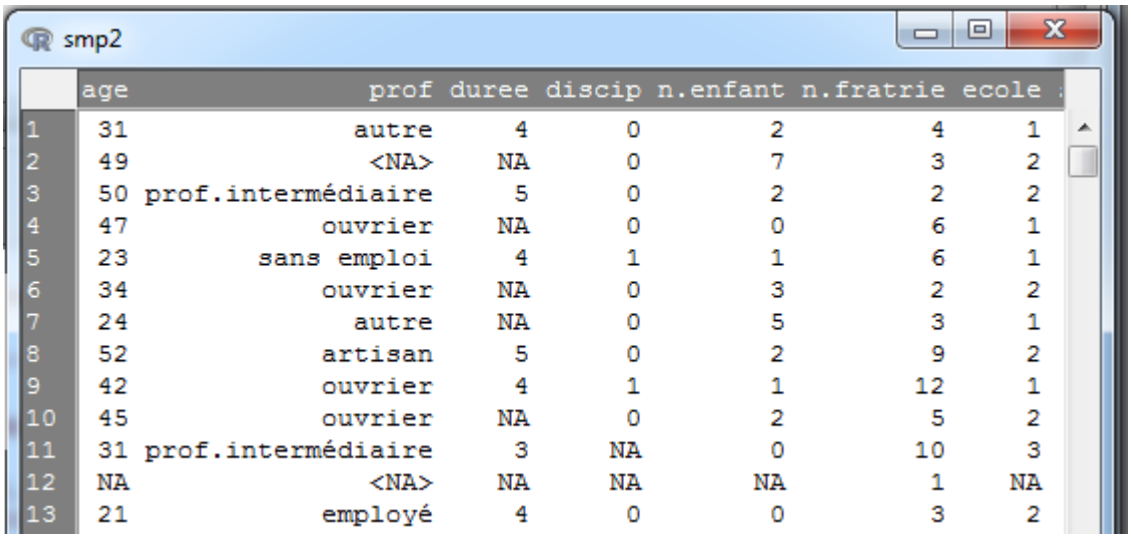


FIGURE 7 – Vérification du nombre d'observations et de variables



smp2

	age	prof	duree	discip	n.enfant	n.fratrie	ecole
1	31	autre	4	0	2	4	1
2	49	<NA>	NA	0	7	3	2
3	50	prof.intermédiaire	5	0	2	2	2
4	47	ouvrier	NA	0	0	6	1
5	23	sans emploi	4	1	1	6	1
6	34	ouvrier	NA	0	3	2	2
7	24	autre	NA	0	5	3	1
8	52	artisan	5	0	2	9	2
9	42	ouvrier	4	1	1	12	1
10	45	ouvrier	NA	0	2	5	2
11	31	prof.intermédiaire	3	NA	0	10	3
12	NA	<NA>	NA	NA	NA	1	NA
13	21	employé	4	0	0	3	2

FIGURE 8 – Visualisation des variables et observations de la table smp2

5 Manipulation des données

5.1 Visualisation brève au jeu de données

Une description brève du jeu de données importé peut être obtenue.

Statistiques > Résumés > Jeu de données actif

Pour chacune des variables du jeu de données, nous disposons d'indicateurs de positions (moyenne, médiane, quartiles). Attention, toutes les variables du jeu de données sont par défaut de type quantitatif. Nous verrons donc dans la partie suivante comment convertir ces variables en variables qualitatives.

5.2 Conversion des données quantitatives en qualitatives

L'ensemble des variables issues du jeu de données importé sont de type quantitatif par défaut. Avant d'analyser le jeu de données, il faut donc convertir les variables quantitatives, qui sont supposées être qualitatives, en variables qualitatives.

Données > Gérer les variables du jeu de données actifs > Convertir des variables numériques en facteurs

Par exemple, la variable "ecole" (niveau de formation actuel) est quantitative par défaut. Nous allons donc la convertir en variable qualitative en 5 classes.

La première possibilité est de transformer cette variable avec des modalités en chiffre "1", "2", "3", "4", "5" :

La seconde possibilité est de transformer cette variable avec des modalités en texte : "sans diplôme", "collège", "CAP, BEP", "Lycée", "université" :

5.3 Recodage des données quantitatives en qualitatives

Certaines analyses demandent de recoder des variables quantitatives en variables qualitatives, à 2 ou plusieurs catégories. Lors du recodage, il faut faire attention aux données manquantes.

Données > Gérer les variables du jeu de données actifs > Recoder des variables

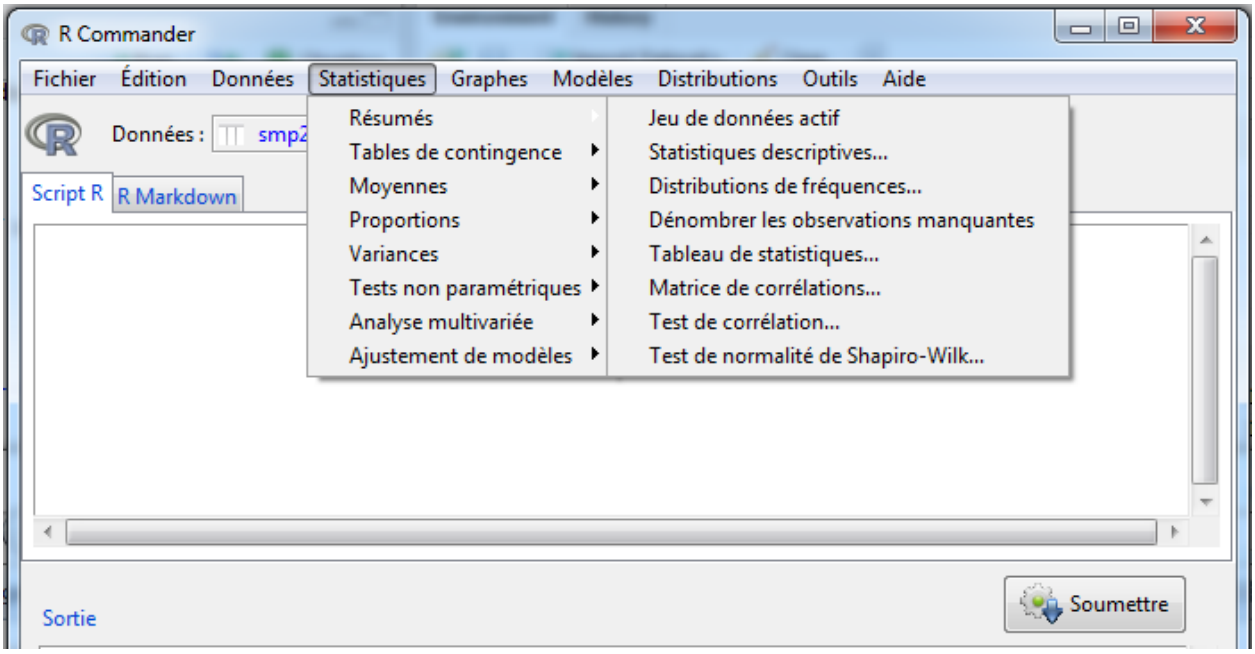


FIGURE 9 – Description de l'ensemble des variables du jeu de données smp2

Dans cet exemple, nous allons recoder la variable quantitative durée d'interview (dur.interv) en variable qualitative à 3 classes (dur.interv_c2). Si dur.interv est compris entre 0 et 60 minutes alors dur.interv_c2="Duree 60-", si dur.interv est vide alors dur.interv_c2=NA, sinon dur.interv_c2 = "Duree 60+".

5.4 Création de nouvelles variables

De nouvelles variables peuvent être créées à partir d'autres variables à l'aide de fonctions mathématiques : des opérateurs (+, -, *, /, ^...) ou des fonctions (log, exp, sin, cos, tan...).

Données > Gérer les variables du jeu de données actifs > Calculer une nouvelle variable

Pour exemple, nous allons créer la variable log(duree), qui représente le logarithme de la durée d'interview :

6 Description des données

6.1 Distribution des variables quantitatives et qualitatives

Dans une étude, il est important de décrire les variables de son jeu de données.

Pour les variables quantitatives, il est intéressant d'obtenir des moyennes, écart-types, médiane...

Statistiques > Résumés > Statistiques descriptives

En moyenne, la durée d'interview est de 23.99 minutes (+/- 10 écart-types). La médiane de la durée d'interview est de 25 minutes, c'est à dire que la moitié de la population a une durée d'interview de 25 minutes.

Pour les variables qualitatives, il est intéressant d'obtenir des proportions et des intervalles de confiance.

Statistiques > Résumés > Distribution de fréquence

La proportion de détenus ayant subi des maltraitances pendant l'enfance est de 27,78% (220).

