Guide d'utilisation de SAS® Studio pour réaliser des analyses statistiques simples et des modèles de régression

Loic Desquilbet

▶ **To cite this version:**

Loic Desquilbet. Guide d'utilisation de SAS® Studio pour réaliser des analyses statistiques simples et des modèles de régression. 2023. hal-04038430v1

HAL Id: hal-04038430

https://hal.science/hal-04038430v1

Preprint submitted on 20 Mar 2023 (v1), last revised 26 Apr 2024 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Guide d'utilisation de SAS® Studio pour réaliser des analyses statistiques simples et des modèles de régression

Loïc Desquilbet, PhD en Santé Publique

Professeur en Biostatistique et en Epidémiologie Clinique
Département des Sciences Biologiques et Pharmaceutiques
Ecole nationale vétérinaire d'Alfort

Version 1.6 (20/03/2023)

Préface

Contrat de diffusion

Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International (BY NC ND 4.0). Le résumé de la licence se trouve ici : https://creativecommons.org/licenses/by-nc-nd/3.0/fr/.

Attribution — Vous devez créditer l'Œuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'Oeuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son Oeuvre.

Pas d'Utilisation Commerciale — Vous n'êtes pas autorisé à faire un usage commercial de cette Oeuvre, tout ou partie du matériel la composant.

Pas de modifications — Dans le cas où vous effectuez un remix, que vous transformez, ou créez à partir du matériel composant l'Oeuvre originale, vous n'êtes pas autorisé à distribuer ou mettre à disposition l'Oeuvre modifiée.

Contact

Si vous avez des suggestions de modifications de ce guide (coquilles dans le texte) ou des souhaits de clarification de certains passages, n'hésitez pas à me les signaler par email (loic.desquilbet(at)gmail.com). Par ailleurs, je peux vous envoyer par email les deux fichiers Excel® que je mentionne dans ce guide.

Table des matières

(cc) BY-NC-ND Licence Creative Commons – Guide d'utilisation de SAS® Studio – Loïc Desquilbet – Version v1.6

I. Préambule

A. Comment lire ce guide

Chaque partie de ce guide d'utilisation du logiciel SAS® Studio pour réaliser des analyses statistiques de base ne peut pas se lire avant d'avoir lu les parties précédentes. Ainsi, si par exemple vous souhaitez utiliser un modèle de Cox pour analyser vos données, vous devrez lire … l'intégralité de ce guide !

B. Objectif d'apprentissage général de ce guide

Ce guide d'utilisation de SAS® Studio a pour objectif d'apprentissage général de vous apprendre à vous servir d'un logiciel de statistique reconnu et utilisé par de très nombreux chercheurs pour réaliser des analyses statistiques « simples », sur des données indépendantes (les individus sont considérés comme indépendants les uns des autres), et pour réaliser des modèles de régression univariés et multivariés. Les modèles de régression traités dans ce guide sont la régression linéaire, la régression logistique, et le modèle de Cox.

Les analyses statistiques sur données non indépendantes (comme par exemple dans le cas de séries appariées ou bien dans le cas de recueil longitudinal de données pour un même individu) ne seront pas traitées dans ce guide.

Ce guide n'a pas non plus pour objectif de vous apprendre à vous servir de SAS® pour modifier une base de données (création de nouvelles variables) ou pour créer de nouvelles bases de données à partir de base de données existantes. Enfin, ce guide n'a pas non plus pour objectif de vous présenter de façon exhaustive toutes les options possibles pour analyser d'une certaine façon les données.

C. Présentation rapide du logiciel SAS® Studio

1. Informations préliminaires

SAS® Studio est la version en ligne du logiciel SAS® (pour Statistical Analysis System) dont la première version date de 1976. En tant que membre d'une université (en tant que membre du personnel ou en tant qu'étudiant), il est possible d'accéder gratuitement à SAS® Studio : SAS® OnDemand for Academics. Pour utiliser SAS® OnDemand for Academics, il faut créer un compte. Ensuite, tout se déroule sur Internet, à partir de n'importe quel navigateur Internet. Il n'y a donc aucun logiciel à télécharger puis à installer.

L'utilisation de SAS® nécessite de taper un langage de programmation. Dans SAS® Studio, il est aussi possible de cliquer sur des boutons pour générer automatiquement les lignes de programme correspondantes. J'ai cependant pris le parti, dans ce guide, de vous apprendre à utiliser SAS® en vous faisant taper le langage de programmation, car une fois que l'on sait où trouver dans ce guide les lignes de programme pour réaliser telle ou telle analyse statistique (par exemple, pour réaliser un test de Student), il suffit de le copier et le coller dans l'éditeur de programme en l'adaptant à la situation. Les lignes de programme pour réaliser des analyses statistiques dont je vais parler dans ce guide peuvent être modifiées pour y ajouter, ou retirer, des parties de programmes (qui correspondent entre autres à des options de calculs ou d'affichage).

L'énorme avantage de réaliser ses analyses statistiques avec un logiciel comme SAS® (tout comme avec les logiciels R[1] et Stata® par exemple) est que vous *écrivez* vos analyses statistiques dans un éditeur de programme. Et comme à la fin de vos analyses de la journée vous allez (évidemment) enregistrer votre programme, vous saurez refaire quelques jours, semaines, ou mois après, les analyses faites quelque temps auparavant !

Attention, je vous recommande très fortement d'utiliser un autre navigateur que celui que vous utilisez habituellement, et de le paramétrer en langue anglaise (sauf si celui que vous utilisez habituellement est déjà paramétré en langue anglaise). En effet, si vous utilisez un navigateur Internet paramétré en français, SAS® Studio sera en français. Cela peut paraître plus facile d'accès mais (1) les traductions en français des résultats de SAS® ne sont parfois pas très pertinentes, et (2) quasiment tous les articles scientifiques à comité de lecture indépendant sont écrits en anglais et il sera ainsi plus facile de retrouver dans SAS® des termes statistiques déjà lus dans les articles. Toutes les copies d'écran de ce guide proviennent de l'utilisation de SAS® Studio que je fais tourner sur Chrome® paramétré en anglais.

2. Les deux grandes familles de programmes sous SAS®

Il y a deux grandes familles de programmes sous SAS® : les programmes pour créer ou modifier des bases de données (l'étape DATA) et les programmes pour exploiter ou analyser les données. Ces derniers utilisent très majoritairement des « procédures », dont le nom sera toujours précédé de « PROC » dans un programme SAS®. Par exemple, pour dresser un histogramme, on pourra utiliser la procédure « PROC UNIVARIATE ».

Je vais prendre le parti d'écrire chaque procédure d'une seule façon (sauf exceptions), avec les options que je juge pertinentes pour ce guide. Je ne vais en effet pas présenter l'exhaustivité des options de chaque procédure que je vais utiliser dans ce guide. Néanmoins, dans la partie « Liens Internet vers l'aide de SAS® pour les procédures utilisées dans ce guide » (page 81), je fournis le lien Internet vers l'aide de SAS® de chaque procédure décrite dans ce guide. Dans l'aide de SAS® d'une procédure, la procédure est exhaustivement décrite. Ainsi, lorsque vous commencerez à être un peu à l'aise avec SAS®, vous pourrez par vous-même explorer les options d'une procédure en particulier.

3. Vidéos à visionner avant de poursuivre le guide

Les vidéos ci-dessous doivent être impérativement visionnées (dans l'ordre indiqué) avant de poursuivre ce guide (environ 38 minutes de visionnage).

1. Création d'un compte SAS® OnDemand for Academics (~6 min) : https://youtu.be/YcSTDsgHhng
2. Présentation de l'interface web de SAS® Studio (~12 min) : https://youtu.be/sVvlaC3fsWc
3. Introduction au langage SAS® (~10 min) : https://youtu.be/Zc3j-J_3MK4
4. Importation d'un fichier Excel® dans SAS® Studio (~3 min) : https://youtu.be/QO2wZbQgDG0
5. Etape DATA et procédures SAS® (~7 min) : https://youtu.be/Yl-KhPFKUbQ

Pour aller plus loin, les vidéos ci-dessous peuvent être visionnées aussi. Mais le visionnage n'est pas nécessaire pour la compréhension de ce guide.

1. Bibliothèques SAS® : https://youtu.be/iyevNla7k1c
2. Création de variables dans SAS® : https://youtu.be/b-0tygyxCgQ
3. Fonctions SAS® : https://youtu.be/6v8QxKGcEJc

[1] https://www.r-project.org/

D. Environnement de travail sous SAS® pour ce guide et conventions utilisées

1. Environnement de travail

Pour ce guide, j'ai créé un dossier sous « Files (Home) » que j'ai intitulé « Pour Guide pratique SAS Studio » (cf. flèche (a) sur la Figure 1), au sein duquel j'ai enregistré mon programme (cf. flèche (b) sur la Figure 1) qui va, au fur et à mesure de ce guide, se remplir de lignes de programmation.

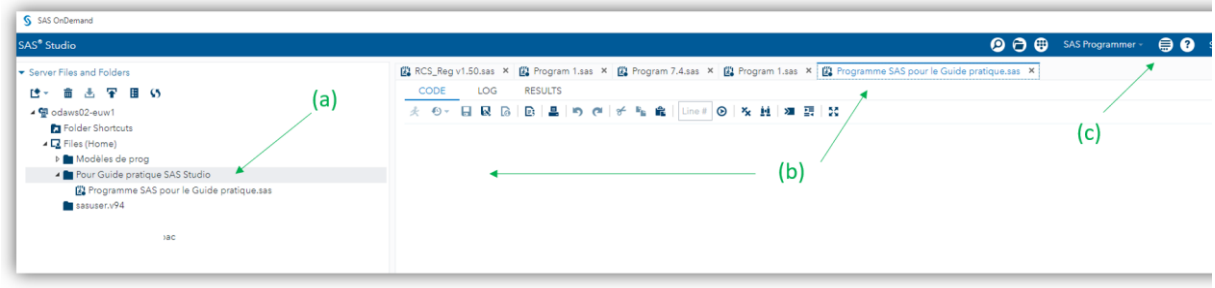


Figure 1

Pour information, j'ai modifié quelques paramétrages sous SAS® Studio, en cliquant sur « More application options », puis « Preferences », Figure 1.c). Dans « Code and Log », j'ai désactivé la saisie semi-automatique (cf. flèche (a) sur la Figure 2).

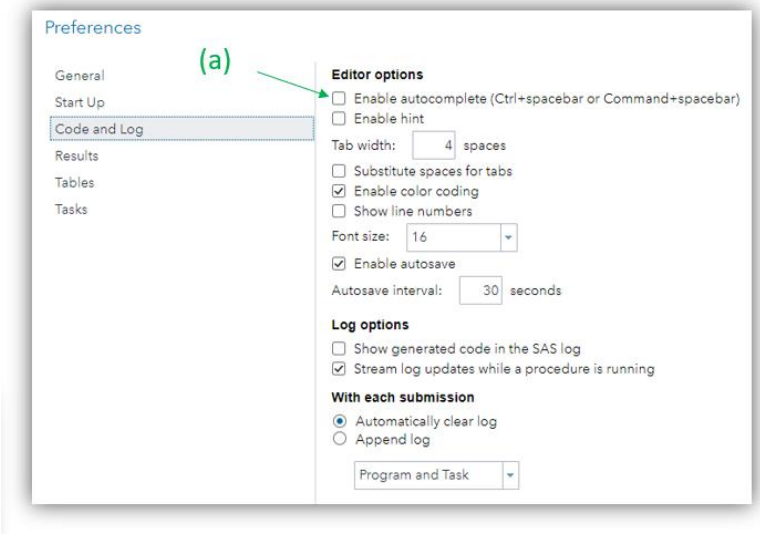


Figure 2

J'ai aussi modifié la mise en forme des résultats dans « Results » (cf. flèches (a) sur la Figure 3).