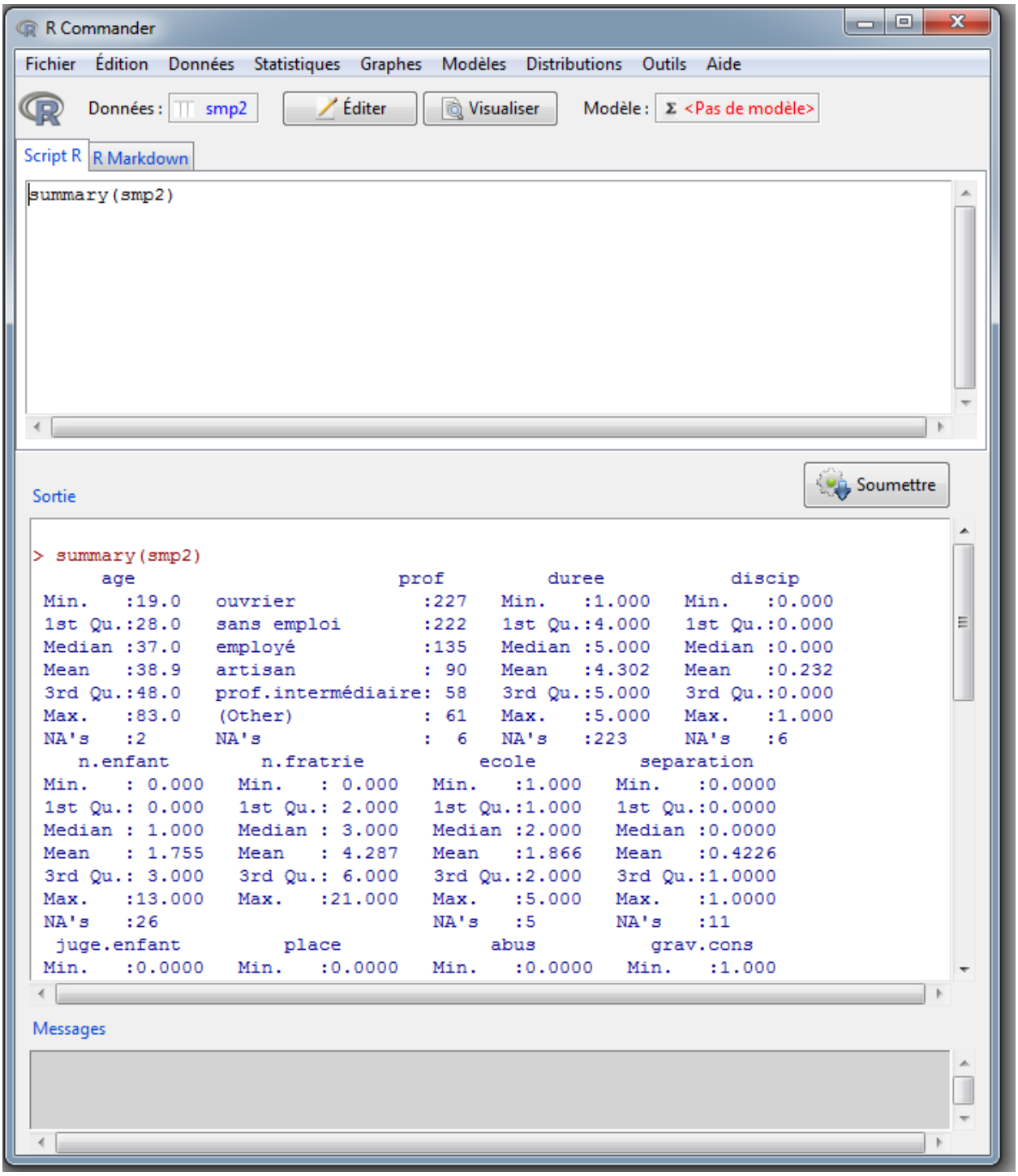


FIGURE 10 – Résultats : description du jeu de données smp2

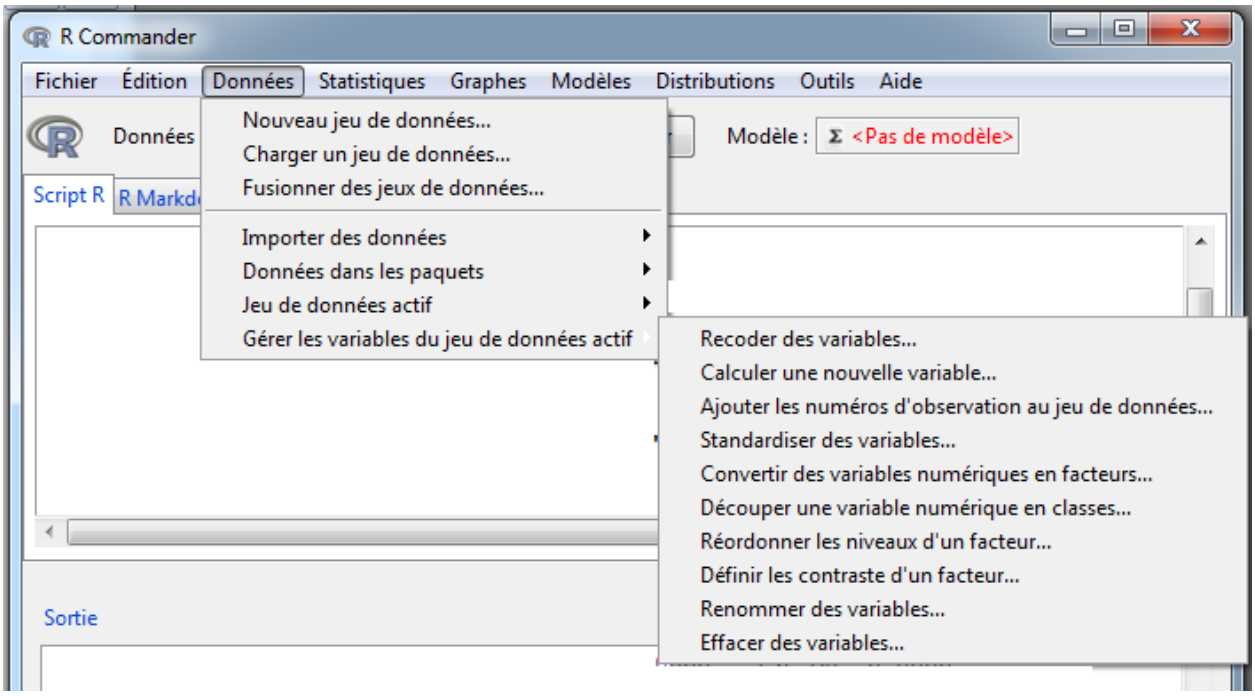


FIGURE 11 – Conversion de variables quantitatives en facteurs

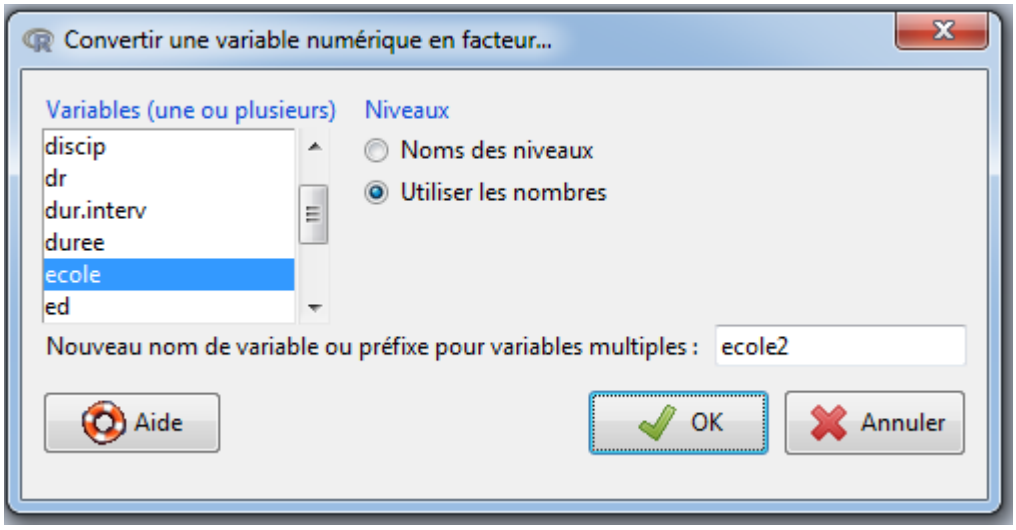


FIGURE 12 – Conversion de la variable ecole en facteurs - Etape 1

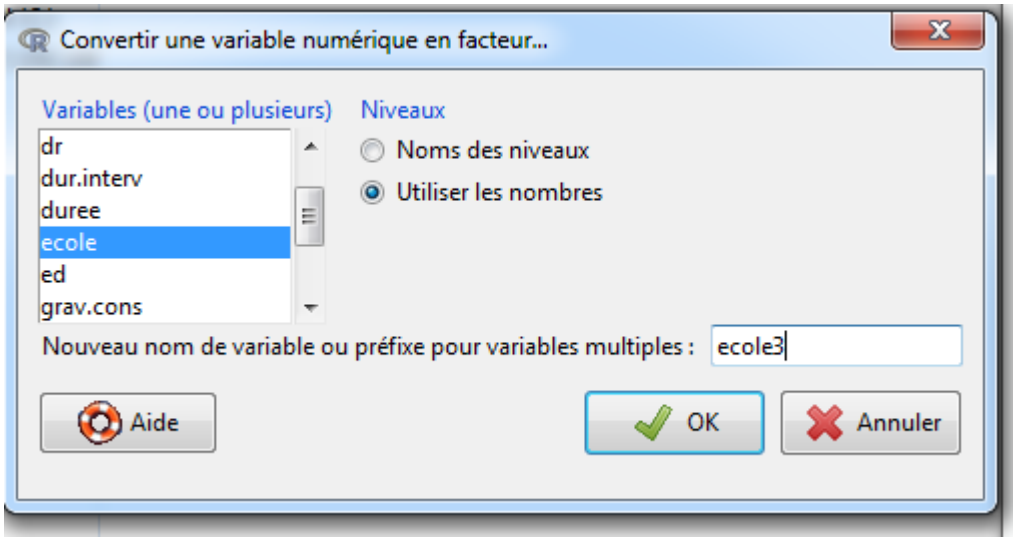


FIGURE 13 – Conversion de la variable ecole en facteurs - Etape 2

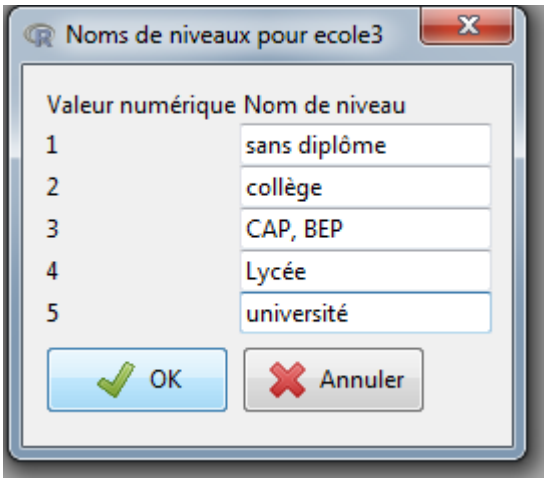


FIGURE 14 – Conversion de la variable ecole en facteurs - Etape 3

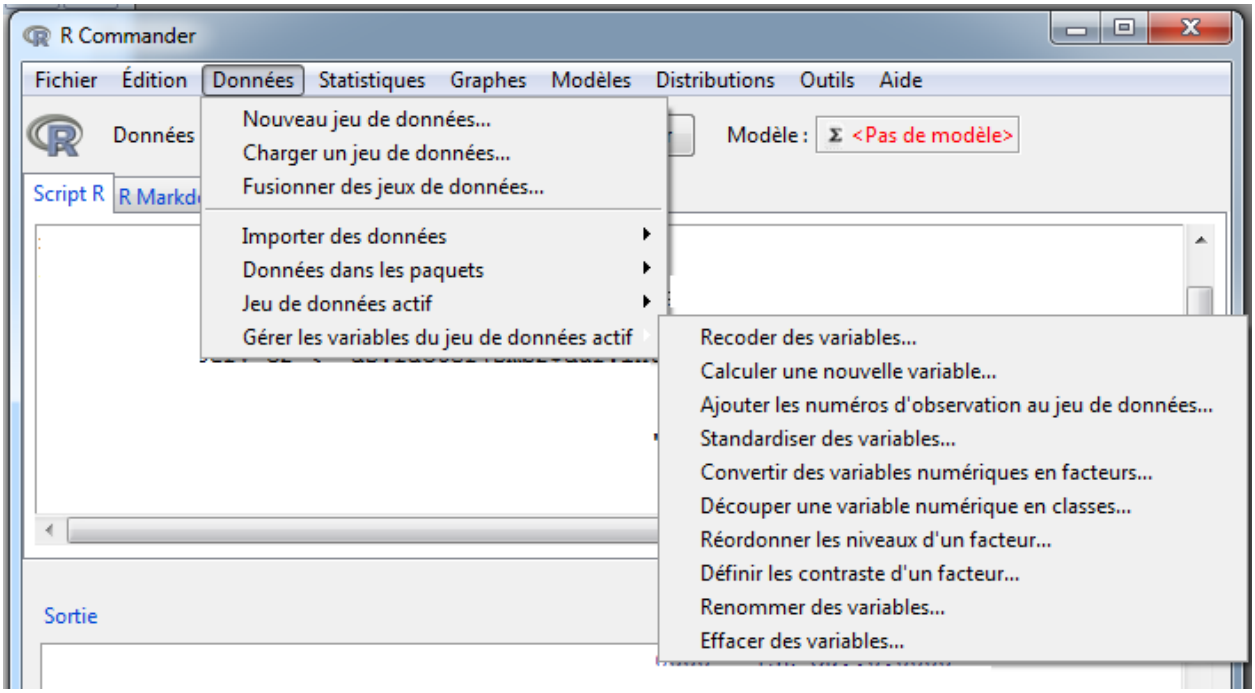


FIGURE 15 – Recodage des variables

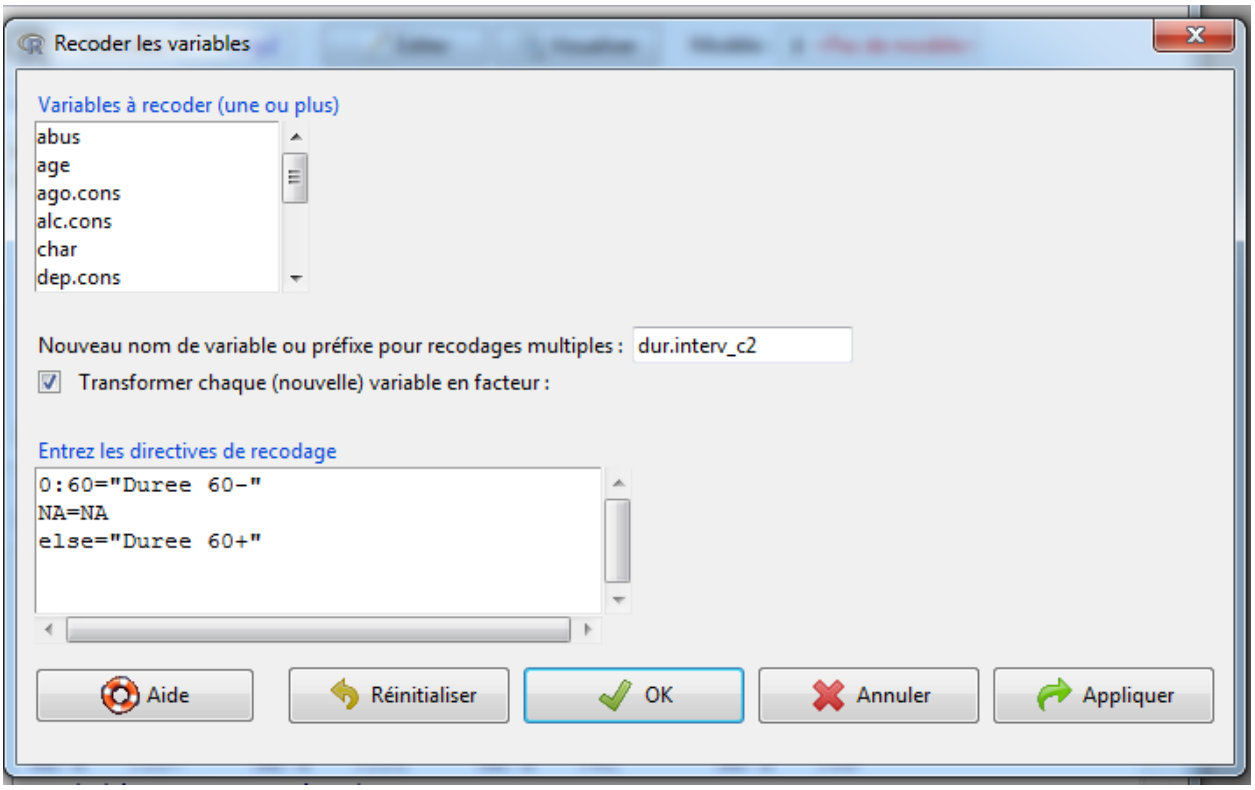


FIGURE 16 – Résultats : recodage de la variable dur.interv

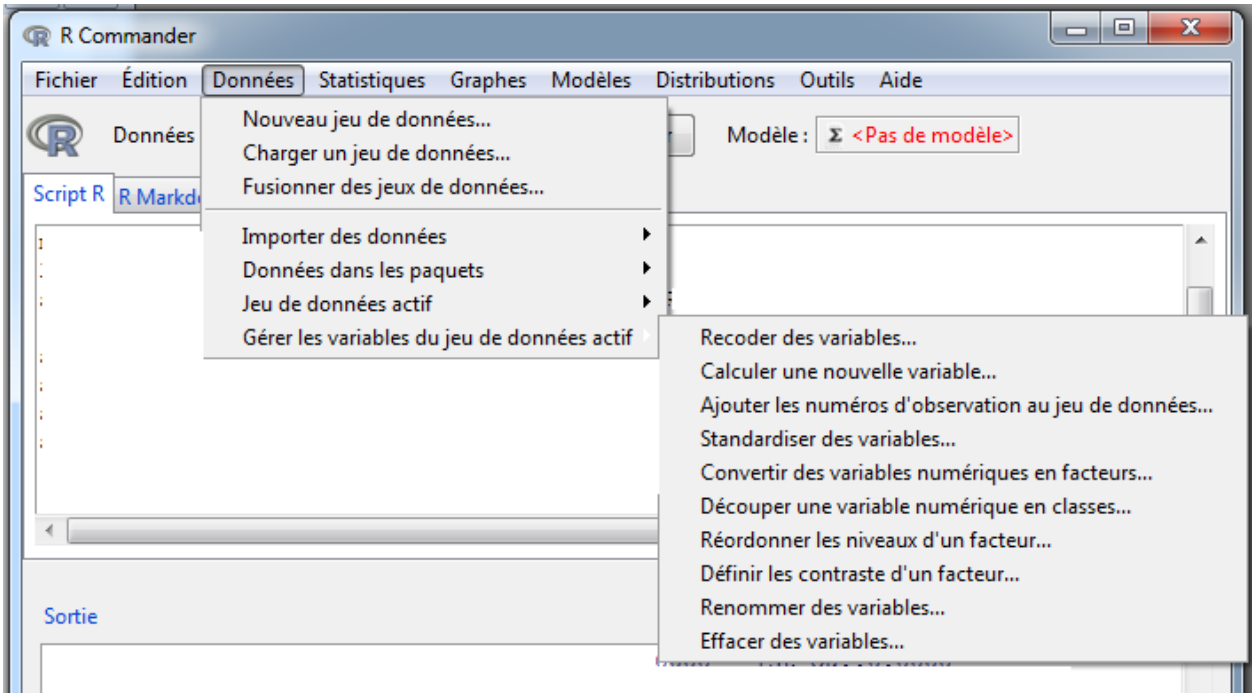


FIGURE 17 – Création de variables

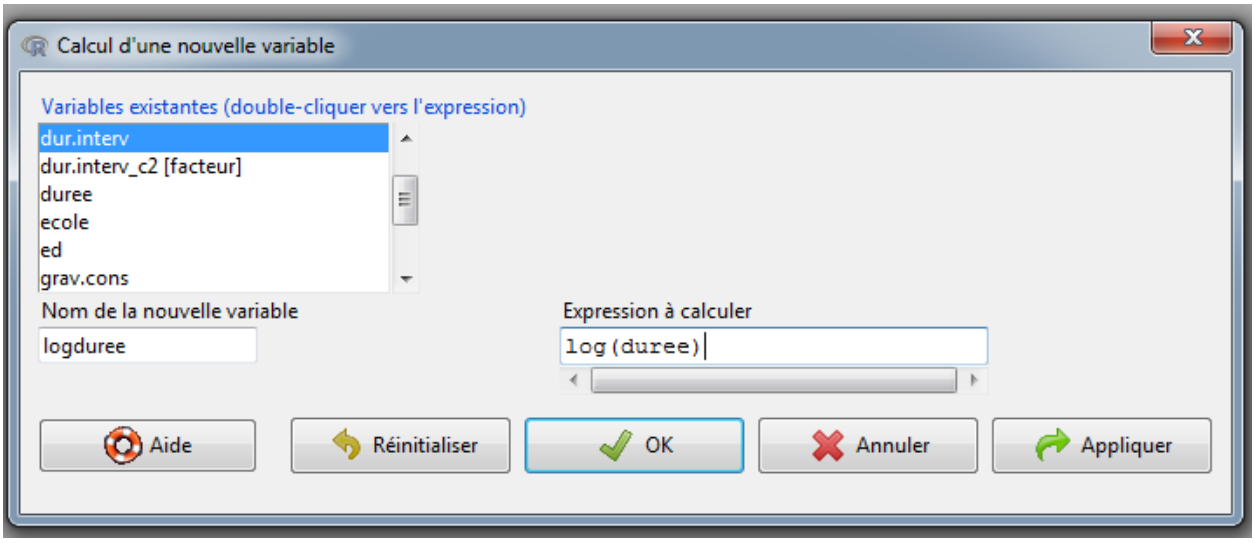


FIGURE 18 – Résultats : création de la variable log(duree)

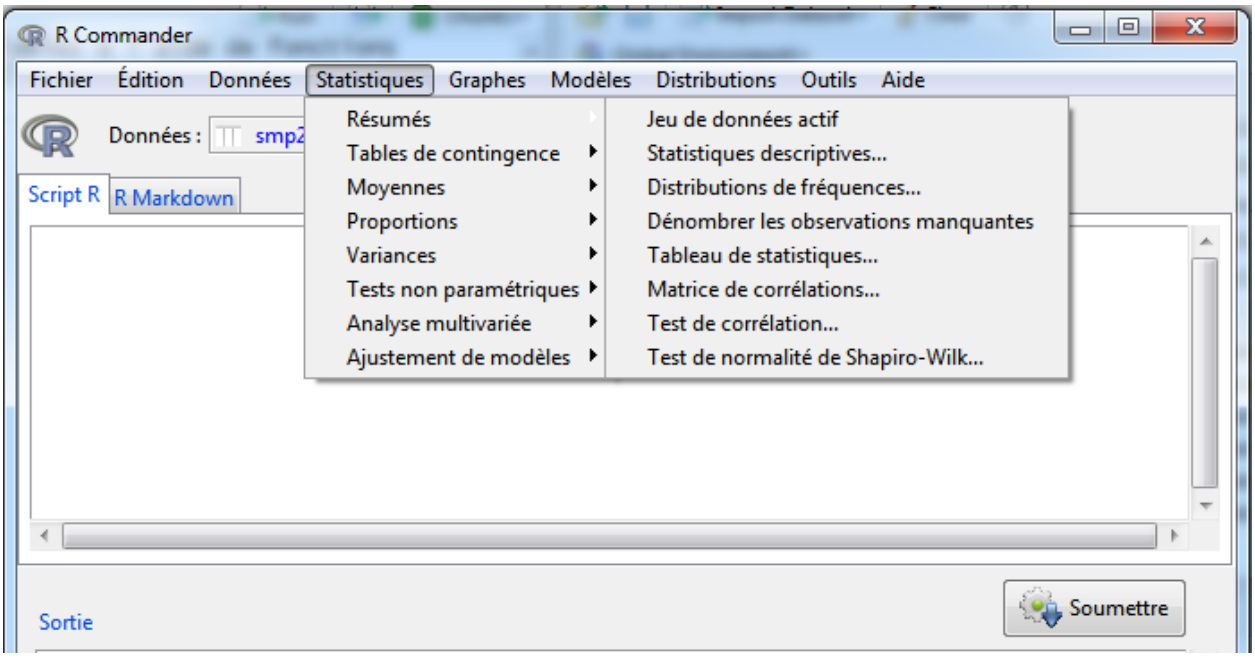


FIGURE 19 – Description des variables quantitatives du jeu de données smp2 - Etape 1

Pour obtenir l'intervalle de confiance d'une proportion, il n'existe pas de commande sous Rcmdr. Il faut taper et soumettre la procédure suivante dans le script : `prop.test(n,t)` où n est le nombre de cas et t est le nombre total d'individus concernés par la variable testée.

... avec un IC95% [24,71% - 31.06%]

6.2 Représentation des variables quantitatives et qualitatives

La distribution des variables quantitatives va être illustrée par des histogrammes ou des boxplots et celle des variables qualitatives par des diagrammes.

Graphes > Histogramme

Par exemple, la distribution de la durée d'interview (dur.interv) est représentée par un histogramme. Dans les options, il est possible de choisir quel type de données nous intéresse (effectifs, pourcentages, densités), de renommer les libellés des axes et de donner un titre au graphique.

Graphes > Boite de dispersion

La distribution de la durée d'interview peut être représentée également par un boxplot.

Graphes > Graphes en barres

Nous allons représenter la variable abus (abus) par un diagramme en barres. Il est possible de renommer les libellés des axes et de donner un titre au graphique.

Sur cet exemple, nous constatons que les détenus ayant subi des maltraitances pendant l'enfance sont moins nombreux.

Pour représenter le diagramme en barre en pourcentage, il faut préalablement créer une variable abus en pourcentage.

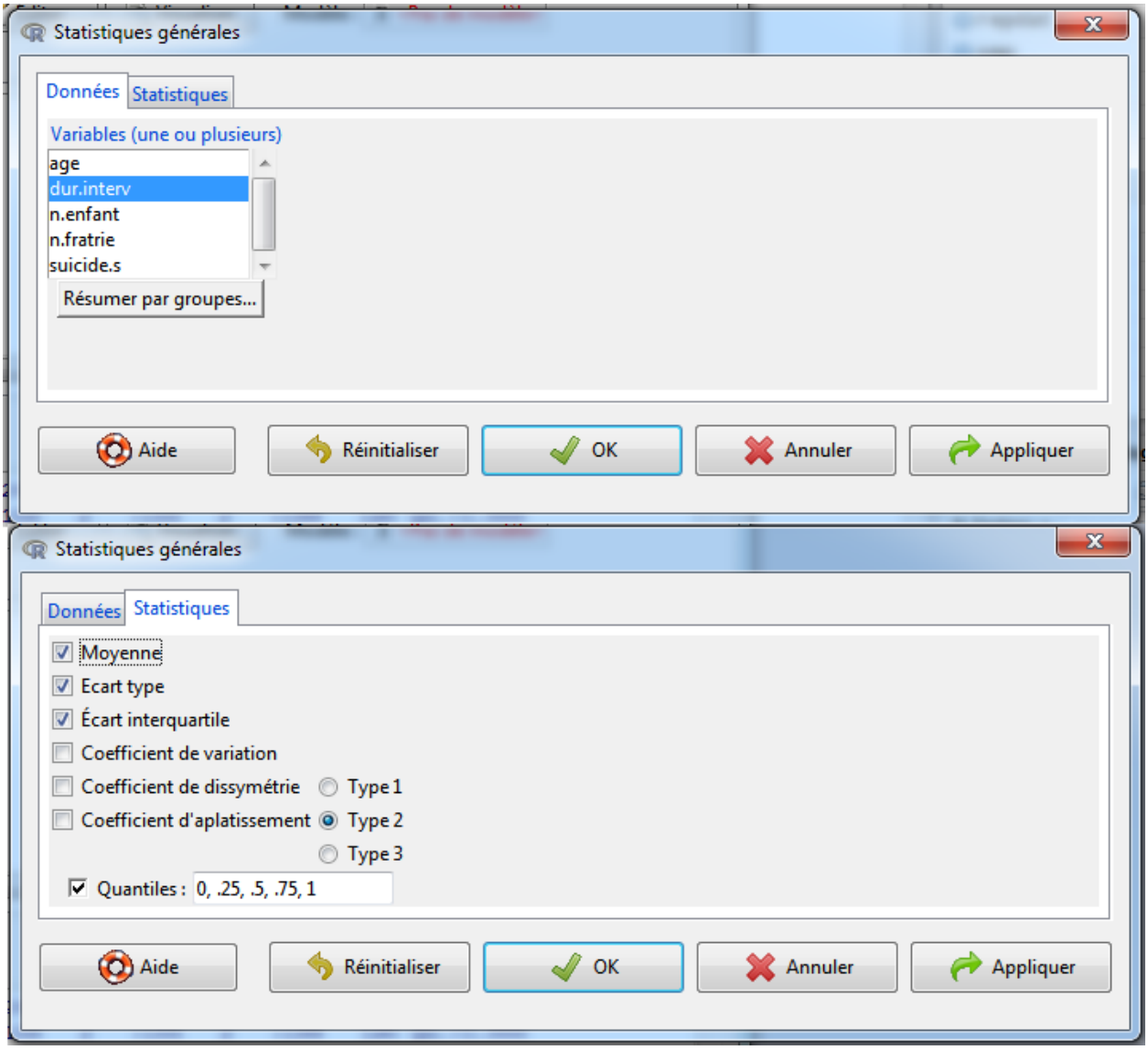


FIGURE 20 – Description des variables quantitatives du jeu de données smp2 - Etape 2

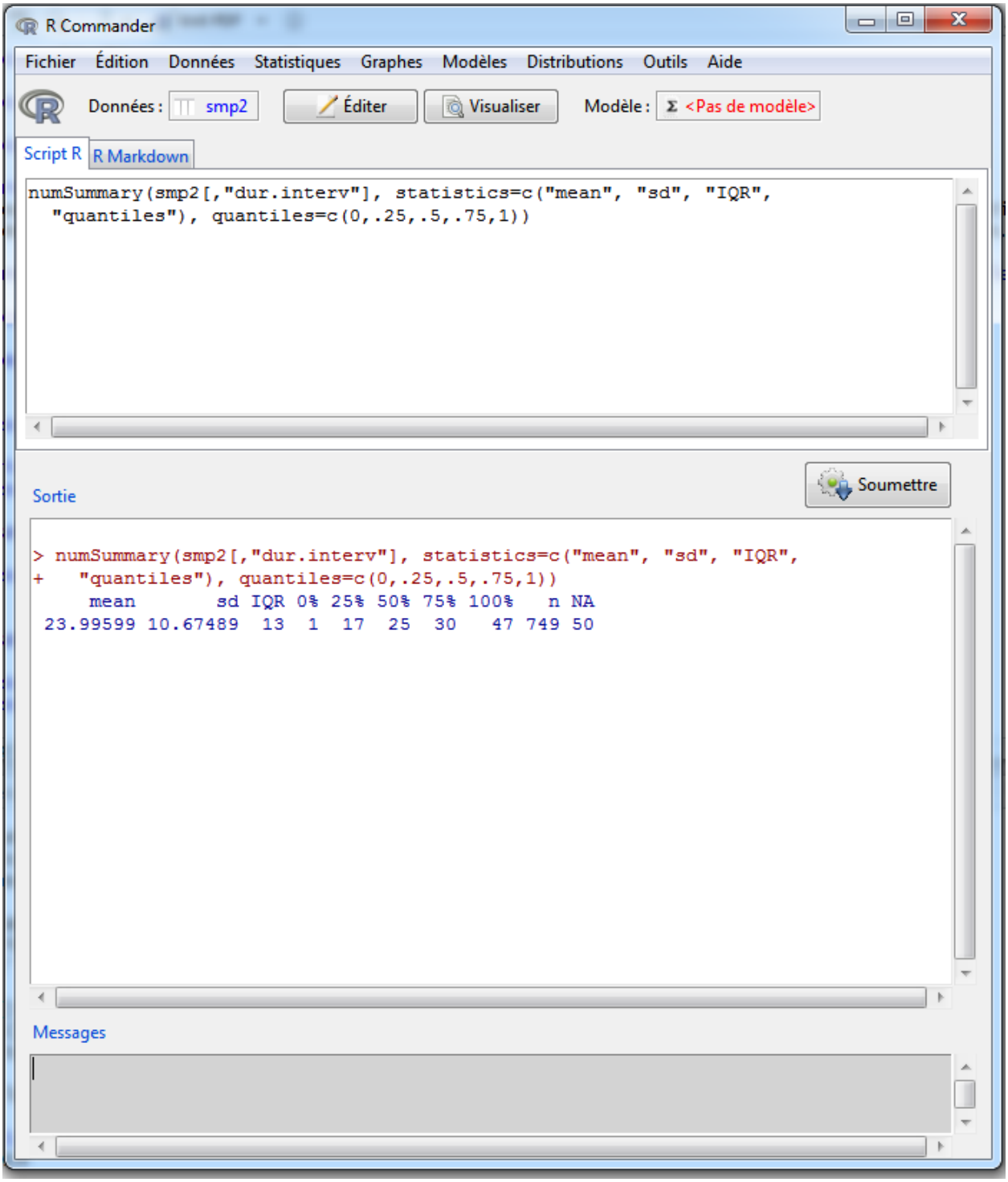


FIGURE 21 – Résultats : description de la variable dur.interv issue du jeu de données smp2

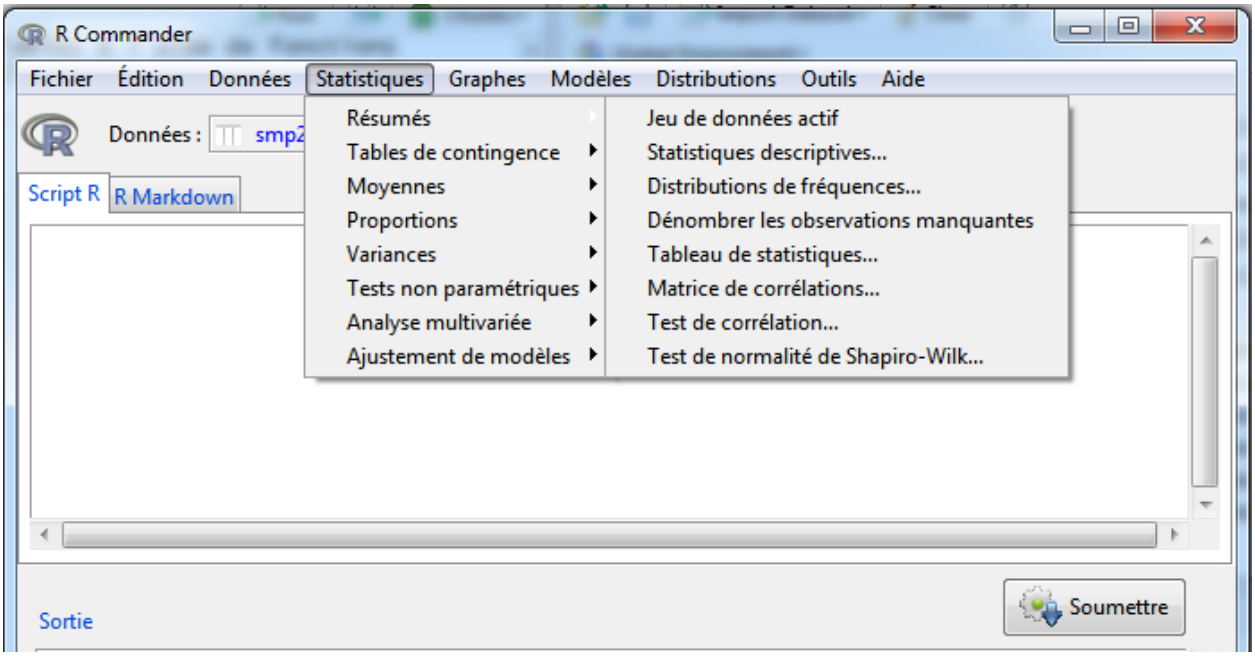


FIGURE 22 – Description des variables qualitatives du jeu de données smp2 - Etape 1

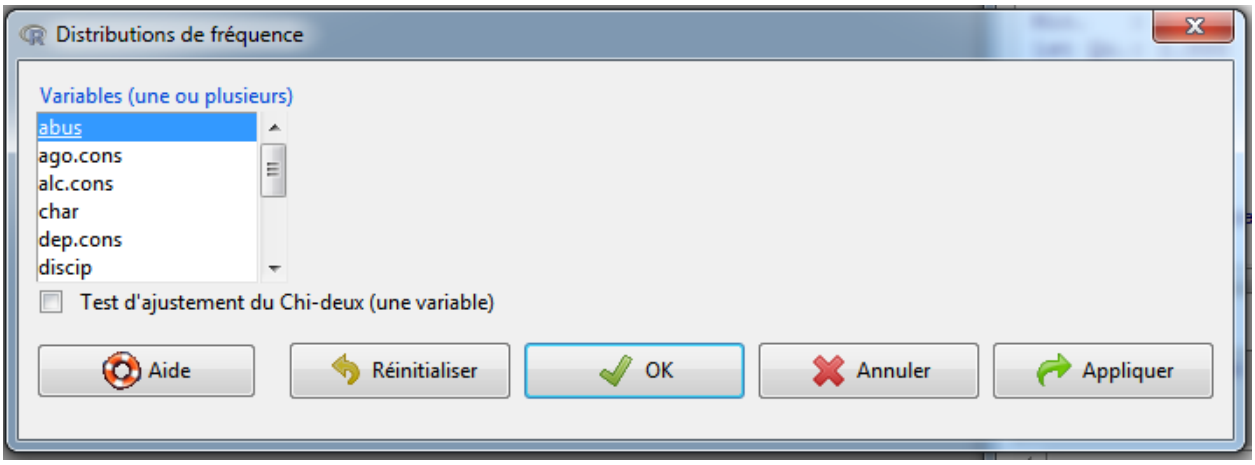


FIGURE 23 – Description des variables qualitatives du jeu de données smp2 - Etape 2

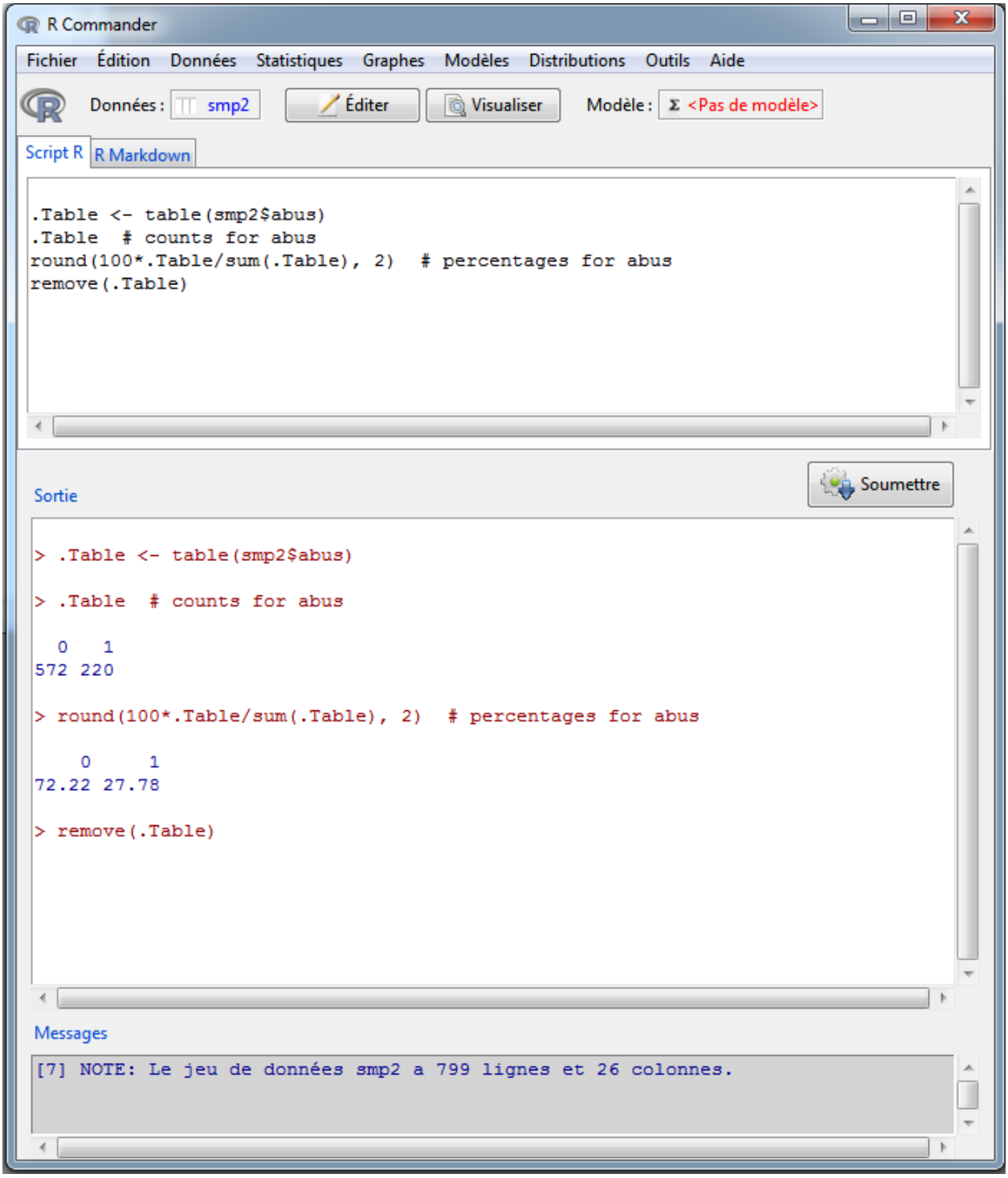


FIGURE 24 – Résultats : description de la variables abus issue du jeu de données smp2 (proportion)

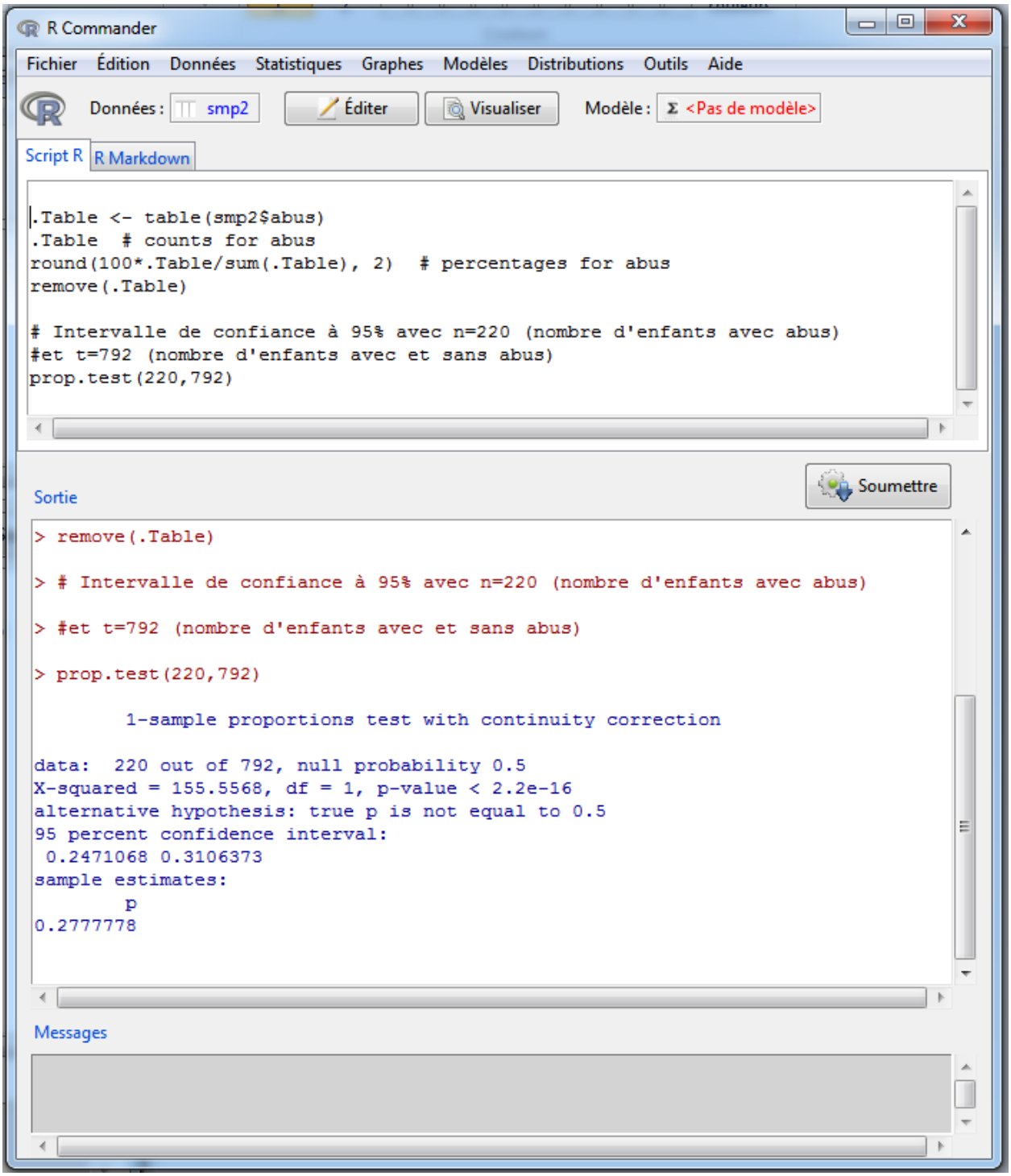


FIGURE 25 – Résultats : description de la variable abus issue du jeu de données smp2 (IC)

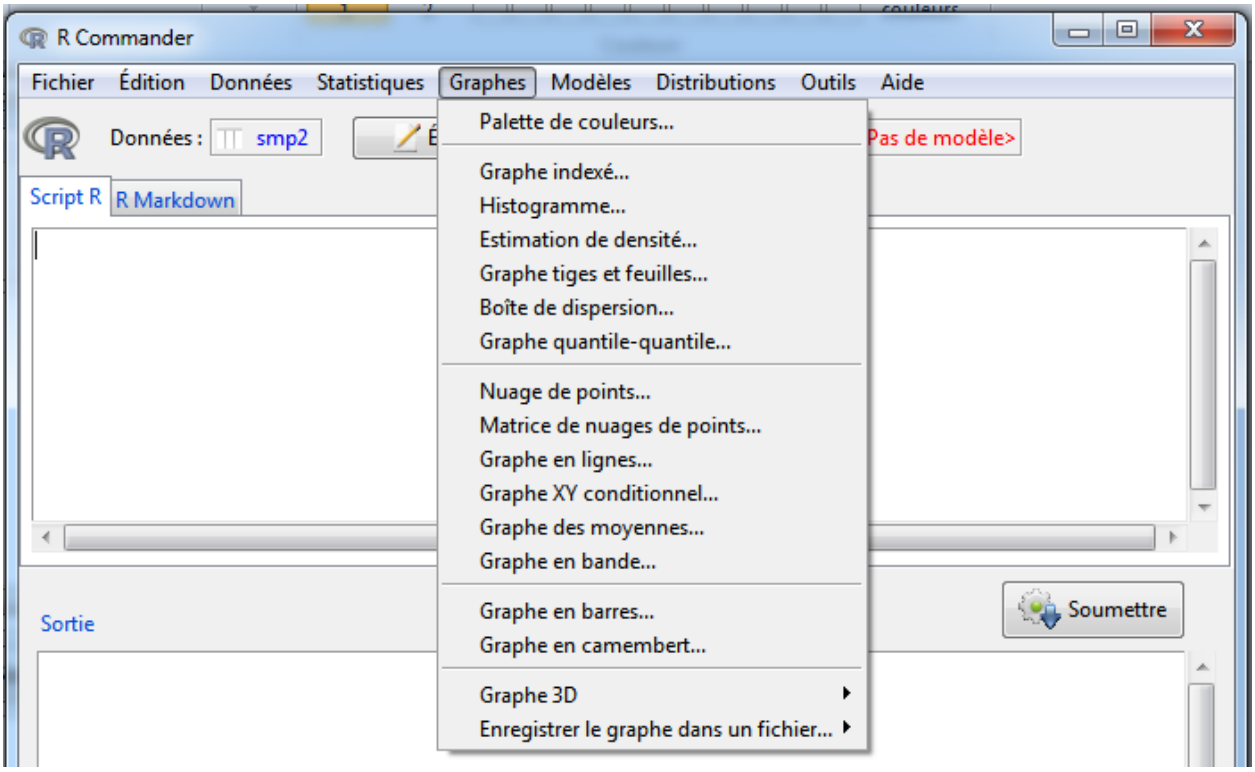


FIGURE 26 – Réalisation de graphiques de différents types de variables

6.3 Représentation de variables quantitatives en fonction d'une variable qualitative

Nous pouvons illustrer par un boxplot si le fait d'avoir subi des maltraitances pendant l'enfance pouvait entraîner des écarts dans la durée de l'interview.

Graphes > Boite de dispersion

La médiane entre les deux groupes semble identique.

7 Tests statistiques

7.1 Comparaison de moyennes d'une variable quantitative entre deux groupes

Nous allons comparer statistiquement la différence de durée d'interview entre les détenus ayant subi des maltraitances pendant l'enfance et ceux n'ayant jamais subi de maltraitance pendant l'enfance.

1) Effectifs par groupe (n1 et n2)

Nous allons regarder la moyenne de durée d'interview entre les deux groupes ainsi que les effectifs par groupe

Statistiques > Résumés > Statistiques descriptives

Les effectifs des deux groupes sont supérieurs à 30. Nous montrons que la médiane dans les deux groupes est de 25 minutes, et avec une moyenne de 23.34 minutes (+/- 10.7) dans le groupe "non abus" et de 25.75 minutes (+/- 10.2) dans le groupe "abus".