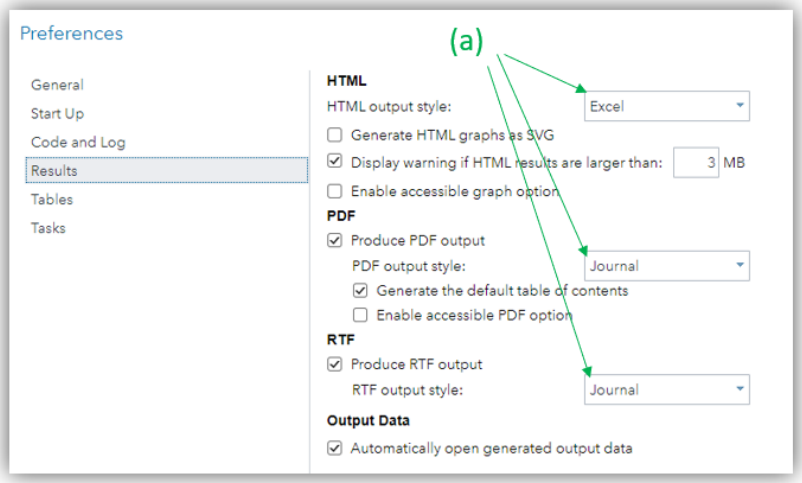


Figure 3

2. Conventions utilisées pour ce guide

Lorsque je vais écrire dans le texte de ce guide les résultats statistiques que SAS® fournit, je vais arrondir à deux chiffres après la virgule si plus de deux chiffres après la virgule sont fournis par SAS®.

Par ailleurs, dans les lignes de programmation que je vais écrire, les mots-clés de SAS® seront écrits en majuscule et en gras.

E. Présentation du fichier de données fictif

Le fichier de données qui va être utilisé dans ce guide est un fichier de données Excel®, nommé « Données pour guide SAS v3.5.xlsx » (que je peux vous envoyer pour reproduire sous SAS® de votre côté tous les résultats présentés dans ce guide). Les données sont contenues dans l'onglet « Donnees » de ce fichier Excel®. Ce fichier de données, fictif, comprend 99 chiens adultes suivis au cours des années à partir d'une consultation chez un vétérinaire qui correspond à l'inclusion dans l'étude (J0). Tous les chiens de l'étude ont été suivis au moins 3 ans, sans aucun perdu de vue. Les 19 variables contenues dans ce fichier de données sont listées alphabétiquement ci-dessous.

AGE : variable correspondant à l'âge du chien à J0, en années entières.

AGE_4CL : variable codée en 0/1/2/3 à partir des quartiles de la variable AGE. Cette variable vaut « 0 » pour les chiens dont l'âge était < 7 ans, « 1 » pour les chiens dont l'âge était compris entre 7 et 9 ans (exclu), « 2 » pour les chiens dont l'âge était compris entre 9 et 11 ans (exclu), et « 3 » pour les chiens dont l'âge était supérieur ou égal à 11 ans.

ALAT : variable correspondant à la concentration en ALAT, en UI/L.

AUTRE_RACE : variable codée en 0/1. Elle vaut « 0 » si le chien était de race Labrador, Golden, ou de race croisée Golden/Labrador, et « 1 » s'il était d'une autre race.

CHOLES_3CL : variable codée en 0/1/2. Elle vaut « 0 » pour les chiens qui présentaient une hypocholestérolémie, « 1 » pour une normocholestérolémie, et « 2 » pour une hypercholestérolémie.

CREAT : variable correspondant à la concentration en créatinine, en mg/L.

CROISEE : variable codée en 0/1. Elle vaut « 0 » si le chien n'était pas de race croisée Golden/Labrador, « 1 » s'il l'était.

DECES : variable codée en 0/1. Elle vaut « 0 » si le chien était toujours en vie à la fin de l'étude, « 1 » s'il était décédé au cours de l'étude.

DECES_3_ANS : variable codée en 0/1. Elle vaut « 0 » si le chien était toujours en vie 3 ans après l'inclusion dans l'étude, « 1 » s'il était décédé dans les 3 ans après J0.

DEMARCHE_ANORMALE : variable codée en 0/1, « 0 » si le chien avait une démarche normale, « 1 » s'il avait une démarche anormale.

FEMELLE : variable codée en 0/1, « 0 » si le chien était un mâle, « 1 » s'il était une femelle.

GOLDEN : variable codée en 0/1. Elle vaut « 0 » si le chien n'était pas de race Golden, « 1 » s'il l'était.

HYPER_CHOLES : variable codée en 0/1. Elle vaut « 0 » si le chien ne présentait pas d'hypercholestérolémie, et « 1 » s'il en présentait une.

LABRADOR : variable codée en 0/1. Elle vaut « 0 » si le chien n'était pas de race Labrador, « 1 » s'il l'était.

OBESE : variable codée en 0/1. Elle vaut « 0 » si le chien n'était pas obèse, « 1 » s'il l'était.

RACE_4CL : variable codée en 0/1/2/3. Elle vaut « 0 » pour les chiens de race Golden, « 1 » pour la race Labrador, « 2 » pour la race croisée Golden/Labrador, « 3 » pour une autre race.

UREE : variable correspondant à la concentration en urée, en g/L.

UREE_4CL : variable codée en 0/1/2/3 à partir des quartiles de la variable UREE. Cette variable vaut « 0 » pour les chiens avec une concentration en urée < 0,24 g/L, « 1 » pour les chiens avec une concentration en urée comprise entre 0,24 g/L et 0,28 g/L (exclu), « 2 » pour les chiens avec une concentration en urée comprise entre 0,28 g/L et 0,33 g/L (exclu), et « 3 » pour les chiens avec une concentration en urée supérieure ou égale à 0,33 g/L.

SURVIE : variable correspondant au délai entre la date du J0 et soit la date de fin d'étude pour les chiens toujours en vie à la fin de l'étude soit la date de décès pour les chiens décédés, exprimée en années (avec un chiffre après la virgule).

II. Avant de commencer à réaliser des analyses statistiques

1. Connaissances théoriques indispensables en biostatistique et en épidémiologie

Ce guide va vous apprendre à *réaliser* certaines analyses statistiques. Mais les tenants (pourquoi faire ou telle analyse statistique) et les aboutissants (ce que l'on peut inférer, ou non, à partir des résultats d'une analyse statistique) ne seront pas traités dans ce guide. C'est la raison pour laquelle il est indispensable d'avoir acquis les connaissances théoriques en biostatistique et en épidémiologie (analytique) avant de réaliser des analyses statistiques[2]. Vous pouvez lire notamment certains polycopiés de formation initiale que je dispense à l'EnvA, en fonction des analyses statistiques que vous souhaitez réaliser.

Pour les analyses statistiques des parties IV et V de ce guide, je vous suggère le polycopié de bases en biostatistique (disponible ici, sous « Bases en Biostatistique »).

[2] De la même façon mais dans un autre contexte, ce n'est pas tout d'avoir les clés d'une voiture qui est en face de vous prête à être conduite, il vous faut auparavant avoir votre permis de conduire. Conduire des analyses statistiques nécessite un « permis de conduire des analyses statistiques », ce qui correspond à l'acquisition des connaissances théoriques en biostatistique et en épidémiologie.

Pour l'analyse de survie (parties VII et VIII.E), je vous suggère le polycopié d'introduction à l'analyse de survie (disponible ici, sous « Analyse de survie »).

Pour la réalisation de modèles de régression (partie VIII), je vous suggère le polycopié d'épidémiologie clinique (disponible ici, sous « Epidémiologie clinique »).

2. Définitions de « critère de jugement », d' « exposition », et de « variable »

Le « critère de jugement »[3] (abrégé « CdJ » à partir de maintenant dans toute la suite de ce guide) est l'état de santé que l'on étudie seul (de façon descriptive), ou bien dont on étudie l'association avec une ou plusieurs « expositions ».

Une « exposition » est une caractéristique intrinsèque d'un animal (âge, sexe, race, concentration en urée, …), ou extrinsèque (environnement, traitements reçus, …), qui ne soit pas le CdJ étudié.

Une « variable » représente une caractéristique d'un individu dans le fichier de données. Par exemple, si l'on veut savoir si, parmi des chiens inclus dans une étude, une alimentation de type humide (*versus* sèche) est associée à la présence d'une obstruction urétérale, l' « exposition » est le type d'alimentation (humide *versus* sèche) et le « CdJ » est la présence (*versus* absence) d'une obstruction urétérale. Dans le fichier de données, l'exposition sera représentée, par exemple, par la variable NOURRITURE_HUMIDE, et le CdJ sera représenté, par exemple, par la variable OBSTRUC_URETERALE. Dans ce guide, j'écrirai en majuscule, sans guillemet, et en écriture normale (c'est-à-dire ni en gras, ni en italique), le nom des variables utilisées. Dans toute la suite de ce guide, je ne vais quasiment plus parler que de « variable » (sauf exception), même si parfois, le terme « exposition » aurait été plus pertinent.

3. Structure d'un fichier de données pour analyses statistiques

Avant une analyse statistique, un fichier de données doit être structuré de façon rigoureuse pour ensuite conduire des analyses statistiques sur ces données. Pour vérifier cette structure, je vous recommande vivement de lire le document « Comment structurer un fichier de données Excel® avant analyses statistiques » en cliquant ici, dans la partie intitulée « Collecte des données d'une enquête épidémiologique et structure d'un fichier de données ». Vous y verrez notamment que le fichier de données doit comporter le nom des variables sur la première ligne, et les individus doivent être présentés en ligne. De plus, si une donnée est manquante pour une variable qui sera utilisée dans les analyses statistiques, il faut laisser la case vide (absolument vide, c'est-à-dire sans espace, ni quoi que ce soit d'autre) dans Excel®.

Pour être analysable statistiquement, je vous recommande fortement que la variable soit numérique. C'est-à-dire qu'elle doit être renseignée pour chaque individu sous forme d'un nombre. Ce nombre affecté à chaque individu varie selon le type de variable (par exemple, « 0 » ou « 1 » pour une variable binaire). SAS® peut quand même réaliser certaines analyses statistiques même si la variable est « alphanumérique » (c'est-à-dire que l'information, par individu, est renseignée en utilisant des lettres, comme par exemple « Oui » ou « Non »), mais encore une fois, je ne le recommande pas.