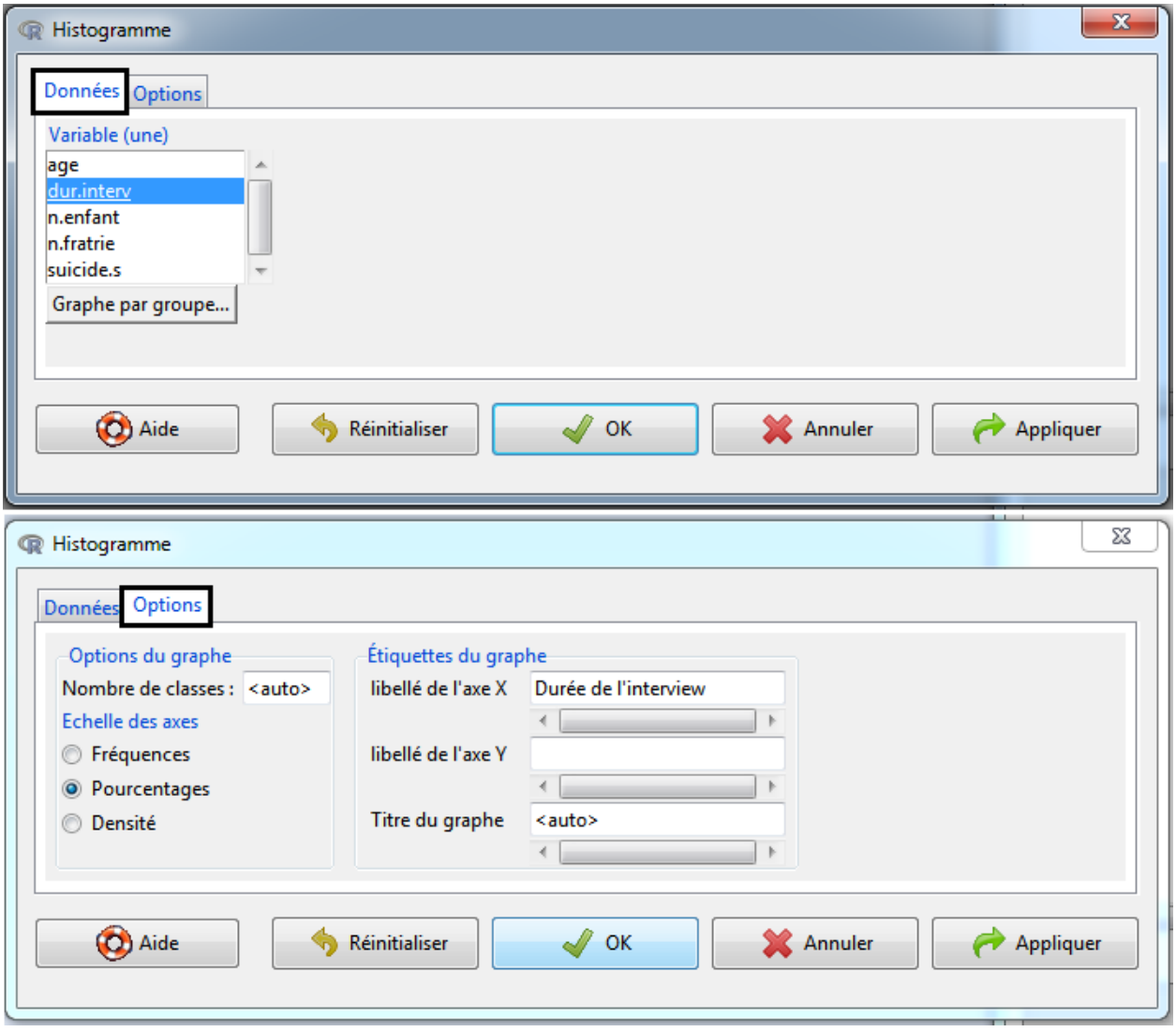


FIGURE 27 – Réalisation d'un histogramme d'une variable quantitative

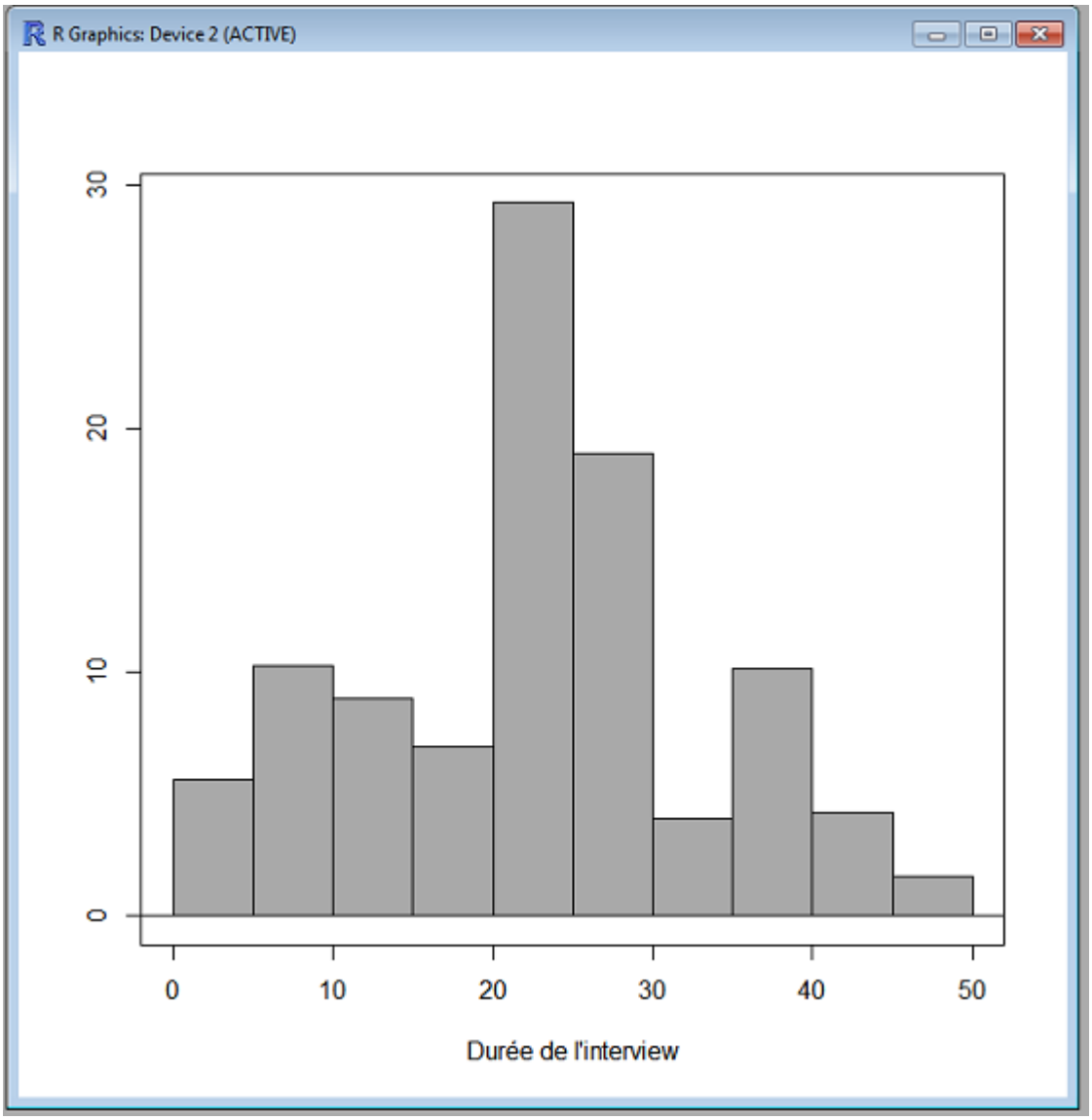


FIGURE 28 – Résultats : histogramme de la variable dur.interv

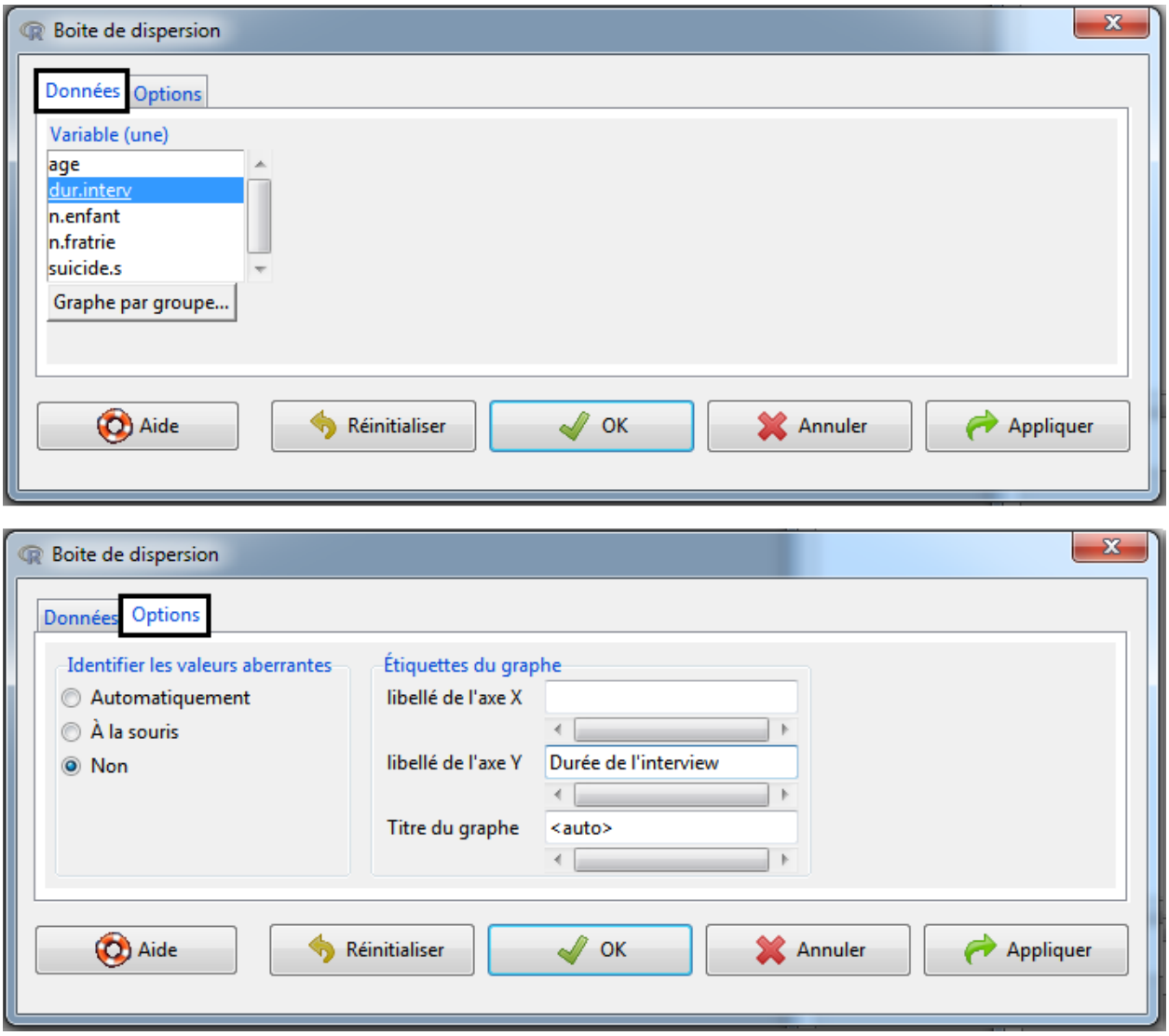


FIGURE 29 – Réalisation d'un boxplot d'une variable quantitative

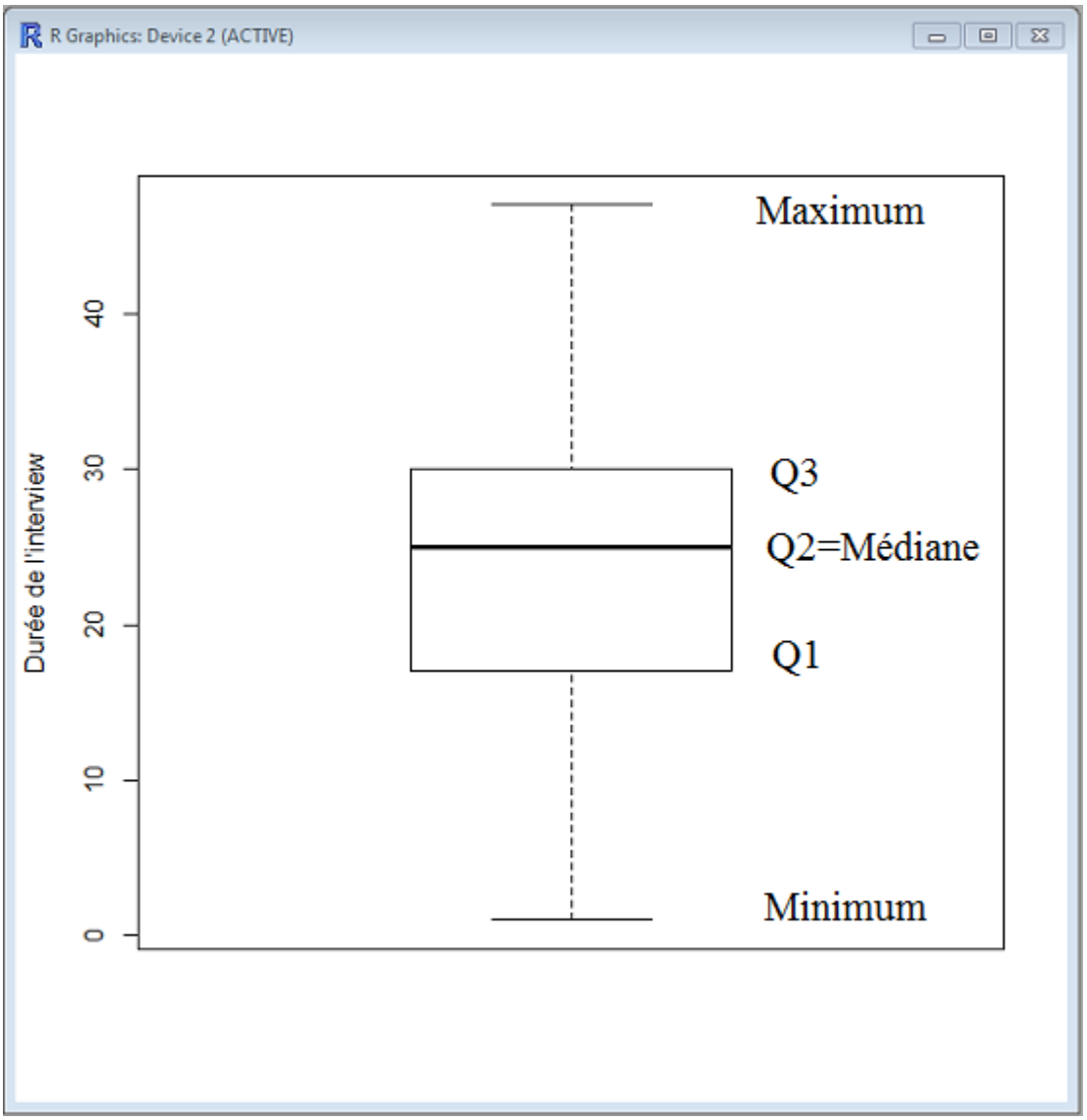


FIGURE 30 – Résultats : boxplot de la variable dur.interv

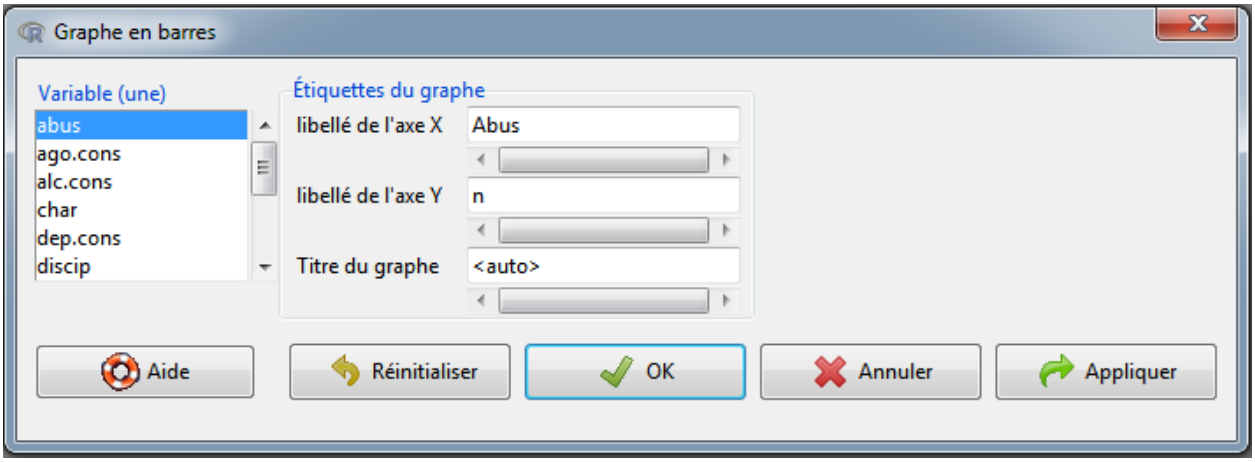


FIGURE 31 – Réalisation d'un diagramme d'une variable qualitative

2) Tester la normalité de la distribution de la variable quantitative (si n1 et n2 <= 30)

Nous allons tester la normalité de la distribution de la durée d'interview.

Statistiques > Résumés > Tests de normalité de Shapiro-Wilk

Le test de Shapiro est significatif ($p<0.05$). Cela signifie que la variable ne suit pas une loi normale.

3) Comparaison de moyennes d'une variable quantitative entre deux groupes

Dans notre exemple, les effectifs dans chaque groupe sont supérieurs à 30, donc nous pouvons réaliser un test de Student.

Si les effectifs étaient inférieurs à 30 et que la durée d'interview suivait une loi normale, alors le test de Student serait toujours valable (seuls les degrés de liberté et la loi seraient modifiés). Si les effectifs étaient inférieurs à 30 et que la durée d'interview ne suivait pas suivi une loi normale, alors le test non paramétrique de Mann-Whitney serait recommandé.

a- Test de Student

Statistiques > Moyennes > t-test indépendant

Le test indique qu'il y a une différence significative ($p<0.05$) de la durée d'interview entre les deux groupes.

b- Test non paramétrique de Wilcoxon

Statistiques > Tests non paramétriques > Test de Wilcoxon bivarié

Le test indique qu'il y a une différence significative ($p<0.05$) de la durée d'interview entre les deux groupes.

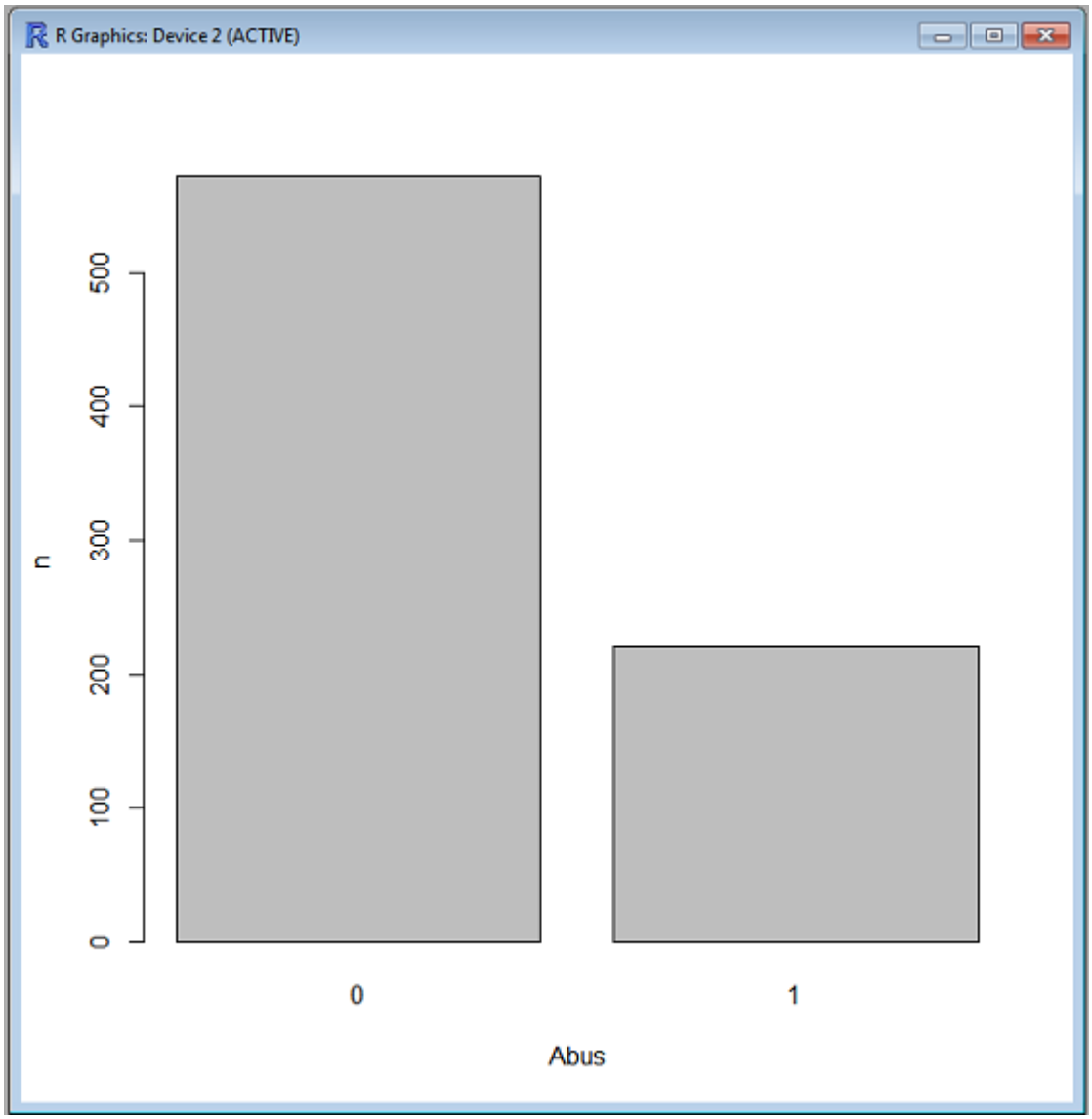


FIGURE 32 – Résultats : diagramme de la variable abus

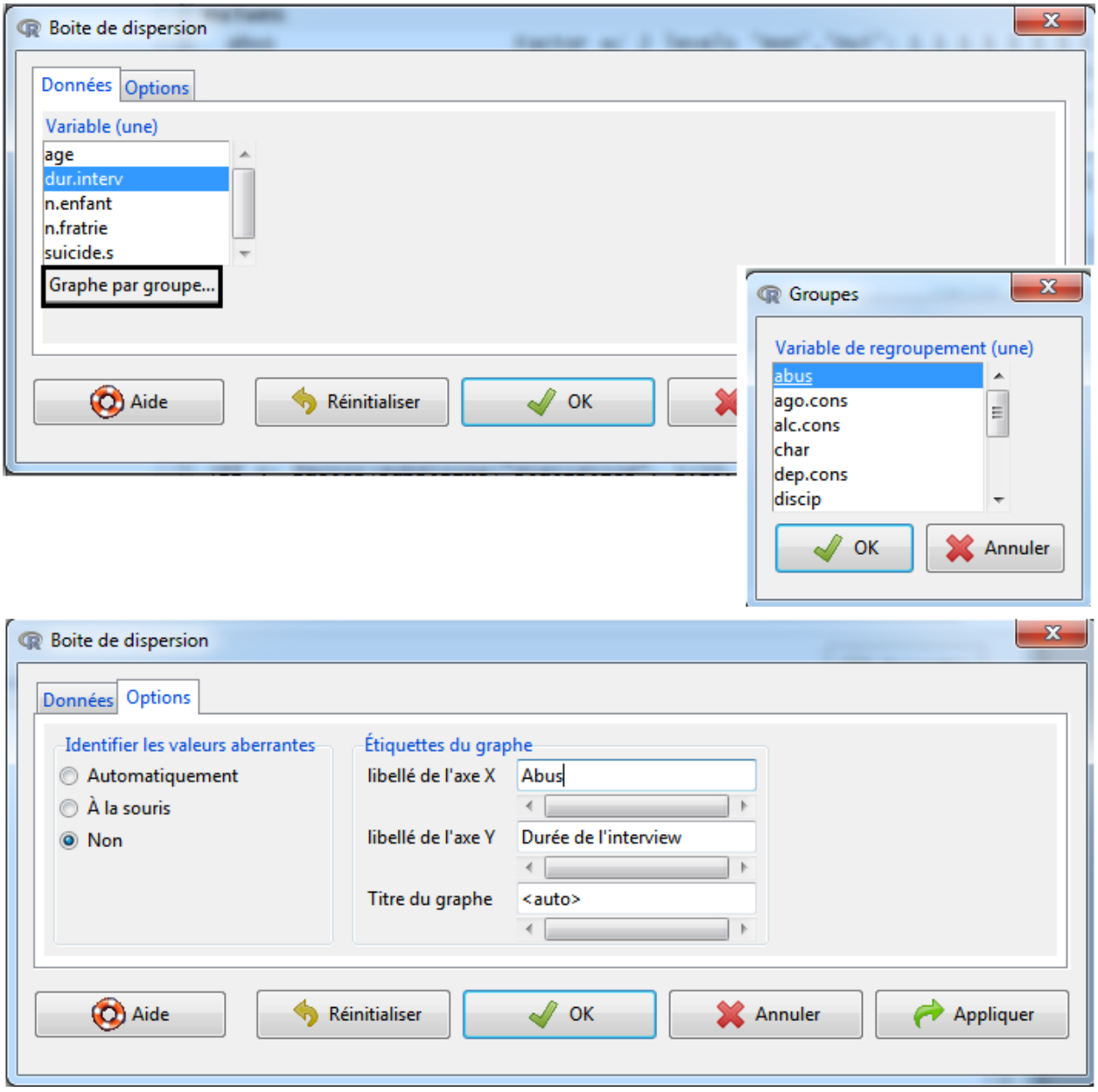


FIGURE 33 – Réalisation d'un diagramme d'une variable quantitative en fonction d'une variable qualitative

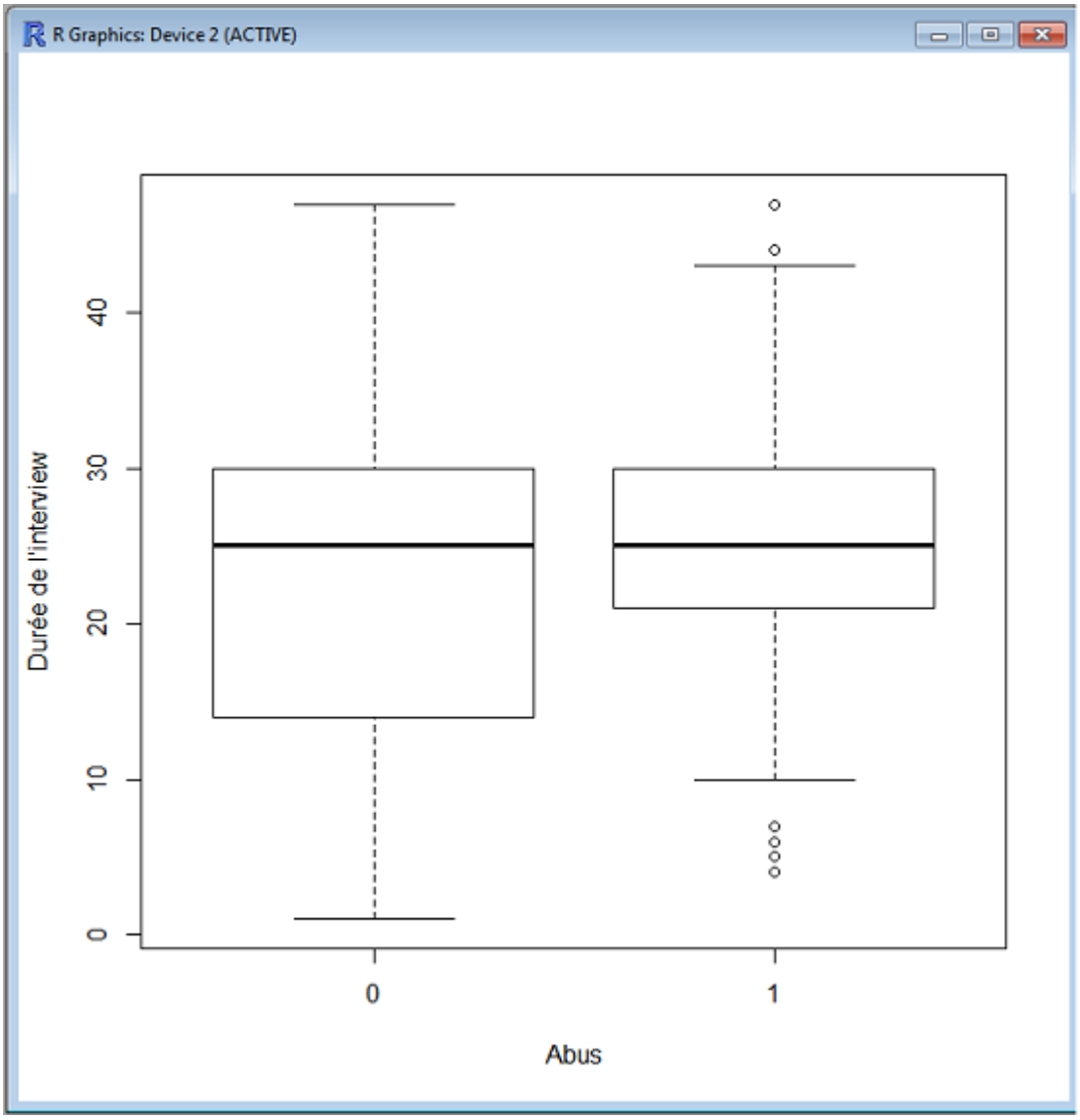


FIGURE 34 – Résultats - diagramme de la variable dur.interv en fonction de la variable abus

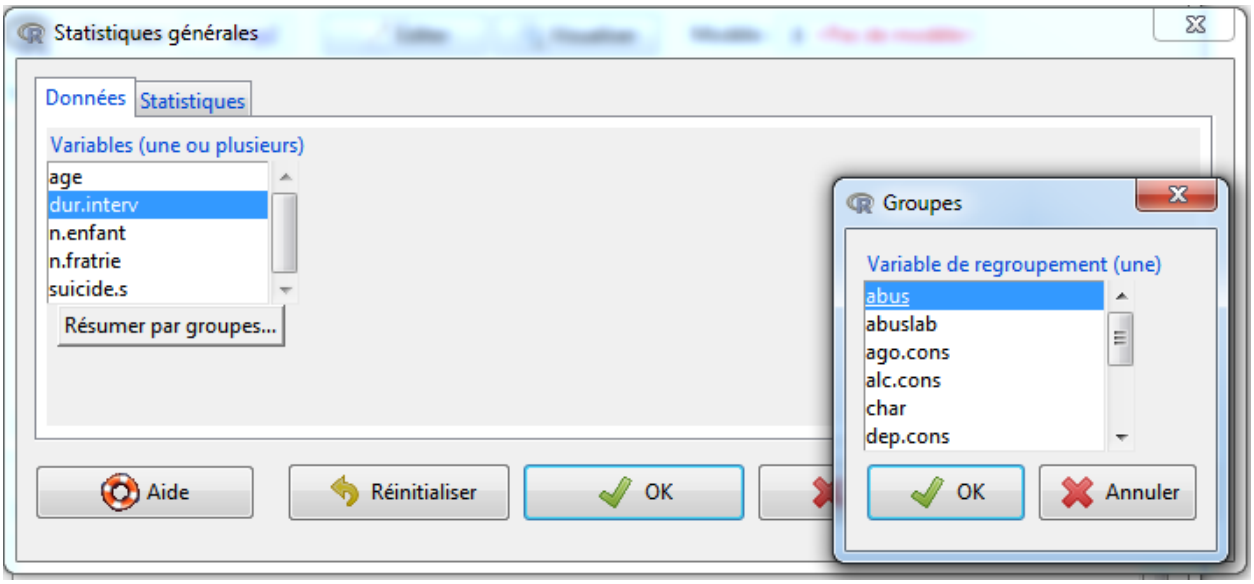
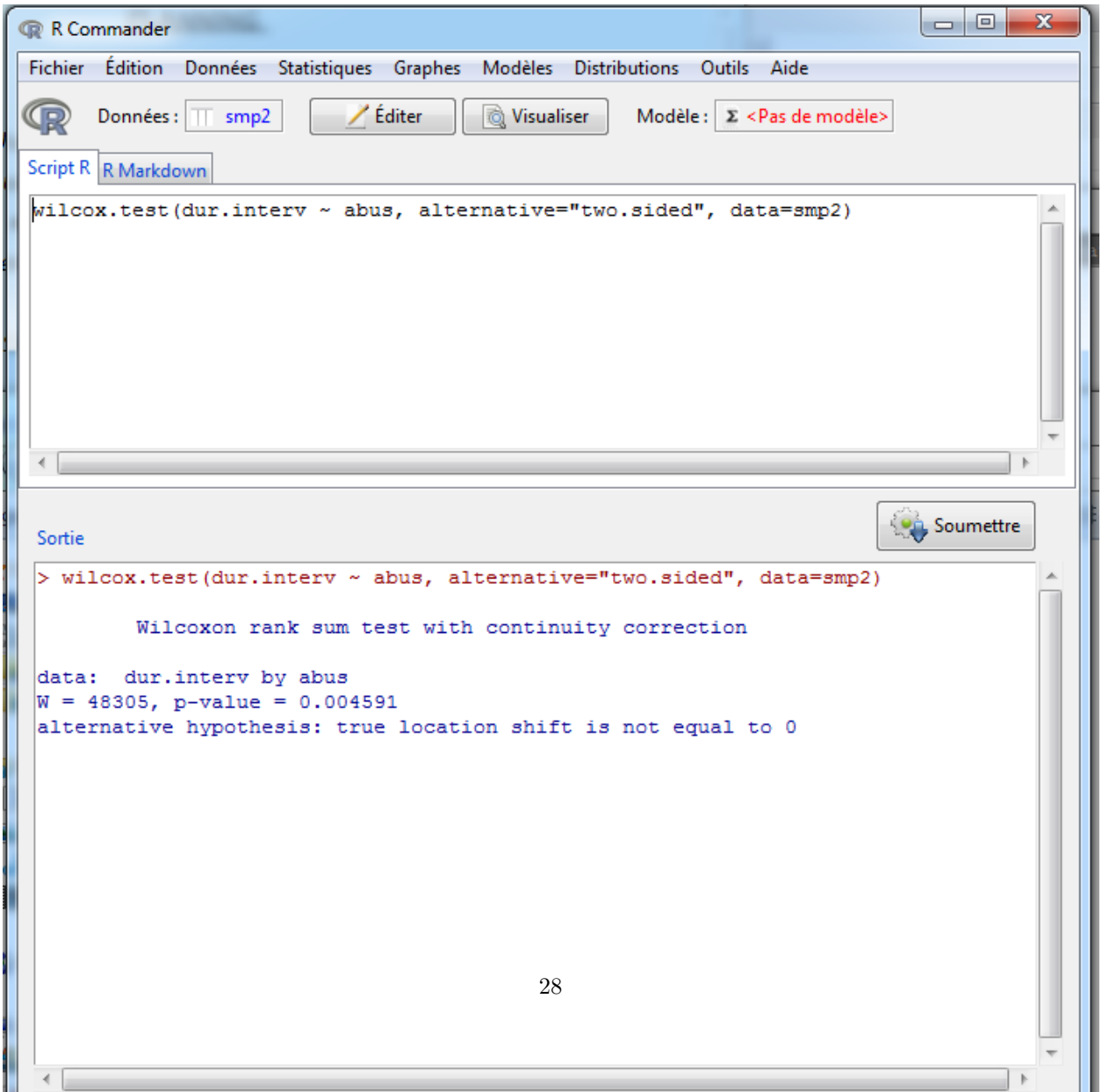


FIGURE 35 – Calcul des moyennes par groupe



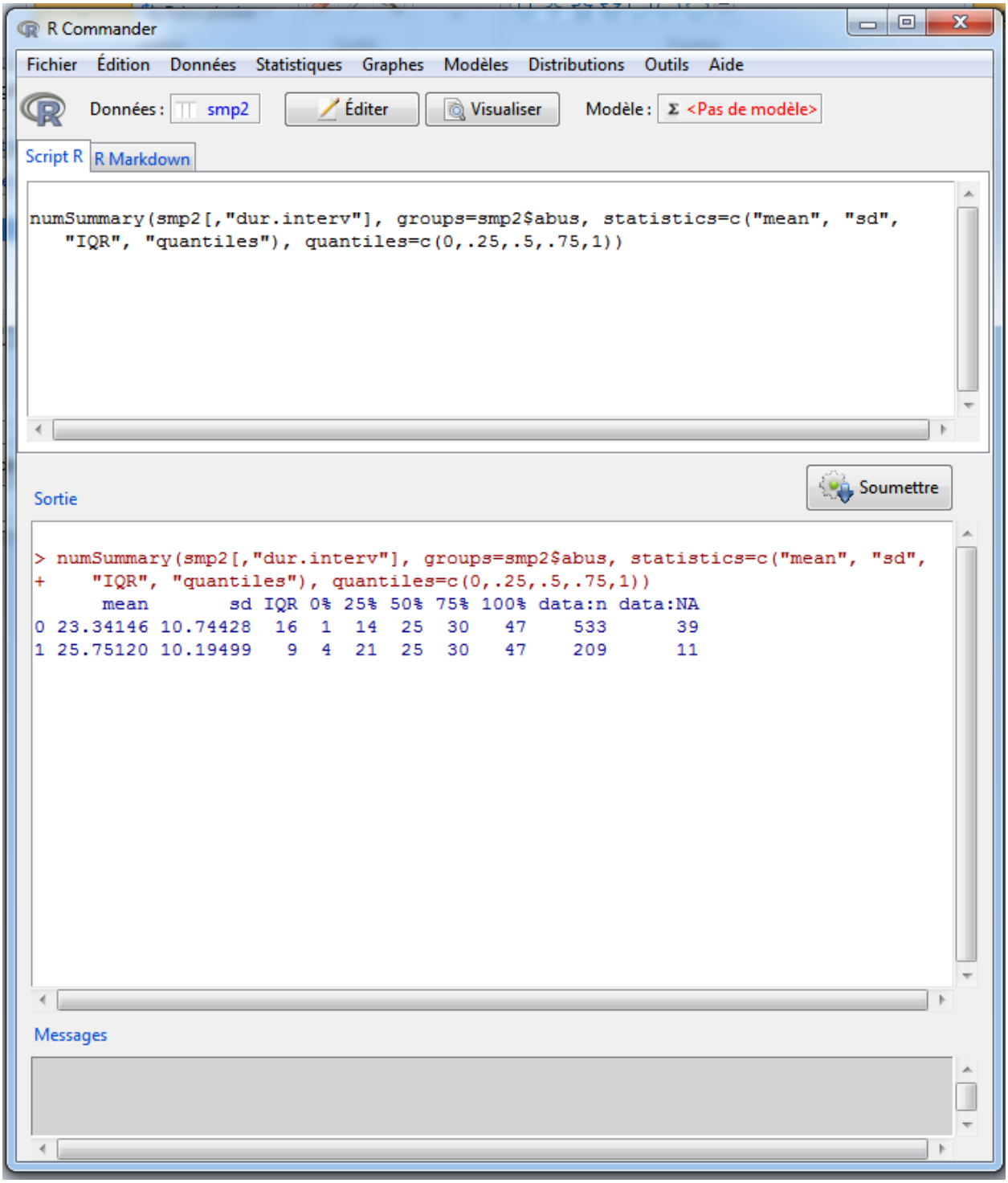


FIGURE 36 – Résultats : moyenne de la variable dur.interv chez les détenus ayant subi des maltraitances et ceux qui n'en ont pas subi

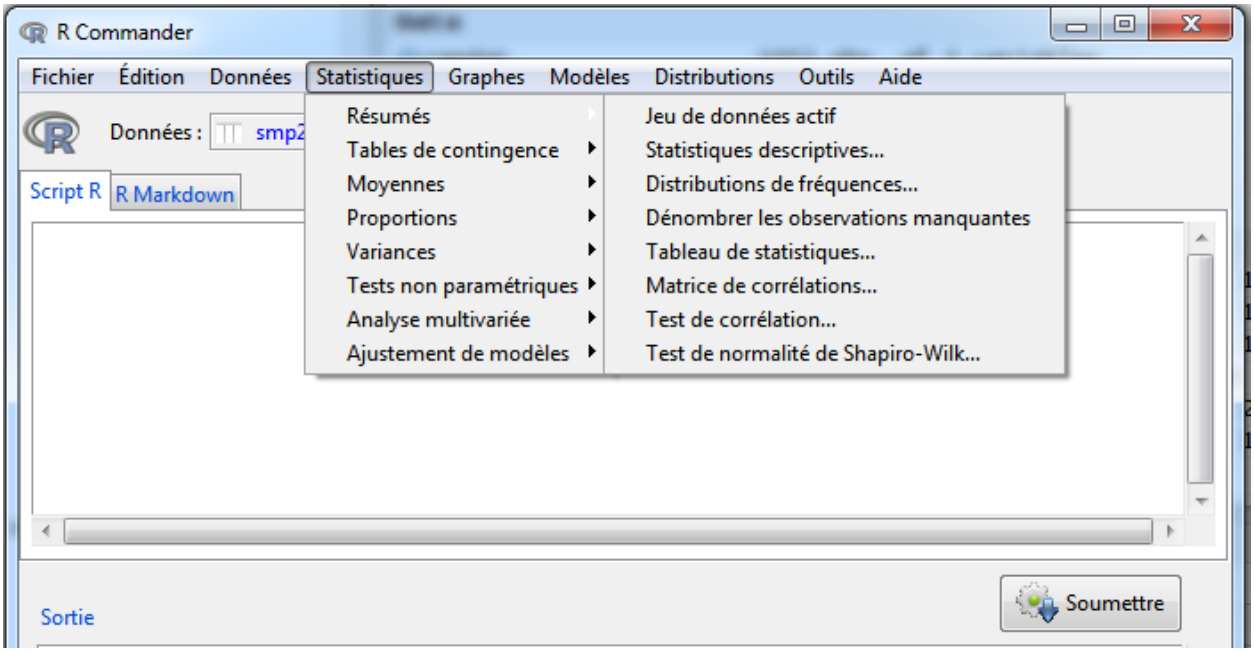


FIGURE 37 – Réalisation d'un test de normalité d'une variable quantitative - Etape 1

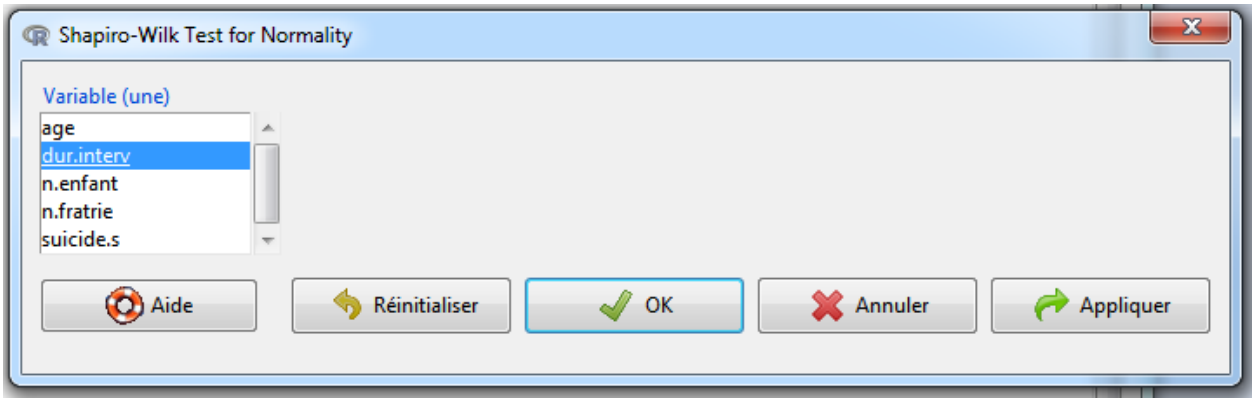


FIGURE 38 – Réalisation d'un test de normalité d'une variable quantitative - Etape 2