Le nettoyage d'une base de données que vous avez par ailleurs créée ou bien que l'on vous a transmise peut prendre des heures, littéralement parlant. (« Nettoyer » ici signifie « faire de telle sorte que la base de données vérifie *tous* les critères cités dans le document « Comment structurer un fichier de données Excel® avant analyses statistiques » cité ci-dessus.)

[3] « outcome » en anglais

4. Lien entre les termes « ligne », « observation », et « individu »

Comme je l'ai écrit ci-dessus, un fichier de données doit contenir une ligne par « individu », avec le nom des variables sur la 1ère ligne du fichier de données. Ainsi, si N est le nombre d'individus, le fichier Excel® contiendra N+1 lignes remplies. Attention, ce que j'entends par « individu » est « l'unité statistique » qui servira aux analyses statistiques. Par exemple, si le sang d'un animal est régulièrement prélevé au cours du temps, et si l'on souhaite modéliser l'évolution de la concentration d'un paramètre sanguin au cours du temps à partir de prélèvements sanguins de plusieurs animaux, l'unité statistique sera le prélèvement sanguin, et chaque ligne du fichier de données correspondra au prélèvement i d'un animal j.

Il se trouve que SAS®, en anglais, appelle « observation » l'unité statistique. S'il y a N+1 lignes dans le fichier de données qui sera importé dans SAS® avec le nom des variables sur la 1ère ligne du fichier de données, SAS® importera N « observations ».

Dans ce guide, je ne vais pas traiter la situation où l'on collecte plusieurs fois au cours du temps des données sur un même animal : le fichier de données comprend les informations collectées seulement à J0, auprès de 99 chiens. Ainsi, dans les résultats présentés, une « observation » SAS® correspondra à un chien.

Dans ce guide, j'utiliserai le terme d' « individu » pour parler du cas général (« individu » devra alors être interprété comme « unité statistique »), et j'utiliserai le terme de « chien » lorsque je parlerai des individus du fichier de données utilisé pour le guide.

5. Quatre types de variables numériques

a) Variable « binaire »

Une variable binaire est une variable à deux classes (ou « modalités »). On peut citer comme exemple la variable correspondant au sexe d'un animal (mâle ou femelle) ou celle correspondant à la présence (*versus* absence) d'une maladie. Par convention, il est recommandé de coder en 0/1 dans le fichier de données une variable binaire (et en l'occurrence, pour interpréter les résultats de SAS®, je vous recommande très fortement de coder les variables binaires en 0/1). Le nom d'une variable binaire devrait être le nom de la classe pour laquelle « 1 » a été attribué (c'est un conseil que je vous donne pour grandement faciliter l'interprétation des résultats fournis par SAS®). Par exemple, si pour le sexe de l'animal, dans le fichier de données, « 1 » a été attribué aux femelles et « 0 » aux mâles, la variable devrait être nommée « Femelle » dans le fichier de données. Dans le cas d'une exposition binaire, la catégorie « exposée » est celle pour laquelle « 1 » a été attribuée à la variable correspondante, et la catégorie « non exposée » est celle pour laquelle « 0 » a été attribuée à cette même variable. De même, dans le cas d'un CdJ binaire, je vous recommande d'attribuer la valeur de « 1 » à la variable correspondant au CdJ pour les individus ayant présenté le CdJ, et « 0 » pour ceux qui ne l'ont pas présenté.

Dans le fichier de données, les variables suivantes sont binaires : AUTRE_RACE, CROISEE, DECES, DECES_3_ANS, DEMARCHE_ANORMALE, FEMELLE, GOLDEN, HYPER_CHOLES, LABRADOR, et OBESE.

b) Variable « qualitative nominale »

Une variable qualitative nominale est une variable avec trois classes ou plus, chacune des classes n'étant *a priori* pas ordonnée les unes par rapport aux autres. Le codage d'une variable qualitative nominale peut être en 0/1/2/etc. ou bien en 1/2/3/etc.

Dans le fichier de données, la variable RACE_4CL est une variable qualitative nominale.

c) *Variable « qualitative ordinale »*

Une variable qualitative ordinale est une variable avec trois classes ou plus, chacune des classes étant ordonnée les unes par rapport aux autres. Une variable qualitative ordinale peut être *originellement* qualitative nominale (par exemple, la variable correspondant aux réponses « jamais », « parfois », « souvent », « très souvent » à une question). Cela dit, le plus souvent, une variable qualitative ordinale provient d'une variable initialement quantitative (par exemple, la variable en quatre classes suivante, correspondant au temps passé par semaine à jouer avec son chien : 0-30 min, 30 min – 2h, 2h – 3h, et > 3h).

Dans le fichier de données, les variables suivantes sont qualitatives ordinales : AGE_4CL, CHOLES_3CL, et UREE_4CL.

d) *Variable « quantitative »*

Une variable quantitative est une variable continue avec (potentiellement) au moins un chiffre après la virgule (chiffre après la virgule existant, ou bien possible si l'instrument de mesure était idéalement très précis) ou bien une variable discrète représentant un dénombrement.

Dans le fichier de données, les variables suivantes sont quantitatives : AGE, ALAT, CREAT, UREE, et SURVIE.

III. Import dans SAS® de données d'un fichier Excel® et vérification de l'import

A. Import des données

Comme vous l'avez vu dans l'une des vidéos que je vous ai fortement recommandé de visionner, pour importer dans SAS® Studio un fichier de données sous Excel®, vous devez tout d'abord l'avoir « uploadé » dans votre dossier de travail (cf. Figure 4.a). Une fois cet « upload » réalisé, vous devez utiliser la procédure PROC IMPORT ci-dessous, en remplaçant ce qui est écrit en italique ci-dessous.

```
PROC IMPORT OUT = Nom_du_fichier_pour_analyses_avec_SAS
DATAFILE = "chemin pour accéder à votre fichier uploadé/fichier Excel.xlsx" DBMS =
XLSX REPLACE;
SHEET = "Nom de l'onglet contenant les données à importer";
GETNAMES = YES;
RUN;
```

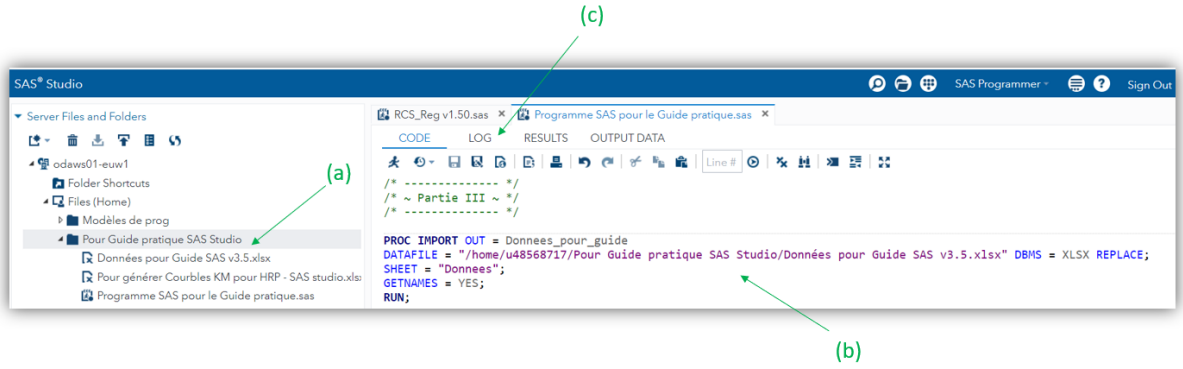


Figure 4

Vous pouvez voir (cf. Figure 4.b) que j'ai choisi d'appeler « Donnees_pour_guide » le fichier de données qui va servir aux analyses statistiques sous SAS®, importé à partir du fichier de données Excel®

« Données pour Guide SAS v3.5.xlsx » que j'avais auparavant « uploadé » dans SAS® Studio. (Vous vous rendez compte que « uploader » un fichier Excel® ne correspond pas à « importer un fichier Excel® ».) Une fois l'import réalisé à l'aide de la procédure PROC IMPORT, je vous recommande de cliquer sur la fenêtre LOG (cf. flèche (c) sur la Figure 4), qui indique combien d'individus et combien de variables le fichier Excel® importé dans SAS® contient (cf. la Figure 5). Je vous recommande ainsi de vérifier que les nombres pointés par les deux flèches sur la Figure 5 correspondent bien au nombre d'individus présents dans le fichier de données et au nombre de variables du fichier de données (ici, respectivement 99 chiens et 19 variables).

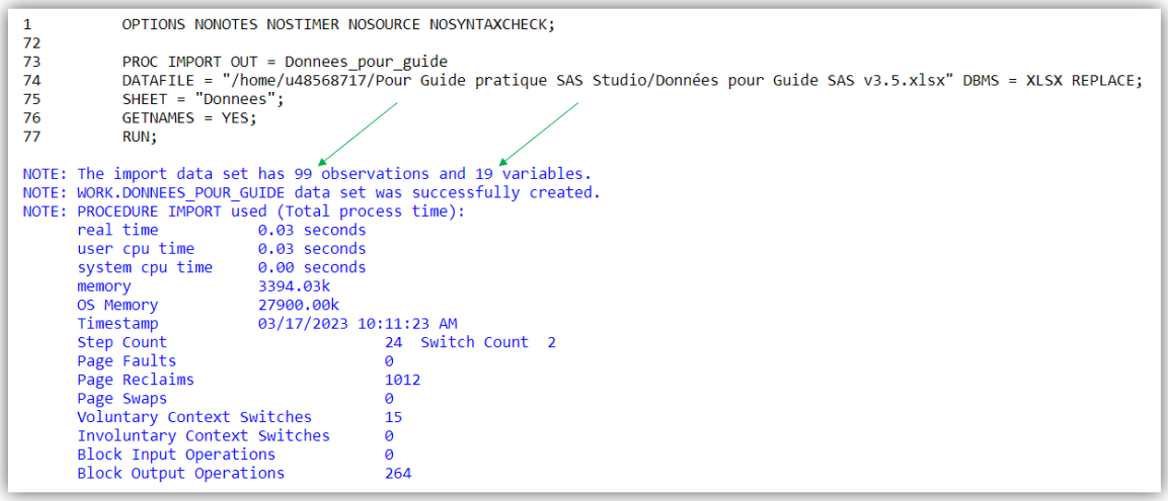


Figure 5

B. Vérification que les données sont prêtes à être analysées

Avant de commencer à analyser vos données, vous devez vérifiées qu'elles sont … analysables ! Notamment, dans la mesure où je vous recommande fortement de ne faire des analyses statistiques que sur des variables numériques, vous devez vérifier que ces variables *a priori* numériques sont effectivement reconnues comme tel par SAS®.

Pour réaliser cette vérification, je vous recommande la procédure PROC CONTENTS suivante, dont les lignes de programme se trouvent ci-dessous.

```
PROC CONTENTS DATA = Donnees_pour_guide;
RUN;
```

Le résultat des lignes de programme ci-dessus se trouve sur la Figure 6.