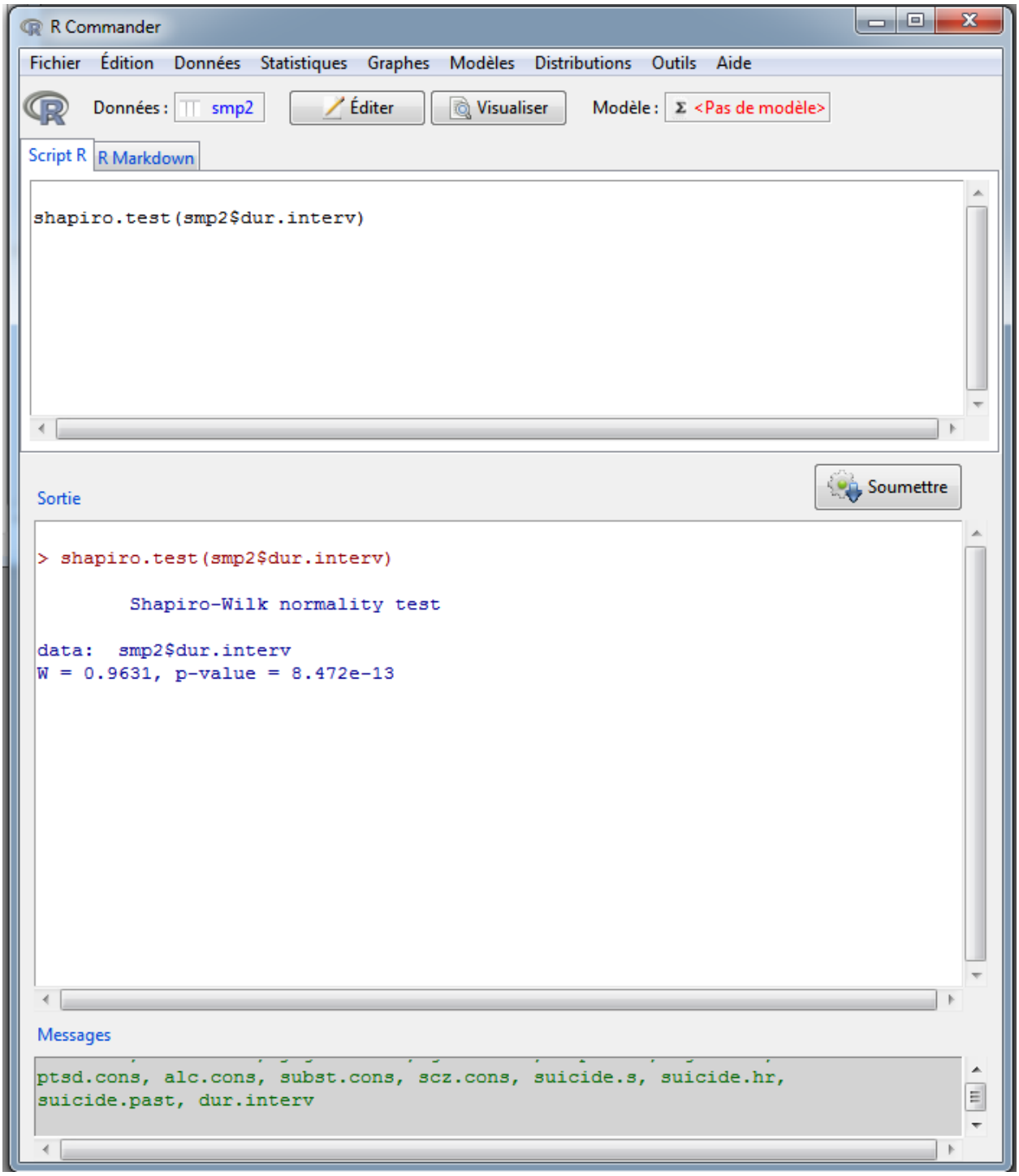


FIGURE 39 – Résultats : test de normalité de la variable dur.interv

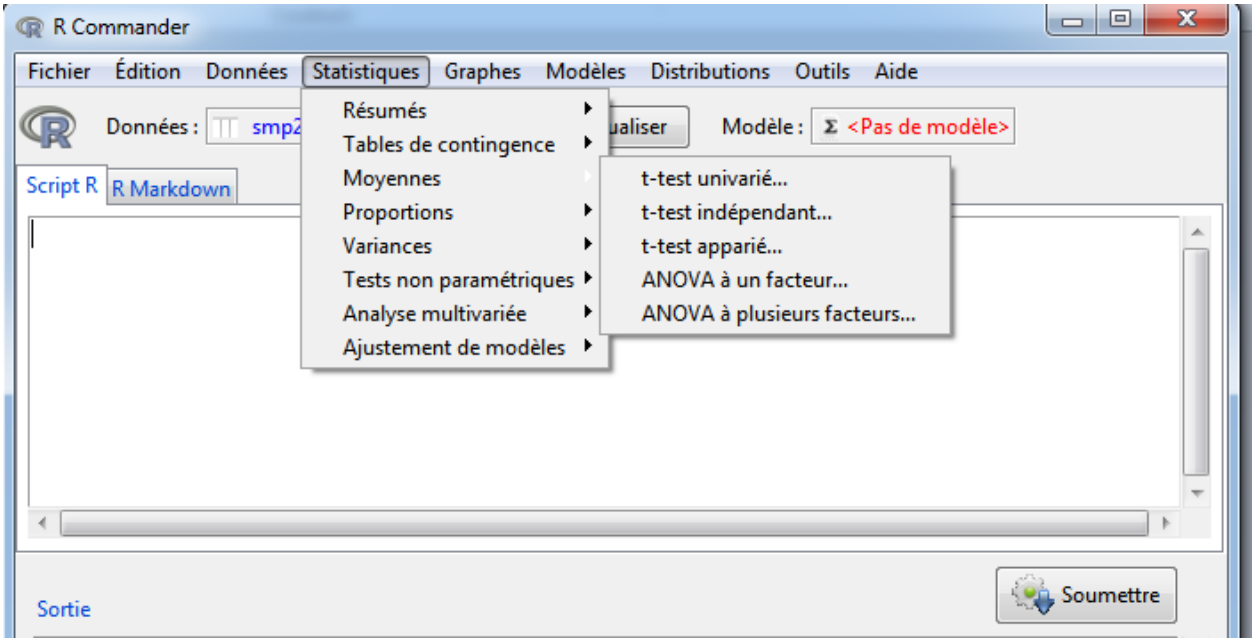


FIGURE 40 – Réalisation du test de Student pour comparer les moyennes entre deux groupes - Etape 1

7.2 *Comparaison de proportions d'une variable qualitative entre deux groupes*

Nous allons comparer statistiquement l'association entre le fait de subir des maltraitances pendant l'enfance et l'existence d'un trouble dépressif.

1) Comparaison de proportions d'une variable qualitative entre deux groupes

Dans notre exemple, le test du khi² est faisable si et seulement si les effectifs espérés sont supérieurs à 5, sinon le test de Fisher est recommandé.

– Test du Khi²

Statistiques > Tables de contingences > Tri croisé

Les effectifs théoriques sont bien supérieurs à 5, le test du Khi² est donc recevable.

Ce test indique qu'il n'y a pas d'association entre le fait de subir des maltraitances pendant l'enfance et l'existence d'un trouble dépressif ($p>0.05$).

– Test non paramétrique de Fisher

Statistiques > Tables de contingences > Tri croisé

Ce test indique qu'il n'y a pas d'association entre le fait de subir des maltraitances pendant l'enfance et l'existence d'un trouble dépressif ($p>0.05$).

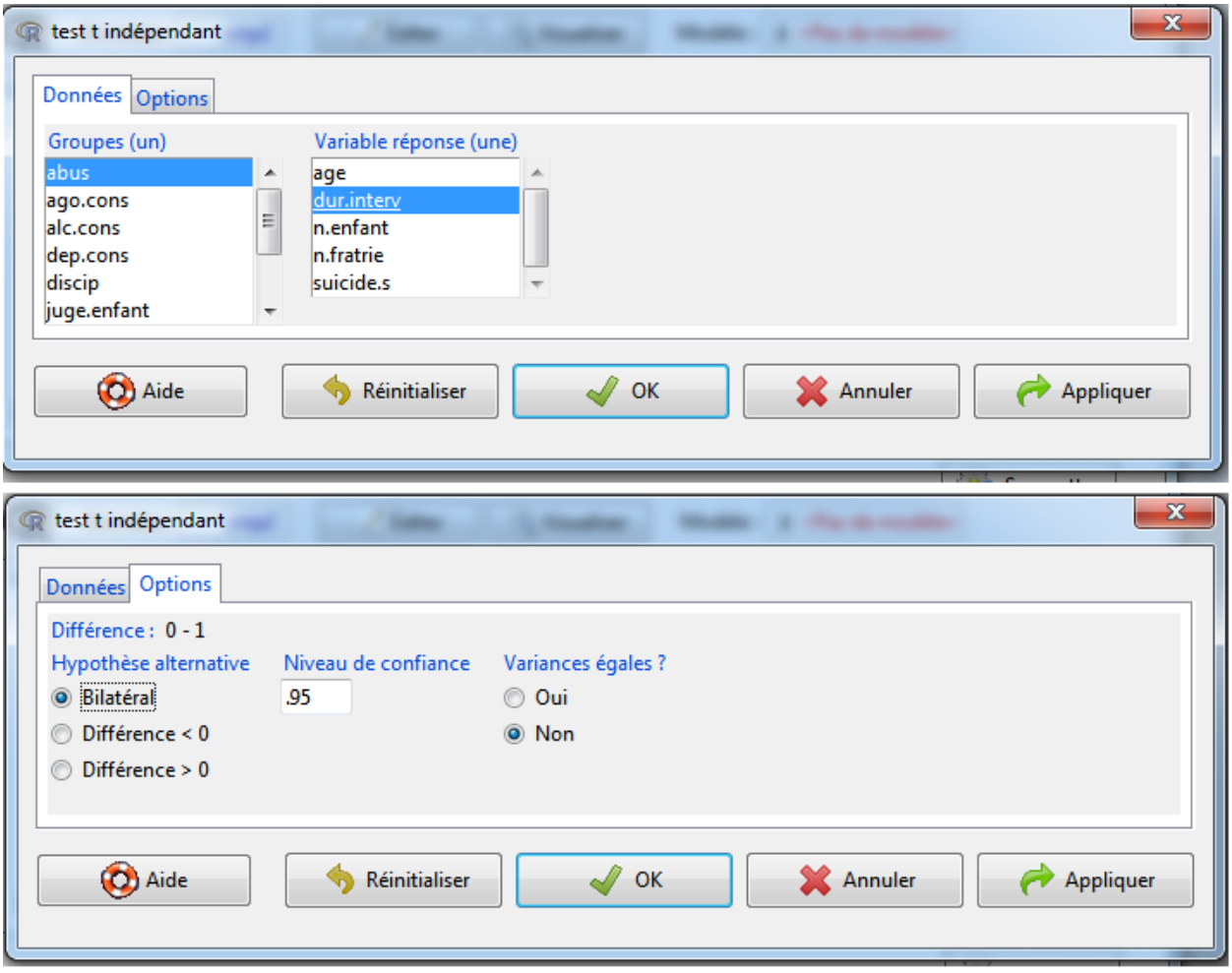


FIGURE 41 – Réalisation du test de Student pour comparer les moyennes entre deux groupes - Etape 2

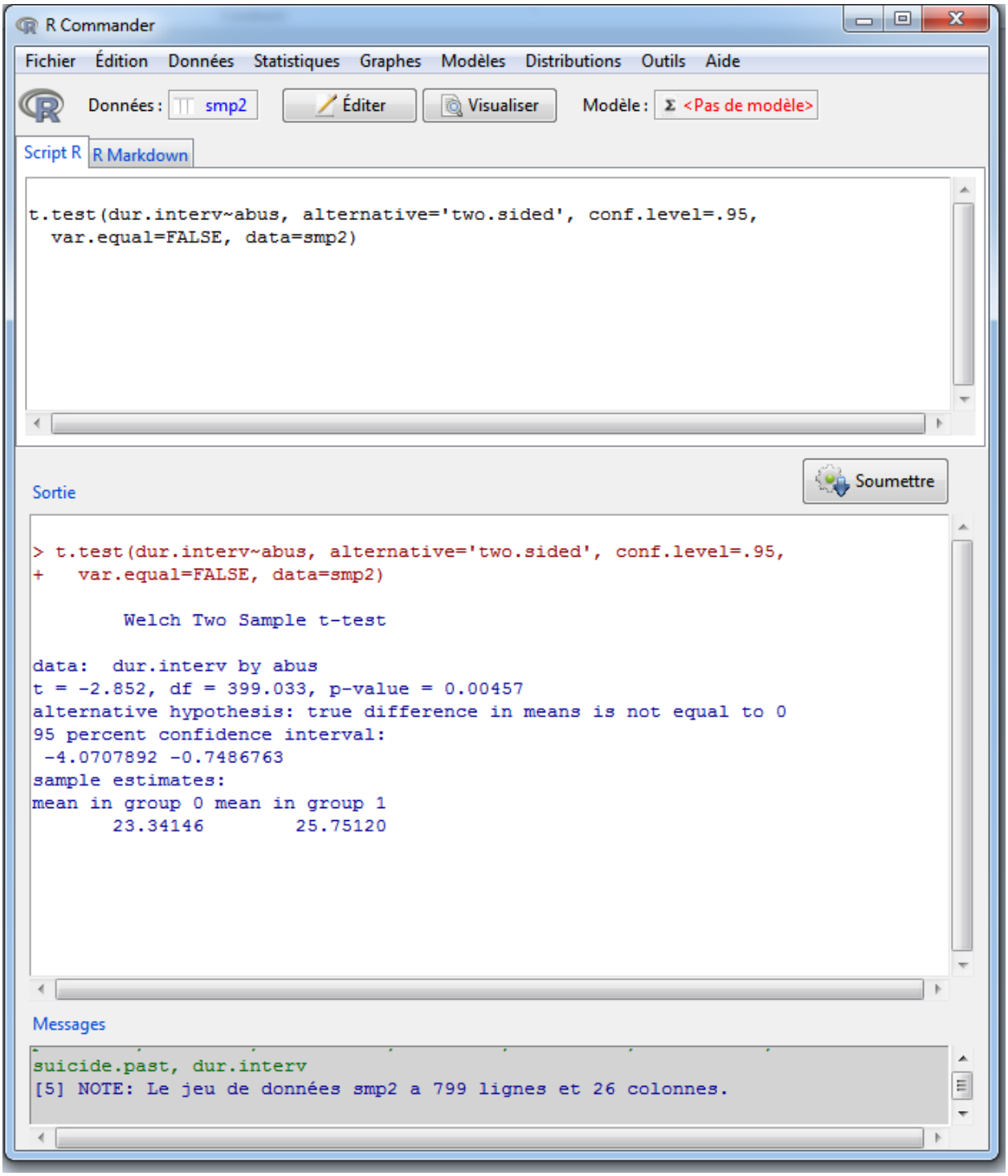


FIGURE 42 – Résultats : test de Student pour comparer les moyennes de la variable dur.interv entre les détenus ayant subi des maltraitances et ceux qui n'en ont pas subi

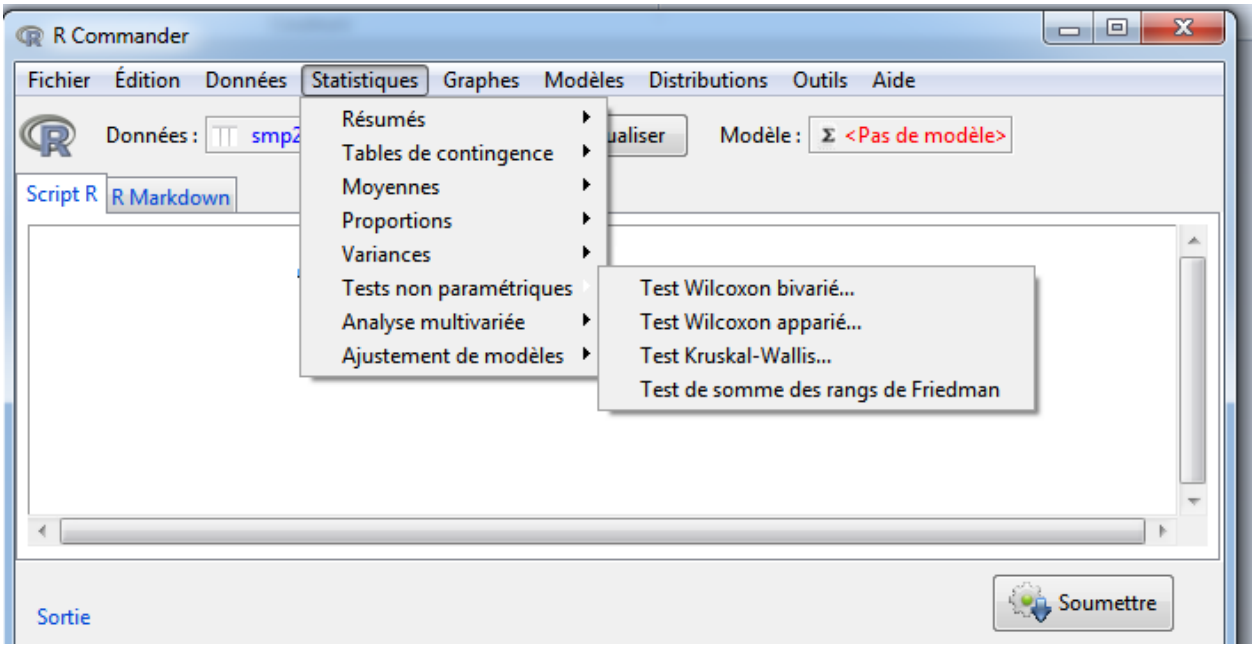
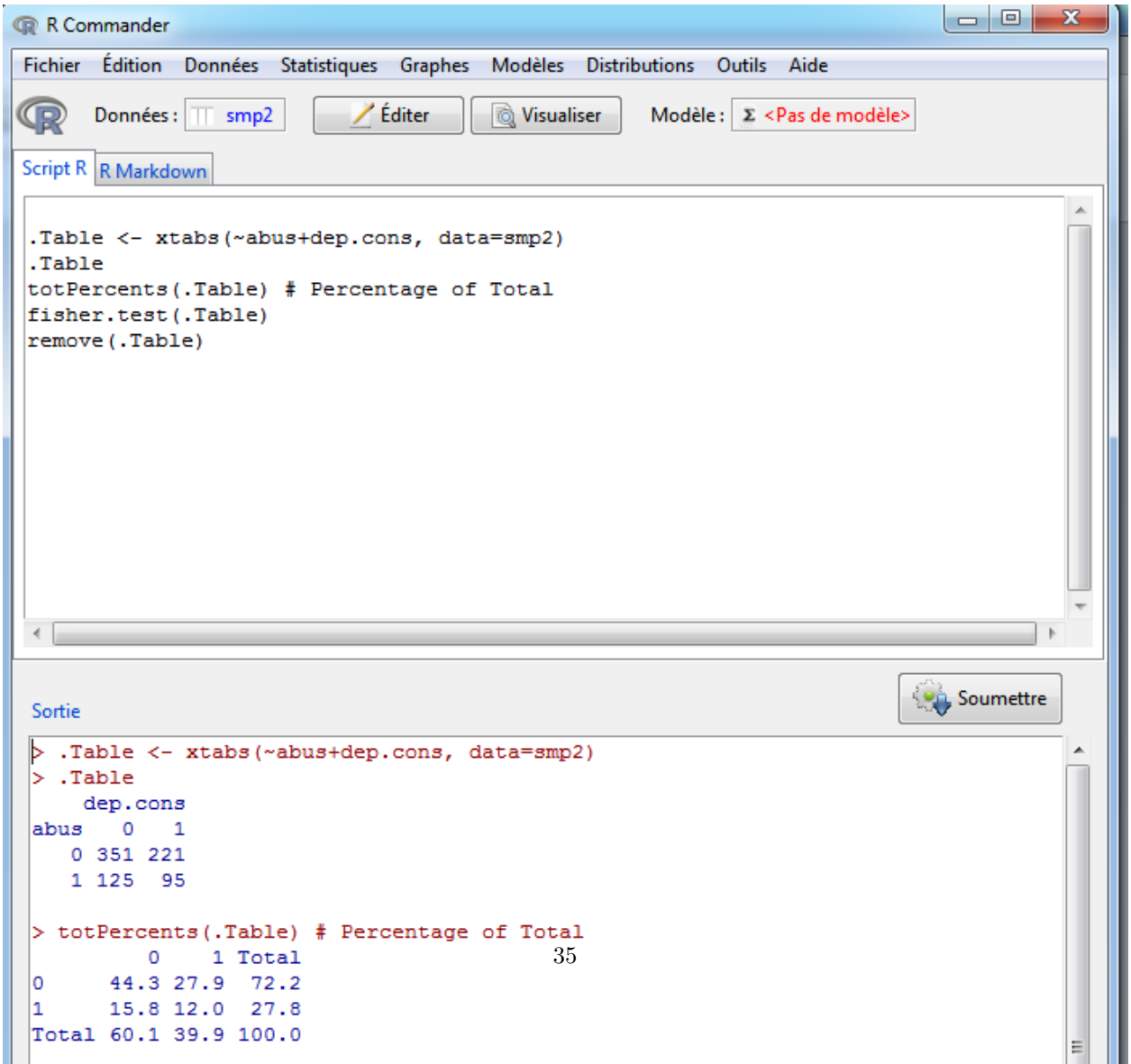