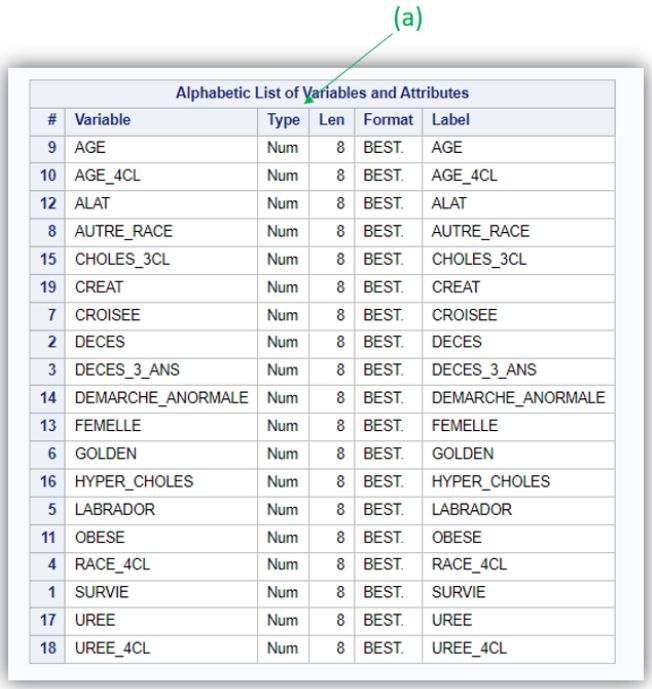
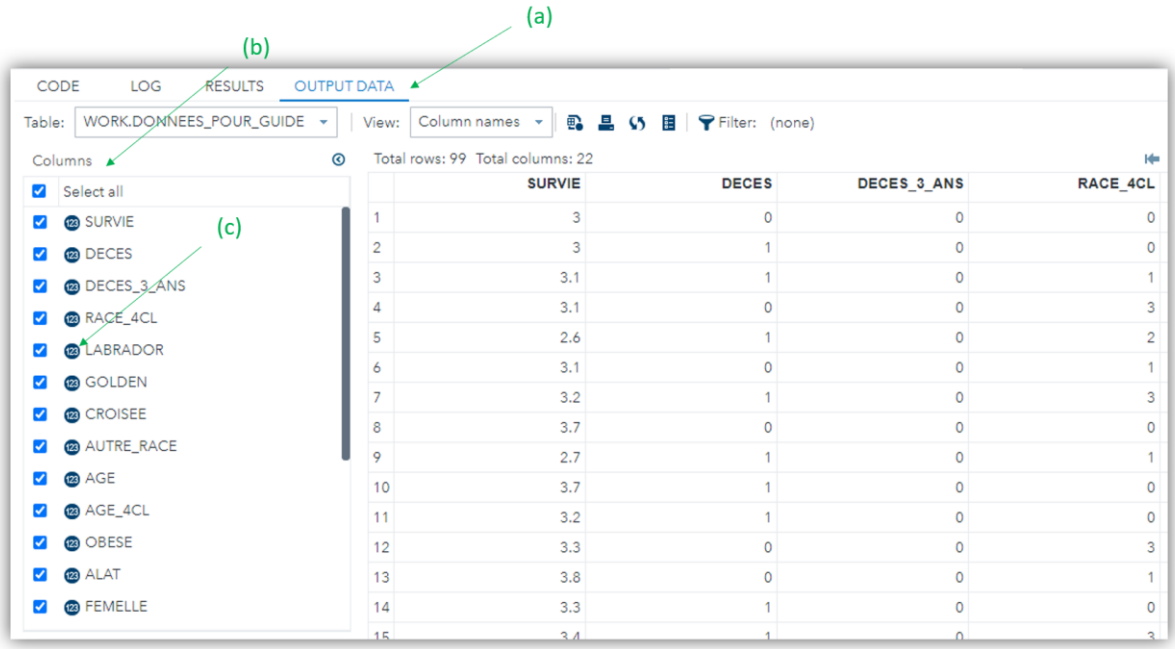


Figure 6

Les variables *a priori* numériques doivent avoir « Num » dans la colonne « Type » (cf. flèche (a) sur la Figure 6). Si « Char » est indiqué dans cette colonne « Type » pour une variable *a priori* numérique, c'est qu'un caractère alphanumérique (un espace, un accent, une lettre, un signe de ponctuation, ou toute autre chose qu'un chiffre ou qu'une case vide) s'est glissé quelque part sur une des lignes de la colonne correspondant à cette variable. Dans cette situation-là, vous devez ouvrir le fichier Excel® de vos données, corriger l'erreur, l'enregistrer puis le fermer, l' « uploader » à nouveau dans SAS®, et enfin l'importer à nouveau avec la procédure PROC IMPORT.

Lorsque vous faites tourner la procédure PROC IMPORT, SAS® va ouvrir spontanément la fenêtre « OUTPUT DATA (cf. Figure 7.a) qui est le fichier de données importé dans SAS®. Ainsi, une autre façon de vérifier vos données (mais qui ne doit pas se substituer à la réalisation de la PROC CONTENTS, car plus rigoureuse) est de parcourir les variables du fichier de données dans la colonne « Columns » (cf. Figure 7.b), de cocher la case « Select all » si elle n'est pas déjà cochée, et de vérifier qu'il y a bien le symbole « 123 » cerclé de bleu à gauche du nom de chaque variable *a priori* numérique (cf. Figure 7.c pour la variable LABRADOR par exemple).



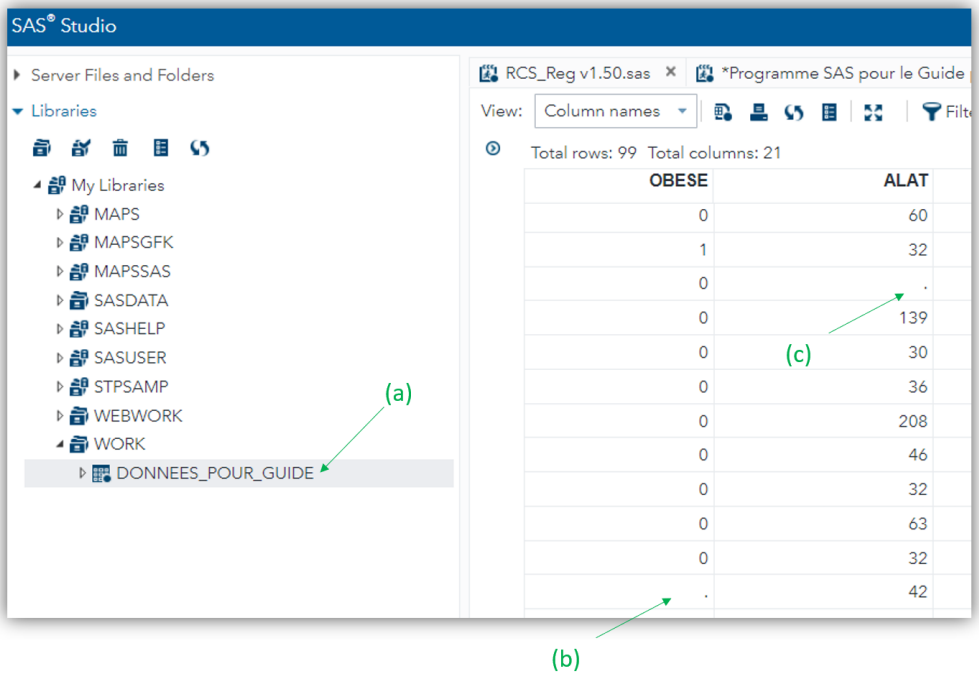
	SURVIE	DECES	DECES_3_ANS	RACE_4CL
1	3	0	0	0
2	3	1	0	0
3	3.1	1	0	1
4	3.1	0	0	3
5	2.6	1	0	2
6	3.1	0	0	1
7	3.2	1	0	3
8	3.7	0	0	0
9	2.7	1	0	1
10	3.7	1	0	0
11	3.2	1	0	0
12	3.3	0	0	3
13	3.8	0	0	1
14	3.3	1	0	0
15	3.4	1	0	3

Figure 7

IV. Statistique descriptive et association statistique entre deux variables

A. Les données manquantes pour une variable numérique

Avant même de commencer à analyser statistiquement un fichier de données, vous devez savoir que pour toutes les variables numériques, SAS® attribue le signe « . » dans le fichier de données importé. Ainsi, lorsque l'on double-clique sur le fichier de données de travail (cf. Figure 8.a), on peut voir un « . » pour les variables OBESE et ALAT (cf. flèches (b) et (c) sur la Figure 8).



OBESE	ALAT
0	60
1	32
0	.
0	139
0	30
0	36
0	208
0	46
0	32
0	63
0	32
.	42

Figure 8

Ensuite, il se trouve que dans SAS®, une donnée manquante pour une variable numérique a … une valeur, et elle vaut -∞. Cette information est très importante notamment quand on souhaitera sélectionner des individus sur des valeurs inférieures à un certain seuil d'une variable quantitative.

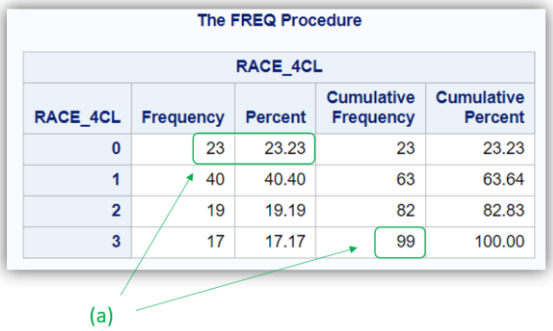
B. Statistiques descriptives

1. Décrire une variable binaire ou qualitative

Pour décrire une variable binaire ou qualitative, il faut dresser un « tableau de fréquences », qui est un tableau qui présente les effectifs et les pourcentages pour chaque classe d'une variable binaire ou qualitative. Je vais prendre pour l'exemple la variable qualitative nominale RACE_4CL en quatre classes. Pour obtenir le tableau de fréquences, je vous suggère d'utiliser la procédure PROC FREQ, dont les lignes de programme sont celles ci-dessous.

```
PROC FREQ DATA = Donnees_pour_guide;
TABLES RACE_4CL;
RUN;
```

Le résultat des lignes de programme ci-dessus est présenté sur la Figure 9.



The FREQ Procedure

RACE_4CL

RACE_4CL	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	23	23.23	23	23.23
1	40	40.40	63	63.64
2	19	19.19	82	82.83
3	17	17.17	99	100.00

Figure 9

On peut lire (cf. Figure 9.a) que les chiens de race Golden (RACE_4CL = 0) sont au nombre de 23 (colonne « Frequency »), ce qui représente 23,23% (colonne « Percent ») de l'échantillon des 99 chiens (dernière ligne dans la colonne « Cumulative Frequency »).

Notez que l'on peut demander à SAS® de décrire plusieurs variables binaires ou qualitatives à la fois. Pour cela, on liste les variables à étudier les unes à la suite des autres, séparées par un espace, dans l'instruction « TABLES ». Les lignes de programme ci-dessous demandent à SAS® de décrire d'un seul coup les variables RACE_4CL et OBESE. L'instruction « / MISSING » permet aussi de demander à SAS® de faire apparaître les données manquantes éventuelles.

```
PROC FREQ DATA = Donnees_pour_guide;
TABLES RACE_4CL OBESE / MISSING;
RUN;
```

Les résultats des lignes de programme ci-dessus sont présentés sur la Figure 10.