FIGURE 43 – Réalisation du test de Wilcoxon pour comparer les moyennes entre deux groupes - Etape 1



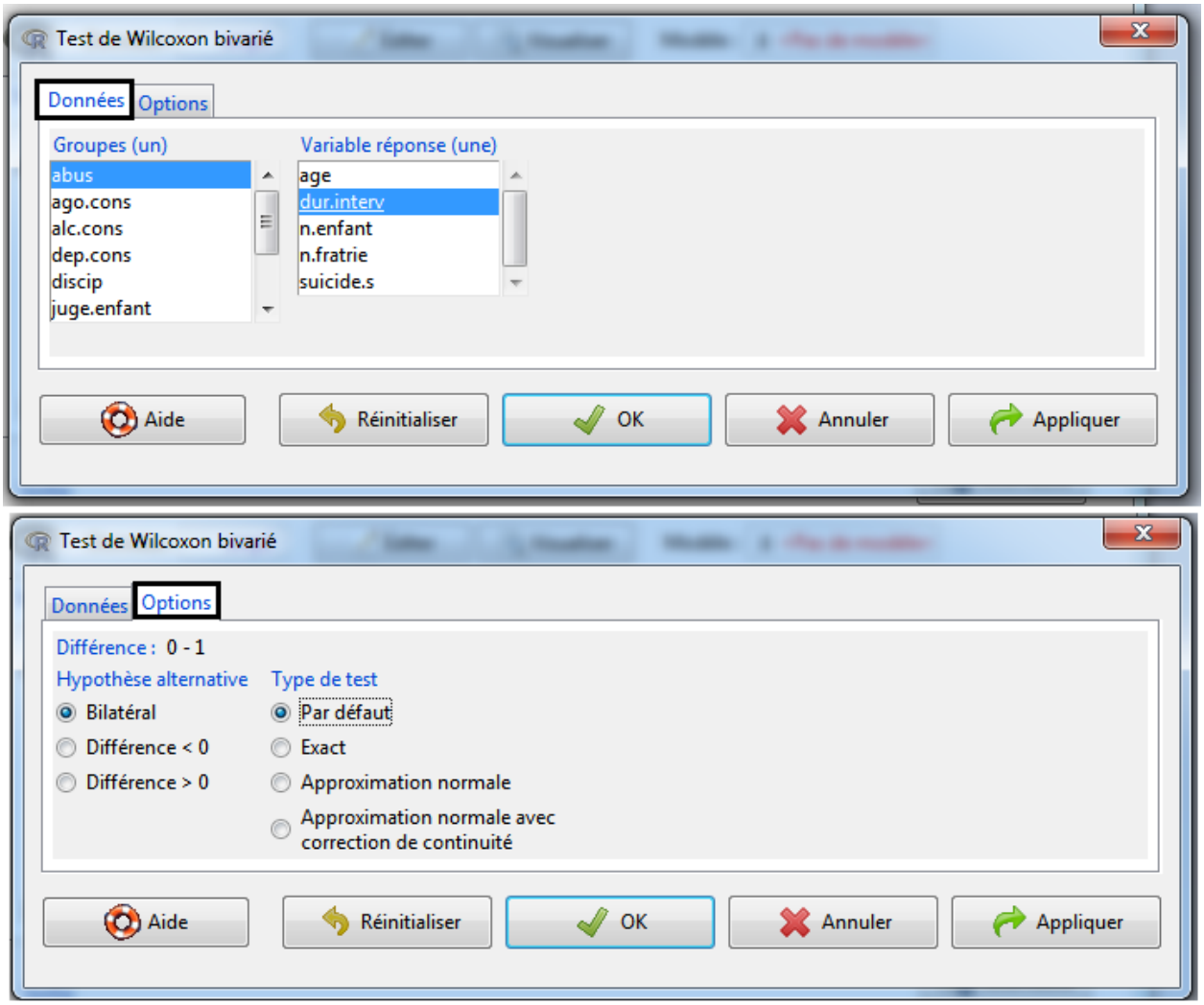


FIGURE 44 – Réalisation du test de Wilcoxon pour comparer les moyennes entre deux groupes - Etape 2

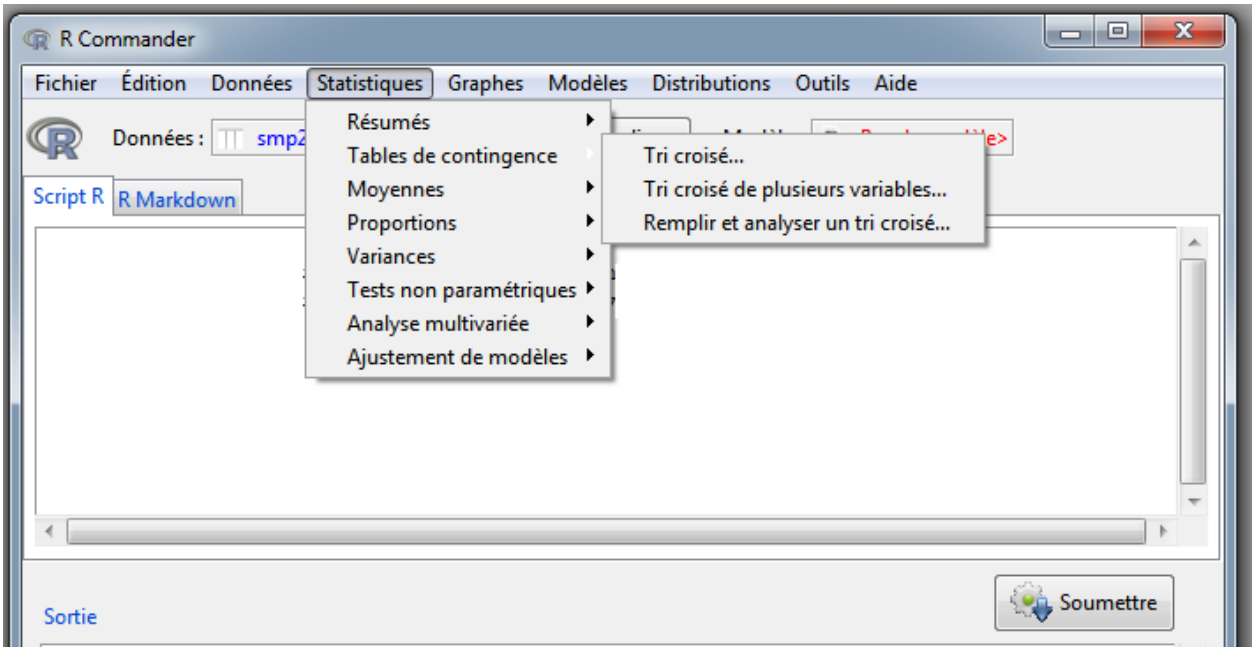


FIGURE 45 – Réalisation du test du Khi² pour comparer les proportions entre deux variables quantitatives - Etape 1

8 Modèles statistiques

Statistiques > Ajustement de modèles

8.1 Modèles linéaires

Nous allons étudier l'association entre la durée de l'interview et l'abus, ajusté sur l'âge du détenu

Statistiques > Ajustement de modèles > Modèle linéaire

Il faut double-cliquer sur les variables afin qu'elles s'affichent dans la formule du modèle.

Le p-value global du modèle est de 0,0005366. Cela signifie qu'une des variables (abus ou âge) est significative.

Il y a une association significative entre la durée d'interview et abus (p=0,003), après ajustement sur l'âge du détenu.

8.2 Modèles logistiques

Nous allons étudier entre le fait de subir des maltraitances pendant l'enfance et l'existence d'un trouble dépressif, ajusté sur l'âge du détenu

Statistiques > Ajustement de modèles > Modèle linéaire généralisé

Il faut double-cliquer sur les variables afin qu'elles s'affichent dans la formule du modèle. Pour indiquer qu'il s'agit d'un modèle logistique, il faut indiquer que la famille est binomiale et la fonction de lien logit.

Il n'y a pas association significative entre le fait de subir des maltraitances pendant l'enfance et l'existence d'un trouble dépressif, après ajustement sur l'âge du détenu.

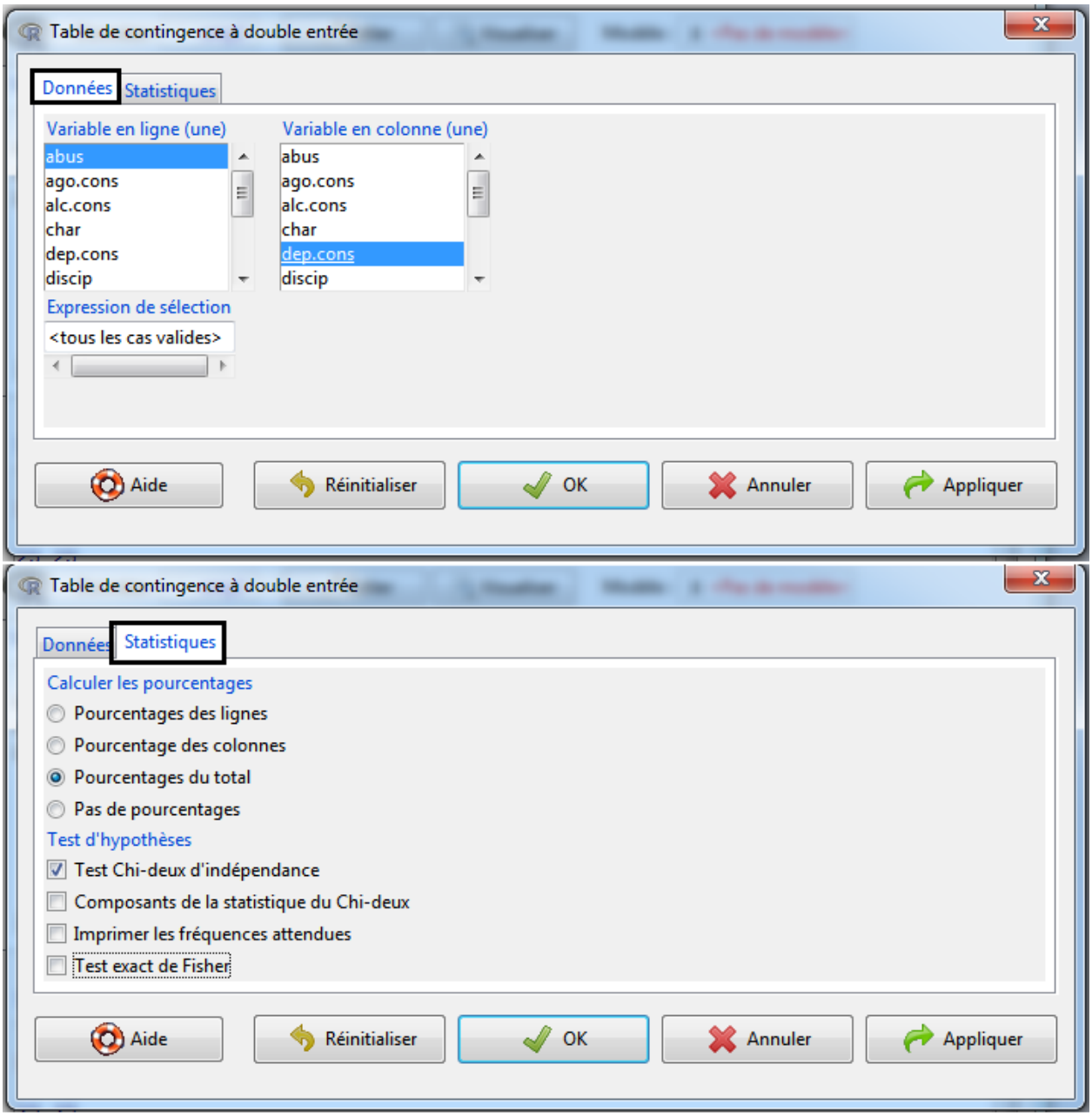


FIGURE 46 – Réalisation du test du Khi² pour comparer les proportions entre deux variables quantitatives - Etape 2

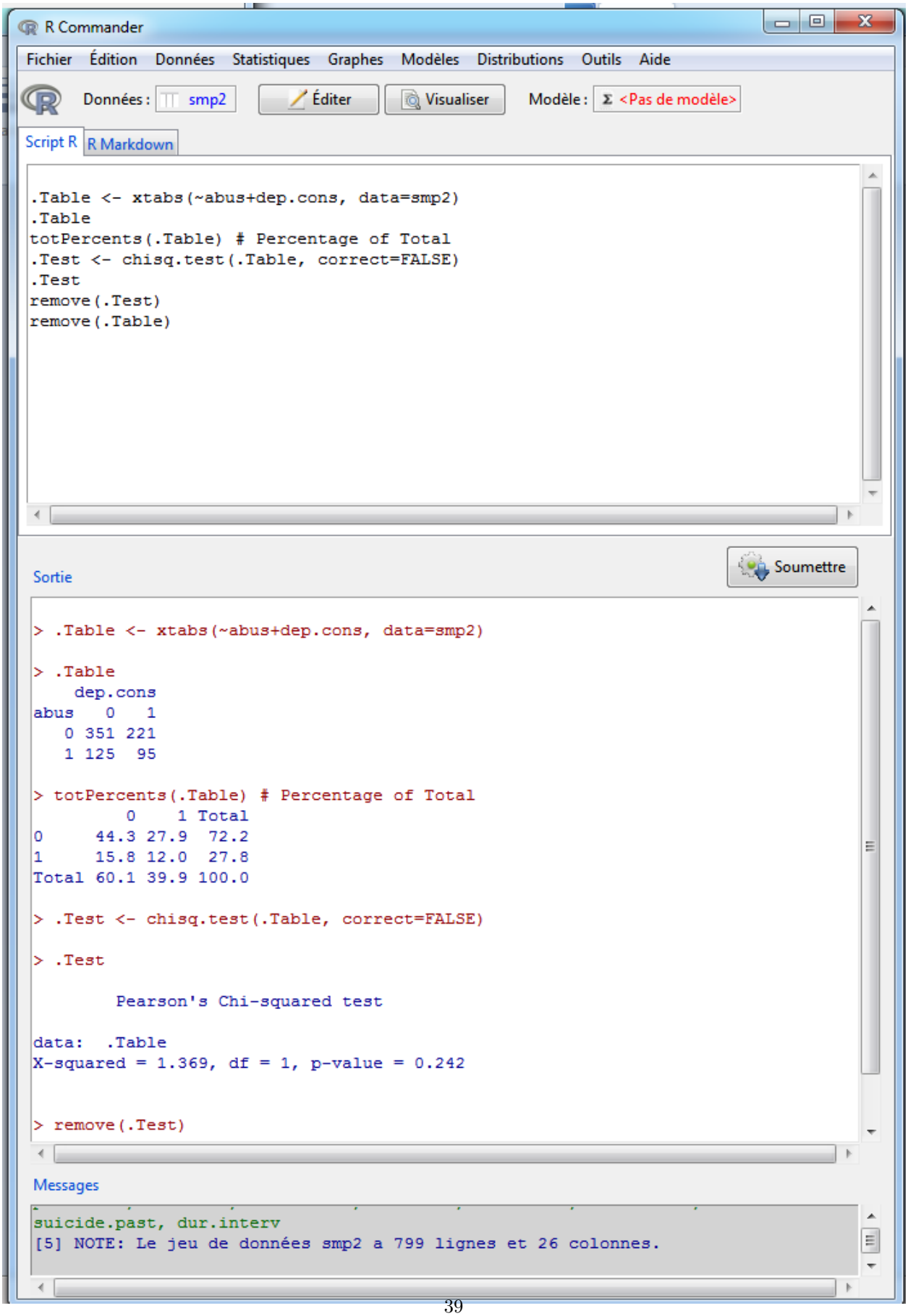


FIGURE 47 – Résultats : test du Khi² pour comparer les proportions entre la variable abus et dep.cons

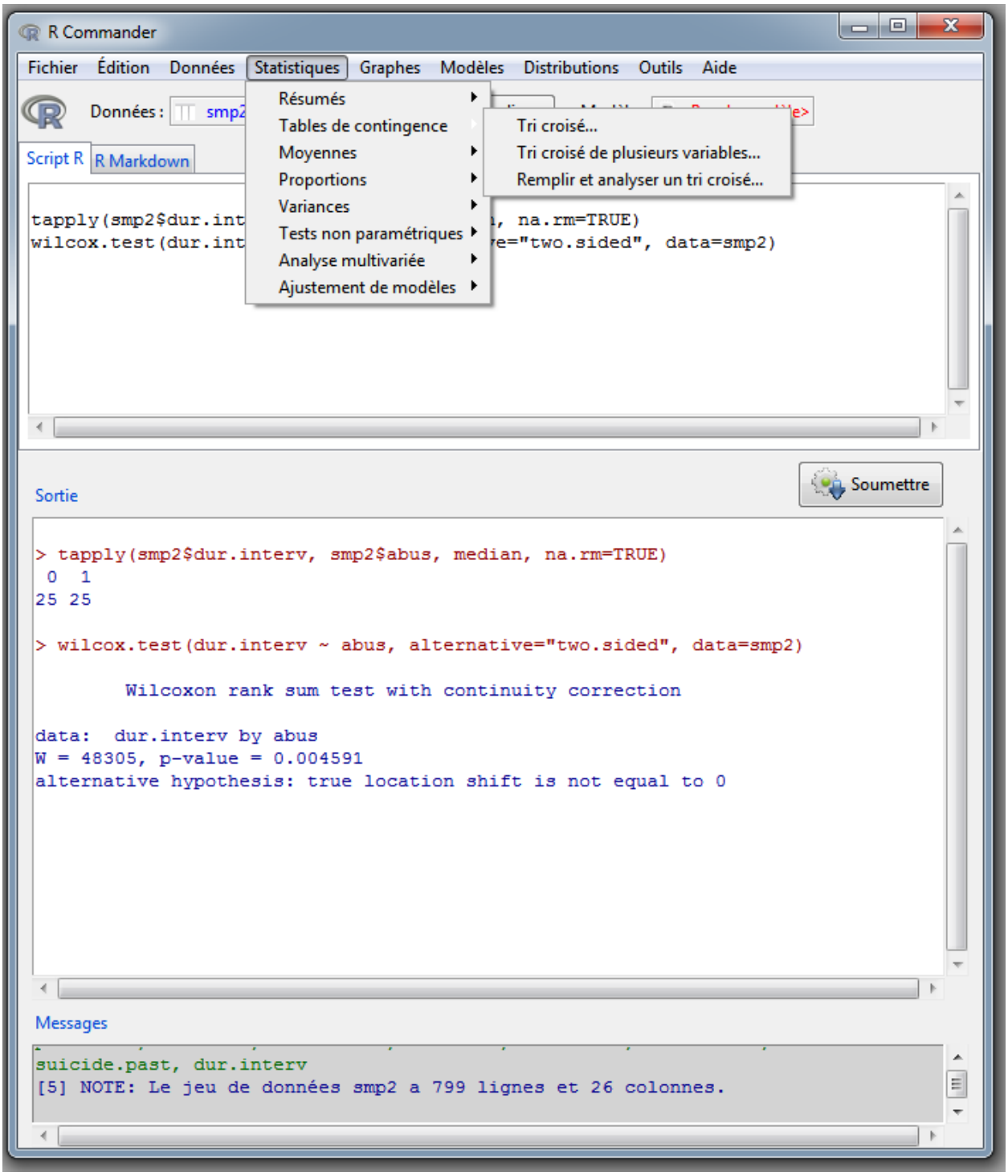


FIGURE 48 – Réalisation du test de Fisher pour comparer les proportions entre deux variables quantitatives - Etape 1

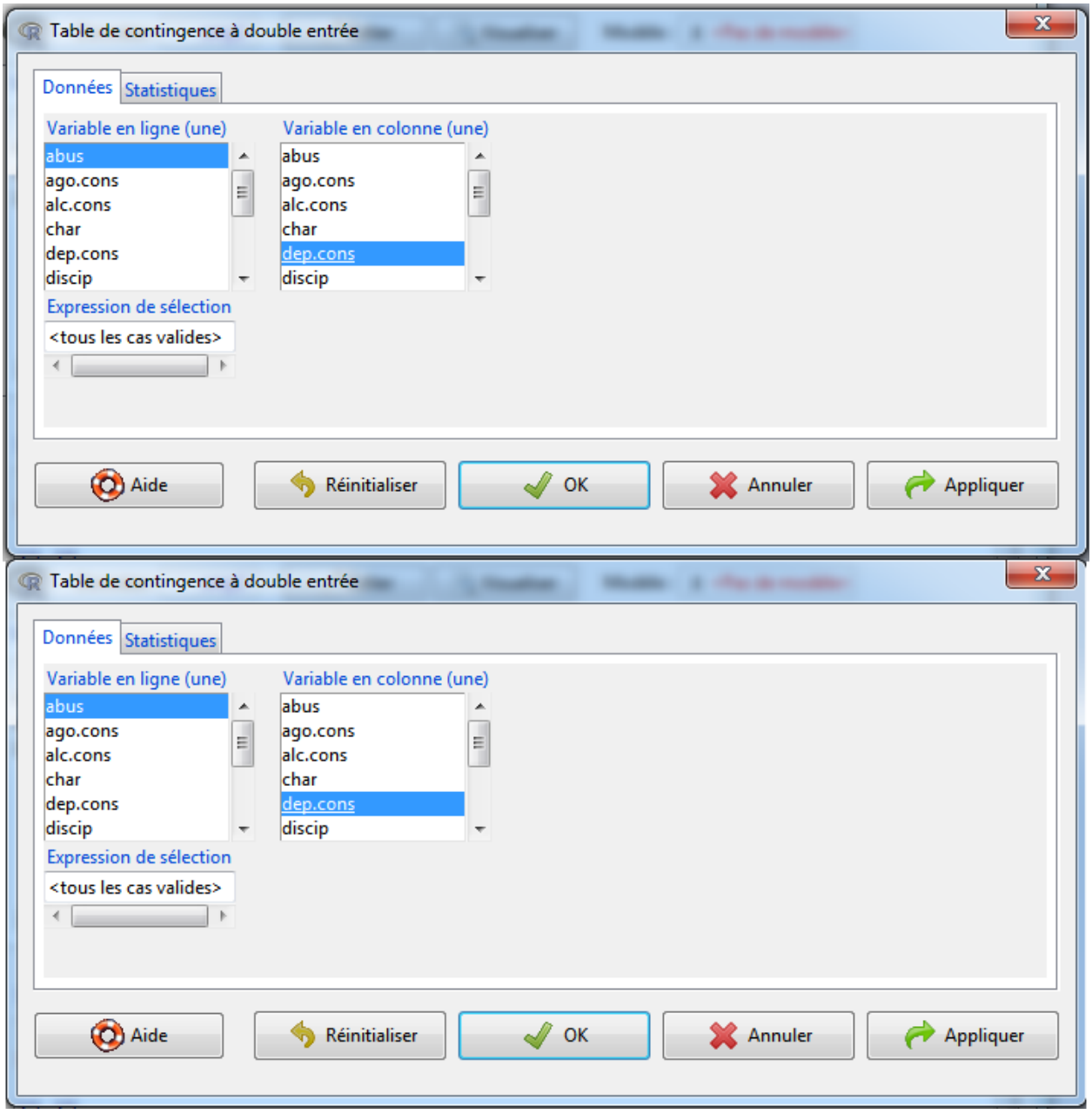


FIGURE 49 – Réalisation du test de Fisher pour comparer les proportions entre deux variables quantitatives - Etape 1 - Etape 2

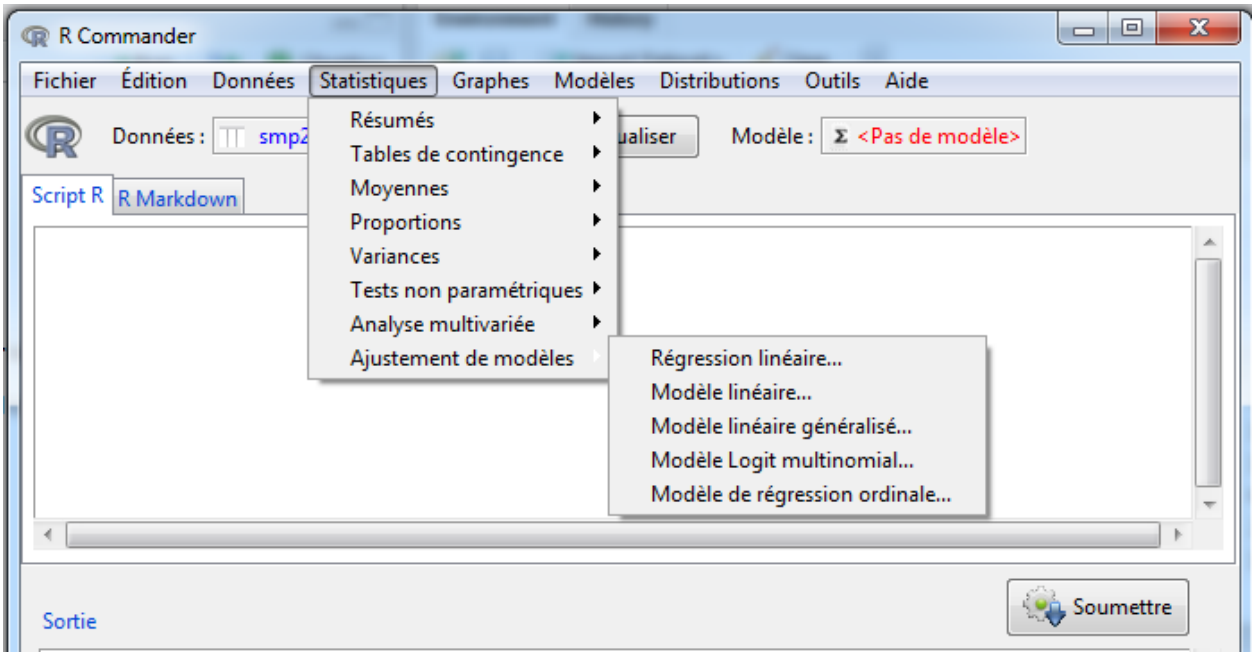


FIGURE 50 – Modèles linéaires et logistiques

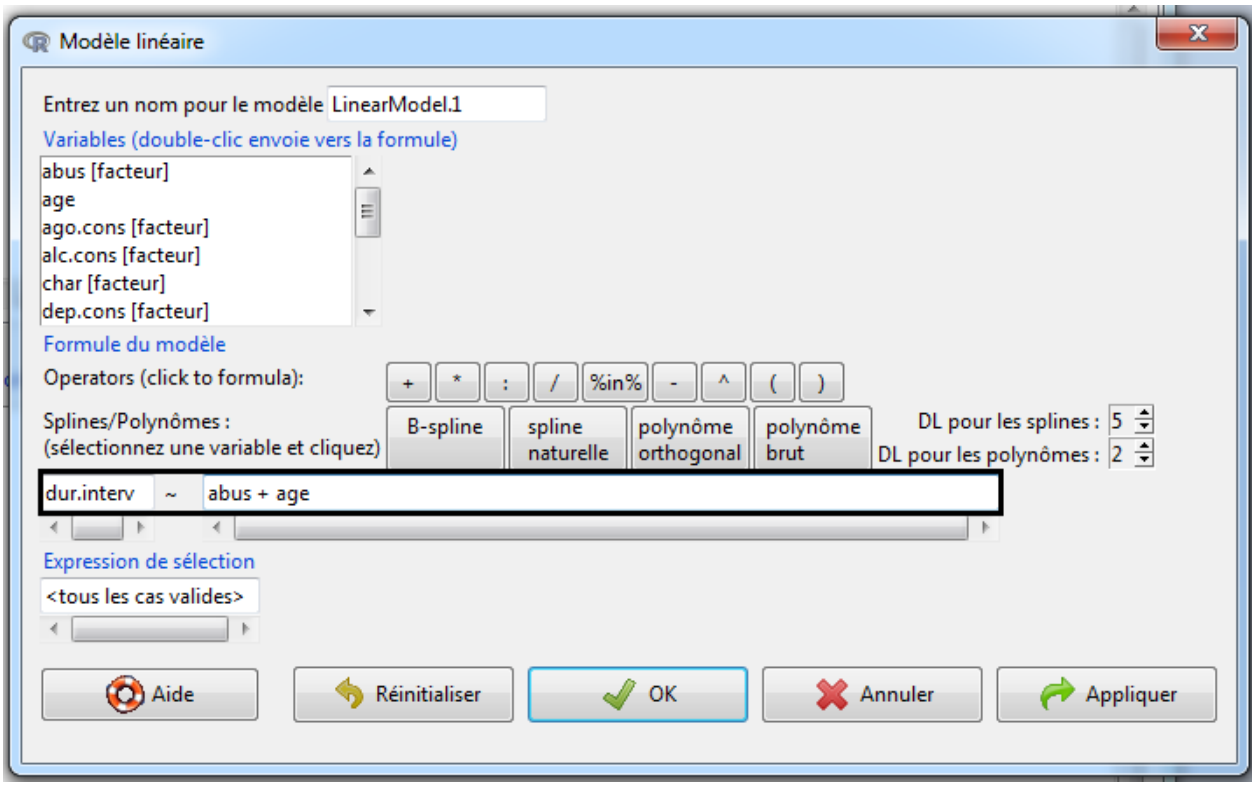


FIGURE 51 – Réalisation d'un modèle linéaire entre une variable quantitative et une variable qualitative

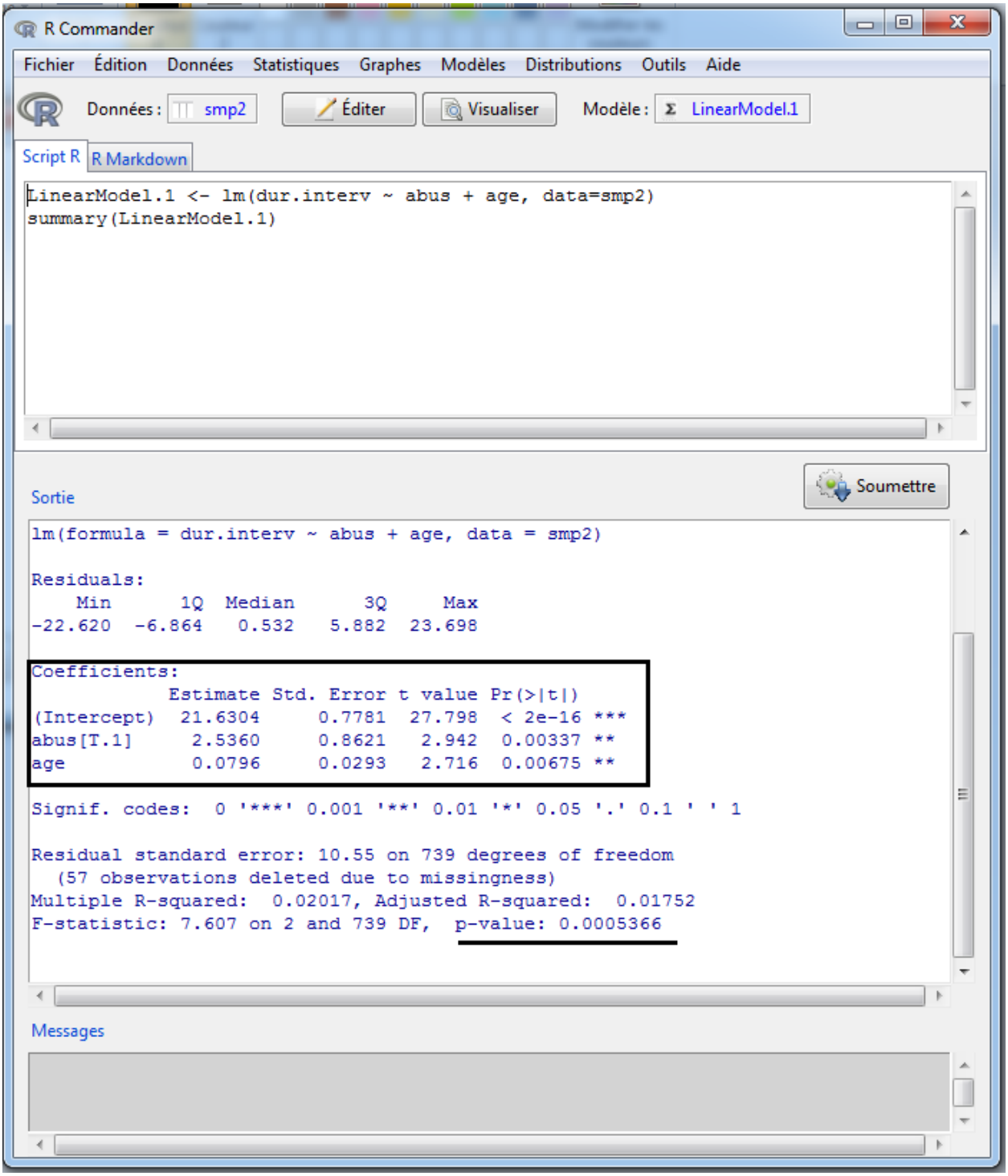


FIGURE 52 – Résultats : modèle linéaire entre la variable dur.interv et abus

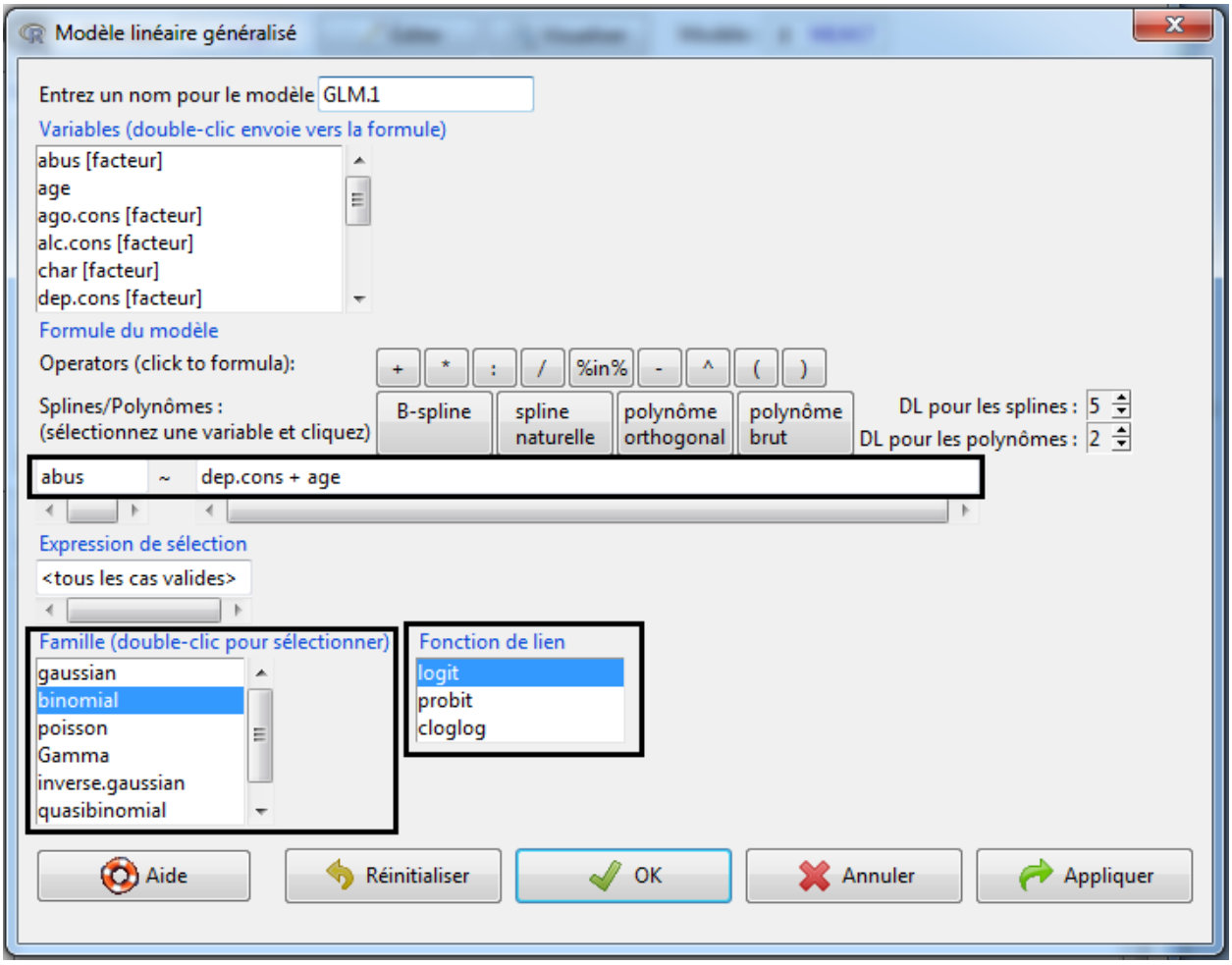


FIGURE 53 – Réalisation d'un modèle linéaire entre deux variables qualitatives

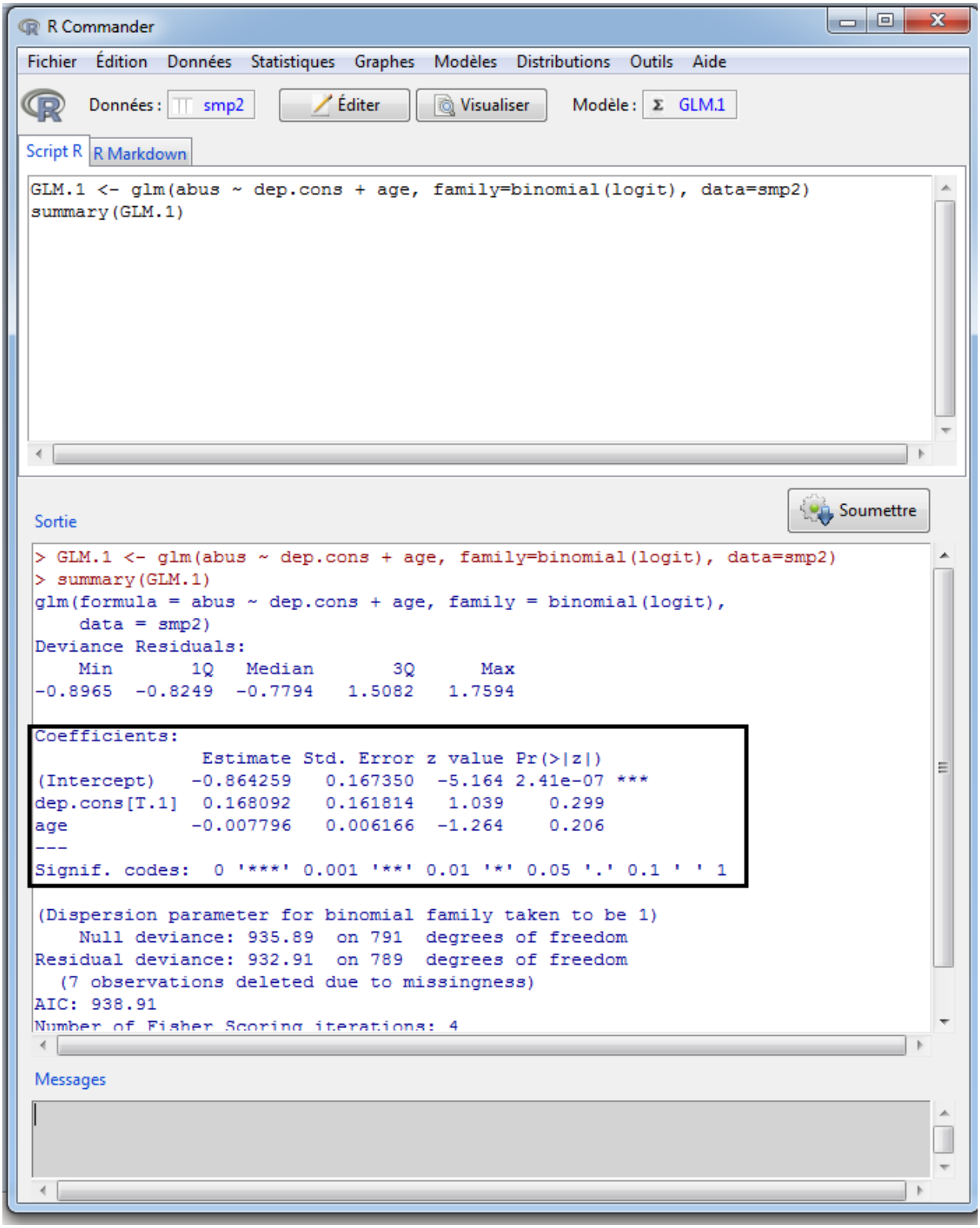


FIGURE 54 – Résultats : modèle logistique entre la variable abus et dep.cons