(a)

The FREQ Procedure

RACE_4CL

RACE_4CL	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	23	23.23	23	23.23
1	40	40.40	63	63.64
2	19	19.19	82	82.83
3	17	17.17	99	100.00

(b)

OBESE

OBESE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	3	3.03	3	3.03
0	79	79.80	82	82.83
1	17	17.17	99	100.00

Figure 10

On peut voir qu'il n'y a pas de donnée manquante pour la variable RACE_4CL (cf. Figure 10.a) mais qu'il y en a 3 pour la variable OBESE (cf. Figure 10.b).

2. Décrire une variable quantitative

a) Description au sein de l'échantillon en entier

Pour vous montrer comment obtenir avec SAS® les différents indicateurs statistiques décrivant une variable quantitative, je vais prendre pour l'exemple la variable quantitative ALAT. Deux procédures peuvent être utilisées : PROC UNIVARIATE et PROC MEANS. (J'ai une préférence pour la PROC MEANS, qui ne fournit que ce qu'on lui demande.)

Je vais commencer par vous décrire les résultats produits par la procédure PROC UNIVARIATE, dont les lignes de programme sont ci-dessous.

```
PROC UNIVARIATE DATA = Donnees_pour_guide;
VAR ALAT;
RUN;
```

Une sélection des nombreux résultats des lignes de programme ci-dessus se trouve sur la Figure 11.

Le bloc « Basic Statistical Measures » fournit entre autres la moyenne, la médiane, et la Standard Deviation (SD) de la variable quantitative ALAT (cf. Figure 11.a). Le bloc « Quantiles (Definition 5) » fournit entre autres les valeurs minimales et maximales, et les 1er et 3ème quartiles (cf. Figure 11.b). Le bloc « Extreme Observations » fournit dans les colonnes « Value » les cinq valeurs les plus faibles (cf. Figure 11.c) et les cinq valeurs les plus élevées (cf. Figure 11.d). Le bloc « Missing Values » fournit le nombre de données manquantes (cf. Figure 11.e ; ici, une valeur de la concentration en ALAT manque dans le fichier de données).

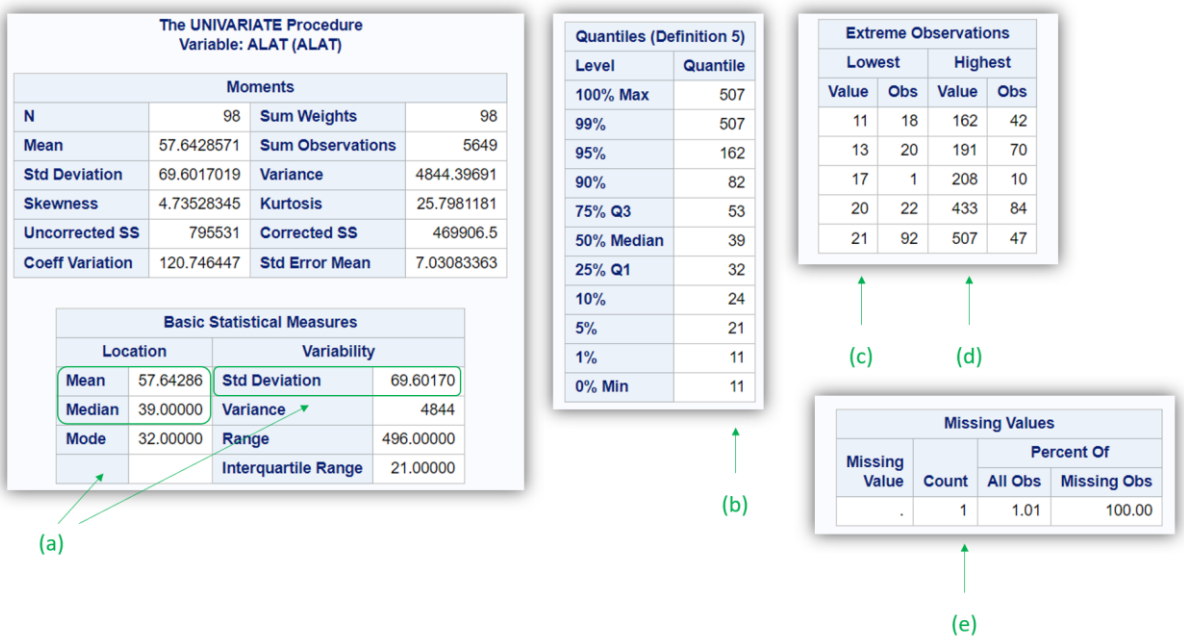


Figure 11

Comme pour la procédure PROC FREQ, il est possible de demander de décrire d'un seul coup plusieurs variables quantitatives à la fois, en listant les variables quantitatives à décrire séparées par un espace. Les lignes de programme ci-dessous demandent à SAS® de décrire d'un seul coup les variables ALAT et UREE.

```
PROC UNIVARIATE DATA = Donnees_pour_guide;
VAR ALAT UREE;
RUN;
```

La procédure PROC MEANS décrit une (ou plusieurs) variable(s) quantitative(s) avec les indicateurs souhaités (qu'il faut taper). Supposons que l'on veuille décrire les variables ALAT et UREE en fournissant leur moyenne, SD, médiane, valeurs minimales et maximales, 1er et 3ème quartiles, et les nombres de données manquantes et non manquantes pour chacune de ces deux variables. Les lignes de programme ci-dessous permettent de réaliser l'analyse souhaitée.

```
PROC MEANS DATA = Donnees_pour_guide MEAN STD MEDIAN MIN MAX P25 P75 NMISS N;
VAR ALAT UREE;
RUN;
```

Le résultat des lignes de programme ci-dessus est présenté sur la Figure 12.

The MEANS Procedure

Variable	Label	Mean	Std Dev	Median	Minimum	Maximum	25th Pctl	75th Pctl	N Miss	N
ALAT	ALAT	57.6428571	69.6017019	39.0000000	11.0000000	507.0000000	32.0000000	53.0000000	1	98
UREE	UREE	0.2915354	0.0944838	0.2800000	0.1300000	0.7500000	0.2400000	0.3300000	0	99

Figure 12

Pour la concentration en ALAT, on lit que la moyenne vaut 57,64 UI/L, la SD vaut 69,60 UI/L, la médiane vaut 39,00 UI/L, la valeur minimale vaut 11,00 UI/L, la valeur maximale vaut 507,00 UI/L, les 1er et 3ème quartiles valent respectivement 32,00 UI/Let 53,00 UI/L. On peut y lire qu'il y a une donnée manquante sur la concentration en ALAT (« 1 » dans la colonne « N Miss »), et aucune donnée manquante pour la concentration en UREE. La colonne « N » indique le nombre de valeurs qui ont été utilisées pour calculer les indicateurs demandés (⇔ nombre de données non manquantes). Notamment, puisqu'une donnée manque pour la concentration en ALAT, les indicateurs pour cette variable-là ont été calculés parmi les 98 chiens de l'échantillon pour lesquels la concentration en ALAT était renseignée.

b) *Description selon les classes d'une variable binaire ou qualitative*

Si l'on souhaite décrire une variable quantitative selon les classes d'une variable binaire ou qualitative, il suffit d'ajouter dans la procédure (PROC UNIVARIATE ou PROC MEANS) l'instruction « CLASS ». Je vous recommande là encore l'utilisation de la procédure PROC MEANS. Les lignes de programme ci-dessous permettent de décrire la variable ALAT (à l'aide de la moyenne, de la médiane, du minimum, du maximum, et des 1er et 3ème quartiles) selon les classes de la variable RACE_4CL.

```
PROC MEANS DATA = Donnees_pour_guide MEAN MEDIAN MIN MAX P25 P75;
CLASS RACE_4CL;
VAR ALAT;
RUN;
```

Le résultat des lignes de programme ci-dessus est présenté sur la Figure 13.

The MEANS Procedure

Analysis Variable : ALAT ALAT							
RACE_4CL	N Obs	Mean	Median	Minimum	Maximum	25th Pctl	75th Pctl
0	23	88.0000000	43.0000000	17.0000000	507.0000000	35.0000000	68.0000000
1	40	46.0512821	36.0000000	11.0000000	191.0000000	29.0000000	52.0000000
2	19	33.5789474	32.0000000	13.0000000	56.0000000	27.0000000	39.0000000
3	17	70.0588235	60.0000000	22.0000000	162.0000000	42.0000000	81.0000000

Figure 13

C. Description graphique d'une ou plusieurs variables quantitatives

a) *Histogramme*

Pour dresser un histogramme d'une variable quantitative sous SAS®, je vous suggère d'utiliser la procédure PROC UNIVARIATE. Les lignes de programme ci-dessous permettent de dresser l'histogramme de la variable ALAT en ajoutant comme option la représentation graphique de la loi normale de moyenne la moyenne de la variable ALAT dans l'échantillon, et de SD la SD de la variable ALAT dans l'échantillon (respectivement 57,64 et 69,60 ; cf. Figure 12).

```
PROC UNIVARIATE DATA = Donnees_pour_guide NOPRINT;
VAR ALAT;
HISTOGRAM ALAT / NORMAL;
RUN;
```

La Figure 14 présente l'histogramme issu des lignes de programme ci-dessus. SAS® indique la moyenne et la SD de la loi normale représentée (cf. flèche (a) sur la Figure 14).

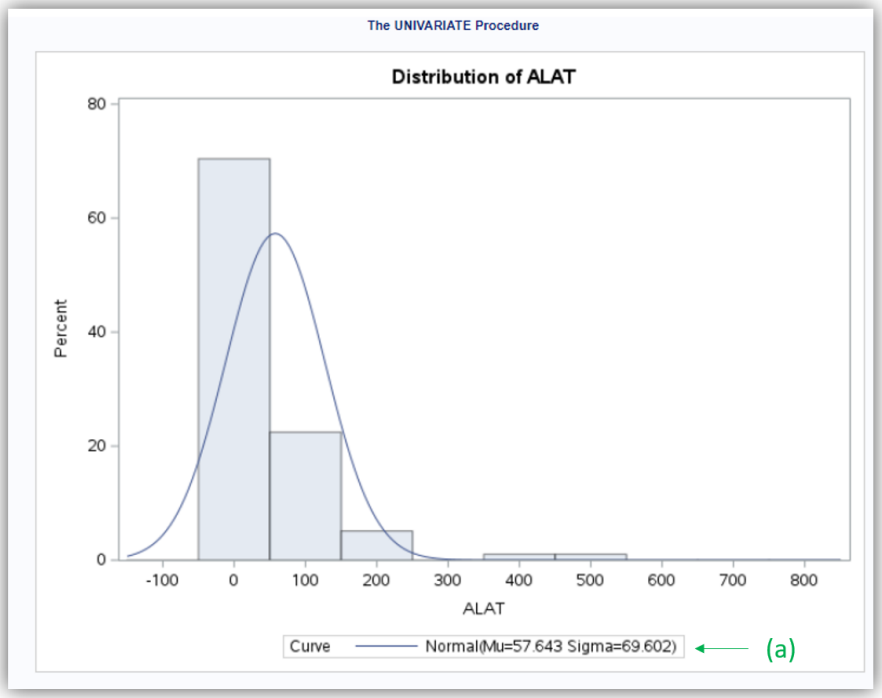


Figure 14

Il est possible de juxtaposer les histogrammes d'une variable quantitative selon les classes d'une variable binaire ou qualitative, là encore en utilisant l'instruction « CLASS ». Les lignes de programme ci-dessous permettent de dresser les histogrammes de la concentration en ALAT d'abord chez les chiens non obèses (OBESE = 0 ; Figure 15.a) puis chez les chiens obèses (OBESE = 1 ; Figure 15.b).

```
PROC UNIVARIATE DATA = Donnees_pour_guide NOPRINT;
CLASS OBESE;
VAR ALAT;
HISTOGRAM ALAT;
RUN;
```

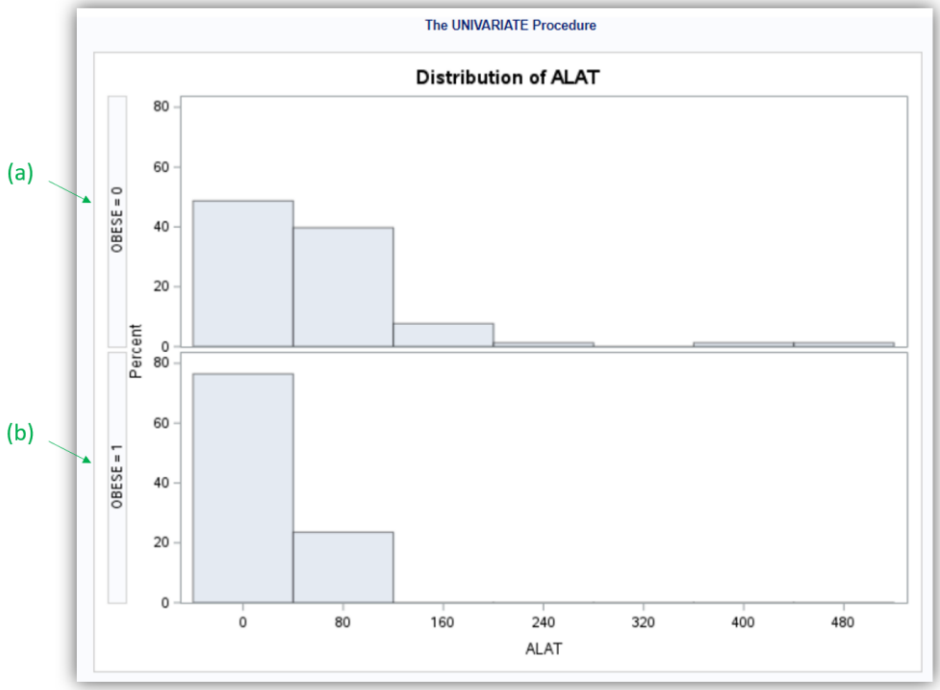


Figure 15

b) *Boite à moustaches (boxplot)*

La représentation graphique de boite à moustaches est intéressante lorsque l'on veut présenter la différence de distribution d'une variable quantitative selon les classes d'une variable binaire ou qualitative, en se focalisant sur les indicateurs de la distribution (moyenne, médiane, 1er et 3ème quartiles, …). La procédure PROC SGPLOT permet de réaliser des boites à moustaches. Les lignes de programme ci-dessous permettent de dresser les boites à moustaches pour la variable ALAT selon les différentes classes de la variable RACE_CL.

```
PROC SGPLOT DATA = Donnees_pour_guide;
VBOX ALAT / CATEGORY = RACE_4CL;
RUN;
```

Le résultat des lignes de programme ci-dessus est présenté sur la Figure 16.

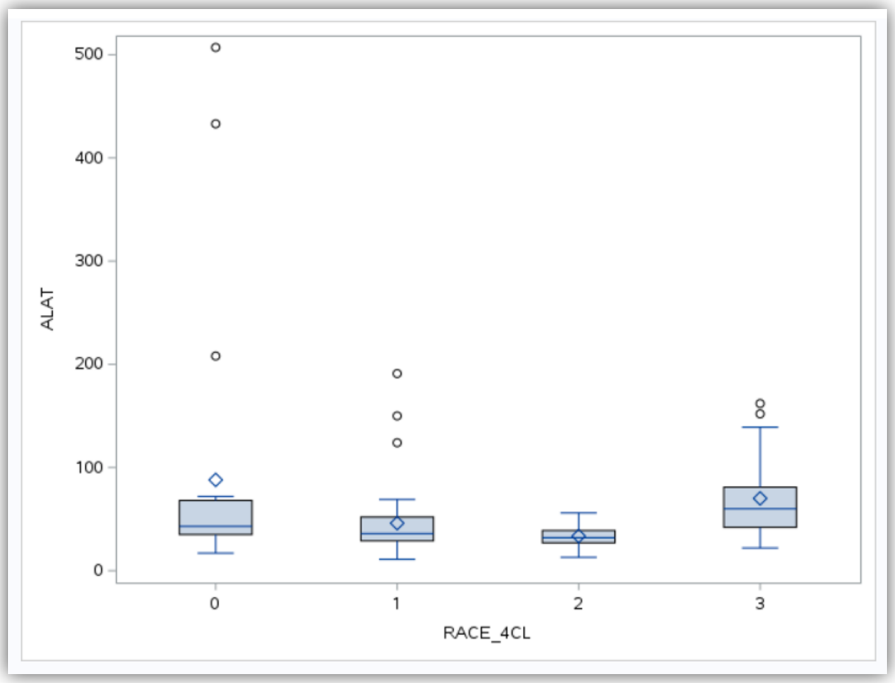


Figure 16

Pour interpréter correctement ces boites à moustaches, il faut savoir à quoi font référence les différentes parties d'une boite à moustaches dans SAS®. Cette information se trouve ici, ainsi que sur la Figure 17.

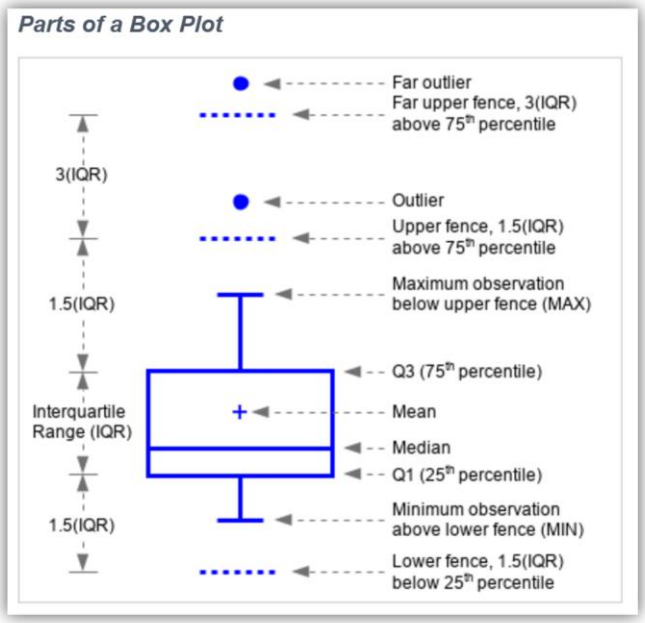


Figure 17

La petite différence entre la représentation graphique fournie par SAS® sur la Figure 16 et celle sur la Figure 17 concerne la représentation graphique de la moyenne : un losange dans SAS® Studio (Figure 16), un « + » sur la Figure 17. En somme, la Figure 16 est la représentation graphique des informations fournies sur la Figure 13 avec la PROC MEANS.

c) *Le nuage de points*

Le nuage de points est une méthode graphique pour explorer l'association entre deux variables quantitatives, et/ou pour identifier des valeurs aberrantes. La procédure PROC SGPLOT permet aussi de réaliser des nuages de points. Les lignes de programme ci-dessous permettent de dresser le nuage de points explorant l'association entre la concentration en créatinine (variable CREAT) et celle en urée (variable UREE), avec la concentration en créatinine sur l'axe des ordonnées et celle en urée sur l'axe des abscisses.

```
PROC SGPLOT DATA = Donnees_pour_guide;
SCATTER X = UREE Y = CREAT;
REG X = UREE Y = CREAT;
RUN;
```

Le résultat des lignes de programme ci-dessus est présenté sur la Figure 18. (Si vous ne souhaitez pas que la droite de régression soit représentée, il ne faut pas taper la 3ème ligne de programme ci-dessus.)

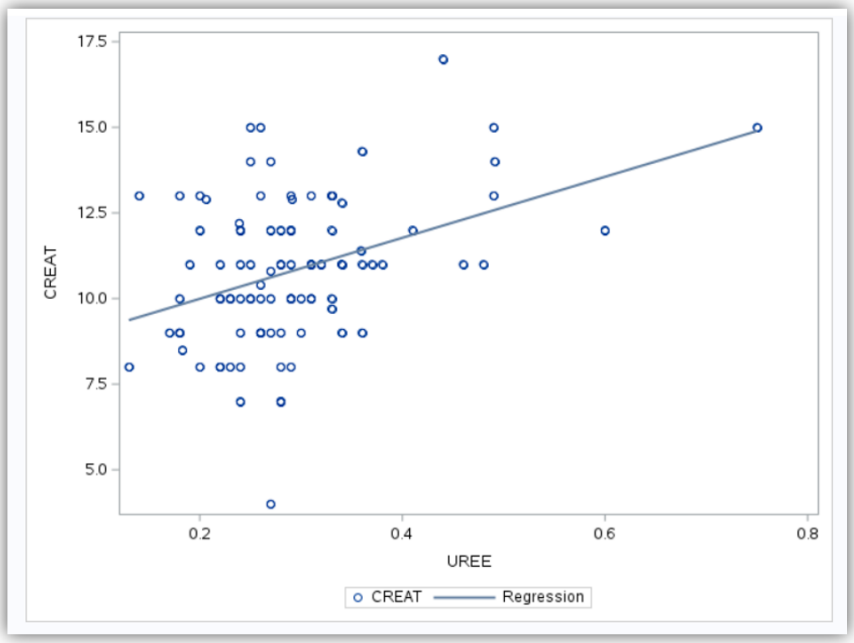


Figure 18

V. Tests statistiques pour tester l'association statistique entre deux variables

A. Association entre deux variables binaires ou qualitatives

1. Introduction

Le croisement de deux variables binaires ou qualitatives permet d'étudier l'association entre ces deux variables. Je ne reviendrai pas dans ce guide sur la façon de correctement lire un tableau croisant deux variables binaires ou qualitatives. Notamment, je ne reviendrai pas sur le fait de savoir faire la distinction entre les « bons » et les « mauvais » pourcentages à citer au moment de dire que ces pourcentages sont, ou ne sont pas, significativement différents (cf. sous-partie « Connaissances théoriques indispensables en biostatistique et en épidémiologie », page 8).

Le croisement de deux variables binaires ou qualitatives s'effectue à l'aide de la procédure PROC FREQ, comme vous allons le voir ci-dessous.

2. Association entre deux variables binaires

Je vais prendre pour l'exemple les deux variables binaires suivantes : FEMELLE et DECES_3_ANS. Les lignes de programme pour croiser ces deux variables et pour savoir si ces deux variables étaient significativement associées dans l'échantillon ($p \leq 0,05$) sont celles ci-dessous.

```
PROC FREQ DATA = Donnees_pour_guide;
TABLES FEMELLE * DECES_3_ANS / CHISQ EXPECTED FISHER;
RUN;
```

Sur la 2ème ligne de programme ci-dessus, le signe « * » permet de croiser deux variables. Les options « CHISQ », « EXPECTED », et « FISHER » permettent respectivement d'obtenir le résultat du test

statistique du Chi-2, de faire apparaître les effectifs attendus sous H_0, et de réaliser le test exact de Fisher. Les résultats des lignes de programme ci-dessus se trouvent sur la Figure 19.

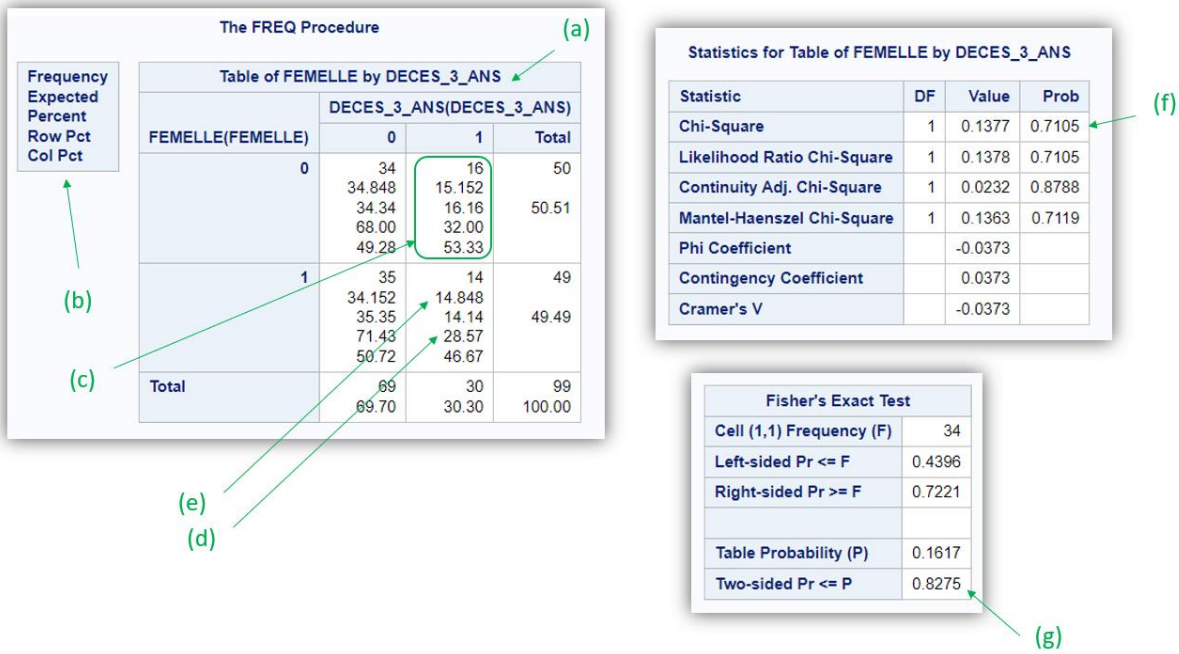


Figure 19

Je vais tout d'abord me focaliser sur le tableau d'effectifs et de pourcentages pointé par la flèche (a) sur la Figure 19. Pour savoir à quoi correspondent tous les chiffres présentés dans l'une des quatre cases du tableau, il faut lire ce qui est indiqué dans le petit tableau pointé par la flèche (b) sur la Figure 19 : les chiffres seront présentés dans cet ordre, du haut vers le bas, à savoir d'abord le nombre observé d'individus (« Frequency »), le nombre d'individus attendus sous H_0 (« Expected »), le pourcentage que représente le nombre observé d'individus par rapport à l'échantillon total (« Percent »), le pourcentage en ligne que représente le nombre observé d'individus (« Row Pct »), et enfin le pourcentage en colonne que représente le nombre observé d'individus (« Col Pct »).

Ainsi, sur la ligne « 0 » et la colonne « 1 », on peut lire (cf. Figure 19.c) qu'il y a 16 chiens qui sont des mâles (puisque FEMELLE = 0) *et* qui sont décédés dans les 3 ans (puisque DECES_3_ANS = 1). Le nombre attendu sous H_0 de chiens mâles décédés dans les 3 ans est égal à 15,15. Les 16 chiens mâles décédés dans les 3 ans représentent 16,16% de l'échantillon total des 99 chiens. Le pourcentage de chiens décédés dans les 3 ans parmi les 50 chiens mâles de l'échantillon est égal à 32,00% (16/50), et le pourcentage de chiens mâles parmi les 30 chiens décédés dans les 3 ans est égal à 53,33% (16/30).

Dans l'échantillon, le pourcentage de chiens décédés parmi les chiens femelles (28,57%, cf. Figure 19.d) était légèrement inférieur à celui parmi les chiens mâles (32,00%). La question est de savoir si ces deux pourcentages étaient significativement différents ou pas.

Dans la mesure où aucun des quatre effectifs attendus sous H_0 n'est inférieur à 5 (la plus petite valeur étant 14,85, cf. Figure 19.e), c'est le test du Chi-2, et non pas le test de Fisher, qu'il faut utiliser.

La valeur du degré de signification du test du Chi-2 se trouve dans le tableau « Statistics for Table of FEMELLE by DECES_3_ANS », ligne « Chi-Square », colonne « Prob » (cf. Figure 19.f) : $p = 0,71$. Puisque ce degré de signification est supérieur à 0,05 (seuil de risque d'erreur de 1ère espèce α, qui sera toujours fixé à 0,05 dans ce guide), les deux pourcentages cités ci-dessus (28,57% et 32,00%) n'étaient pas significativement différents. Si au moins l'un des quatre effectifs attendus sous H_0 avait été inférieur à 5, il aurait fallu lire de degré de signification du test de Fisher dans le tableau « Fisher's Exact Test », à la ligne « Two-sided Pr <= P » (cf. Figure 19.g).

3. Association entre une variable binaire et une variable qualitative

Je vais prendre pour l'exemple les deux variables suivantes : RACE_4CL et DECES_3_ANS. Les lignes de programme pour croiser ces deux variables et pour savoir si ces deux variables étaient significativement associées dans l'échantillon ($p \leq 0,05$) sont celles ci-dessous.

```
PROC FREQ DATA = Donnees_pour_guide;
TABLES RACE_4CL * DECES_3_ANS / CHISQ EXPECTED FISHER;
RUN;
```

Les résultats des lignes de programme ci-dessus se trouvent sur la Figure 20.

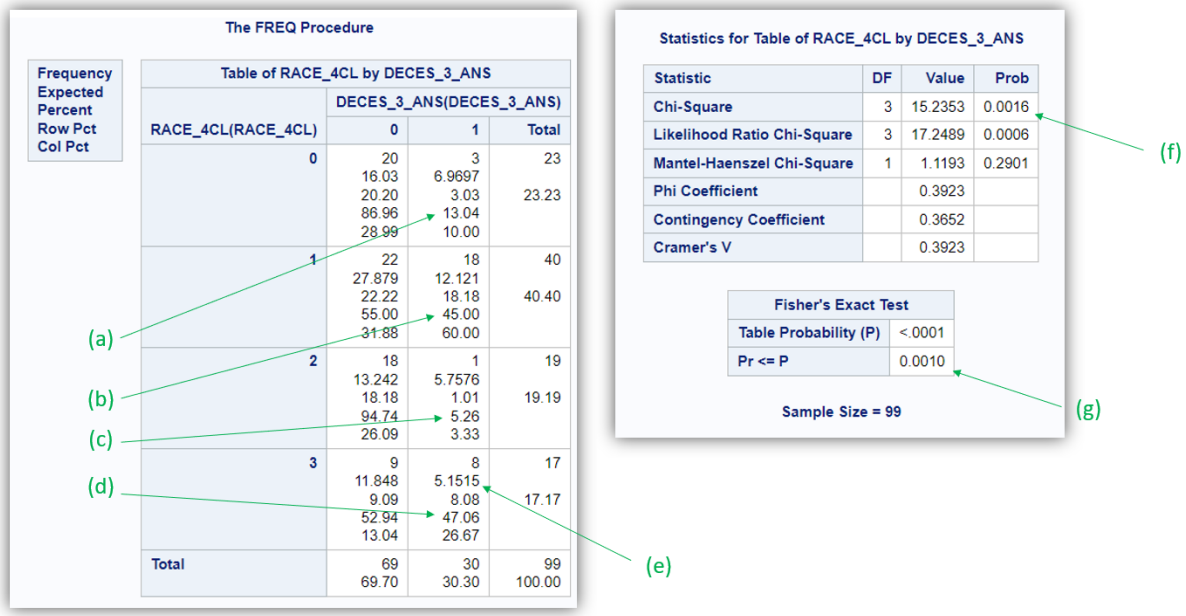


Figure 20

Vous pouvez déjà observer qu'il y a beaucoup moins de résultats statistiques que sur la Figure 20 ! Les pourcentages qui doivent être comparés pour savoir s'il existe une association entre la race et la présence d'un décès dans les 3 ans sont les pourcentages (en ligne) suivants : le pourcentage de chiens décédés parmi les chiens de race Golden (3/23 = 13,04% ; Figure 20.a), le pourcentage de chiens décédés parmi les chiens de race Labrador (18/40 = 45,00% ; Figure 20.b), le pourcentage de chiens décédés parmi les chiens de race croisée Golden/Labrador (1/19 = 5,26% ; Figure 20.c), et le pourcentage de chiens décédés parmi les chiens d'autre race (8/17 = 47,06% ; Figure 20.d).

Dans la mesure où aucun des huit effectifs attendus sous H_0 n'est inférieur à 5 (la plus petite valeur étant 5,15 ; cf. Figure 20.e), c'est le test du Chi-2 qu'il faut utiliser pour tester les quatre pourcentages cités ci-dessus. La valeur de son degré de signification est pointée par la flèche (f) sur la Figure 20 : p = 0,0016. (Si au moins l'un des huit effectifs attendus sous H_0 avait été inférieur à 5, alors il aurait fallu réaliser le test statistique de Fisher, et lire la valeur du degré de signification pointée par la flèche (g) sur la Figure 20.)

4. Association entre deux variables qualitatives

Cette situation produisant un tableau à plus de deux lignes et plus de deux colonnes, elle conduit à des résultats ininterprétables : les pourcentages à comparer, qui seraient testés par le test statistique du Chi-2 ou de Fisher, ne peuvent pas s'exprimer de façon claire et intelligible. Je ne fournirai donc aucun exemple d'une telle situation, et je vous invite plus que fortement à rendre binaire (au moins) une des deux variables lorsque vous souhaitez étudier l'association entre deux variables qualitatives. Par exemple, si vous souhaitiez savoir s'il existe une association entre la race (variable RACE_4CL en quatre

classes) et la cholestérolémie (variable CHOLES_3CL en trois classes), il aurait fallu soit recoder la variable RACE_4CL en une variable binaire, soit recoder la variable CHOLES_3CL en une variable binaire (soit bien entendu recoder de façon binaire ces deux variables !).

B. Association entre une variable binaire ou qualitative et une variable quantitative

1. Introduction

Dans cette situation-là, il faut soit comparer (puis tester) des moyennes (deux si la 1ère variable est binaire, ou trois ou plus si la 1ère variable est qualitative), soit comparer (puis tester) des médianes, selon la distribution de la variable quantitative.

Je vous rappelle les noms des tests statistiques dans les quatre situations suivantes (lorsque les individus sont indépendants) :

- Comparaison de deux moyennes (lorsque la distribution de la variable quantitative peut être considérée comme normale) → test de Student pour séries non appariées
- Comparaison de deux médianes (quelle que soit la distribution de la variable quantitative) → test de Mann-Whitney / Wilcoxon pour séries non appariées
- Comparaison de trois moyennes ou plus (lorsque la distribution de la variable quantitative peut être considérée comme normale) → test de l'ANOVA
- Comparaison de trois médianes ou plus (quelle que soit la distribution de la variable quantitative) → test de Kruskal-Wallis

Je vous incite fortement à calculer (grâce à SAS® bien sûr) les moyennes ou les médianes que *par la suite* vous aller tester. Il se trouve que trois des quatre procédures SAS® que nous allons voir ne fournissent pas les moyennes ou les médianes. Ainsi, la procédure PROC MEANS (cf. page 17) devra être utilisée préalablement.

Pour les exemples ci-dessous, je vais faire l'hypothèse que la distribution de la variable CREAT peut être considérée comme normale.

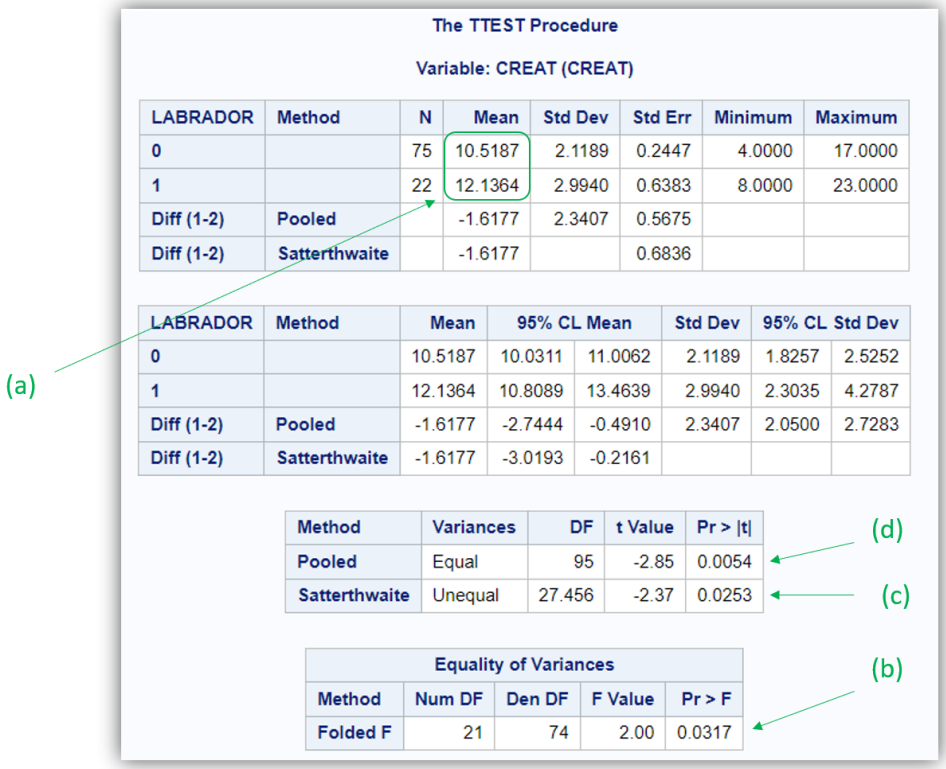
2. Comparaison de deux moyennes

Je vais prendre comme 1er exemple les variables LABRADOR et CREAT. La procédure PROC TTEST est la procédure permettant de réaliser un test de Student pour séries non appariées. Il se trouve que la procédure PROC TTEST fournit aussi les moyennes qui sont comparées puis testées (ce qui évite de réaliser préalablement une procédure PROC MEANS).

Les lignes de programme ci-dessous permettent d'obtenir la moyenne de la concentration en créatinine parmi les chiens qui ne sont pas de race Labrador et celle parmi les chiens qui sont de race Labrador, puis de les tester statistiquement à l'aide du test de Student pour séries non appariées.

```
PROC TTEST DATA = Donnees_pour_guide;
CLASS LABRADOR;
VAR CREAT;
RUN;
```

Les résultats des lignes de programme ci-dessus se trouvent sur la Figure 21.



The TTEST Procedure

Variable: CREAT (CREAT)

LABRADOR	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		75	10.5187	2.1189	0.2447	4.0000	17.0000
1		22	12.1364	2.9940	0.6383	8.0000	23.0000
Diff (1-2)	Pooled		-1.6177	2.3407	0.5675		
Diff (1-2)	Satterthwaite		-1.6177		0.6836		

LABRADOR	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		10.5187	10.0311	11.0062	2.1189	1.8257	2.5252
1		12.1364	10.8089	13.4639	2.9940	2.3035	4.2787
Diff (1-2)	Pooled	-1.6177	-2.7444	-0.4910	2.3407	2.0500	2.7283
Diff (1-2)	Satterthwaite	-1.6177	-3.0193	-0.2161			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	95	-2.85	0.0054
Satterthwaite	Unequal	27.456	-2.37	0.0253

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	21	74	2.00	0.0317

Figure 21

Ainsi, la moyenne de la concentration en créatinine parmi les chiens qui ne sont pas de race Labrador était égale à 10,52 mg/L, et celle parmi les chiens de race Labrador était égale à 12,14 mg/L (cf. Figure 21.a). Ensuite, pour savoir si ces deux moyennes étaient, ou n'étaient pas, significativement différentes, il faut savoir quel test de Student pour séries non appariées doit être réalisé : celui considérant que les variances sont voisines ou celui considérant que les variances ne sont pas voisines. La valeur du degré de signification pointée par la flèche (b) sur la Figure 21 provient d'un test statistique testant l'égalité des variances (p = 0,03). Il est inférieur à 0,05, donc la variance de la concentration en créatinine parmi les chiens de race Labrador ne peut pas être considérée comme voisine de celle parmi les chiens qui ne sont pas de race Labrador. Ainsi, le test de Student pour séries non appariées doit utiliser la méthode « Satterthwaite », et le degré de signification vaut, en arrondissant, 0,03 (cf. Figure 21.c). Ainsi, les deux moyennes de concentration en créatinine (10,52 mg/L et 12,14 mg/L) étaient significativement différentes. Si les deux variances avaient pu être considérées comme voisines (si la valeur du degré de signification pointée par la flèche (b) sur la Figure 21 avait été supérieure à 0,05), alors il aurait fallu utiliser la méthode « Pooled », et le degré de signification aurait été celui pointé par la flèche (d) sur la Figure 21 : p = 0,005 (en arrondissant).

3. Comparaison de deux médianes

Je vais continuer d'utiliser les variables LABRADOR et CREAT. La procédure PROC NPAR1WAY est une procédure permettant de réaliser, entre autres, un test de Mann-Whitney / Wilcoxon pour séries non appariées. Il se trouve que la procédure PROC NPAR1WAY ne fournit pas les médianes qui sont comparées puis testées. Ainsi, la procédure PROC MEANS doit être préalablement utilisée pour fournir les deux médianes qui vont être ensuite testées.

Les lignes de programme ci-dessous permettent d'abord d'obtenir les deux médianes (à l'aide de la procédure PROC MEANS), puis ensuite de tester la différence entre les deux médianes (à l'aide de la procédure PROC NPAR1WAY).

```
PROC MEANS DATA = Donnees_pour_guide MEDIAN;
CLASS LABRADOR;
VAR CREAT;
RUN;

PROC NPAR1WAY DATA = Donnees_pour_guide WILCOXON;
CLASS LABRADOR;
VAR CREAT;
RUN;
```

Les résultats des lignes de programme ci-dessus se trouvent sur la Figure 22.

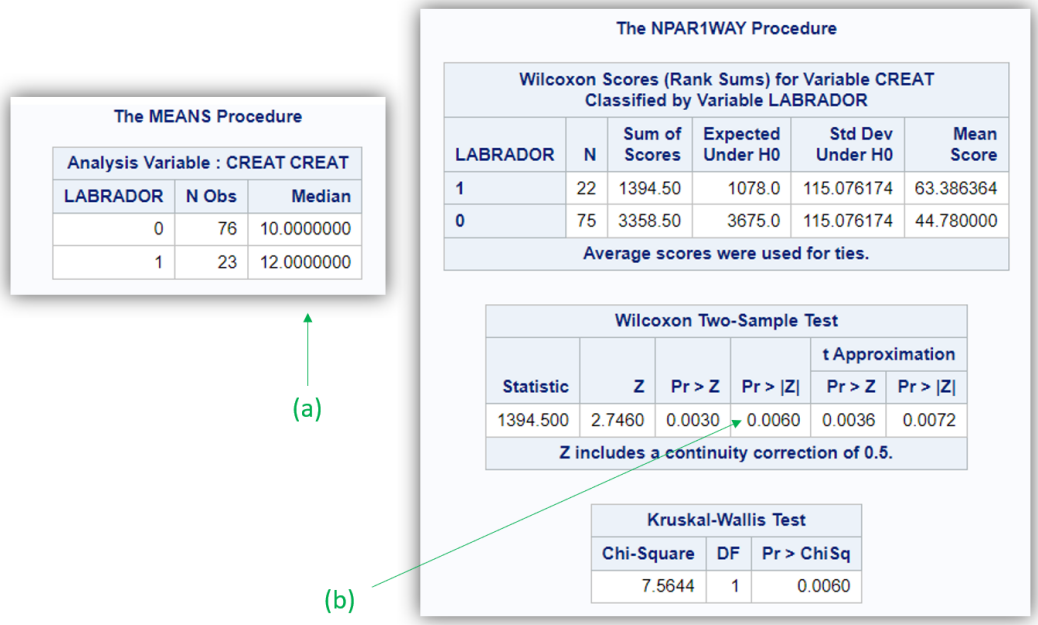


Figure 22

Ainsi, la médiane de la concentration en créatinine parmi les chiens qui ne sont pas de race Labrador était égale à 10,00 mg/L, et celle parmi les chiens de race Labrador était égale à 12,00 mg/L (cf. Figure 22.a). La valeur du degré de signification du test de Mann-Whitney / Wilcoxon pour séries non appariées est celle pointée par la flèche (b) sur la Figure 22 : p = 0,006. Ainsi, les deux médianes de concentration en créatinine (10,00 mg/L et 12,00 mg/L) étaient significativement différentes.

4. Comparaison de trois moyennes ou plus

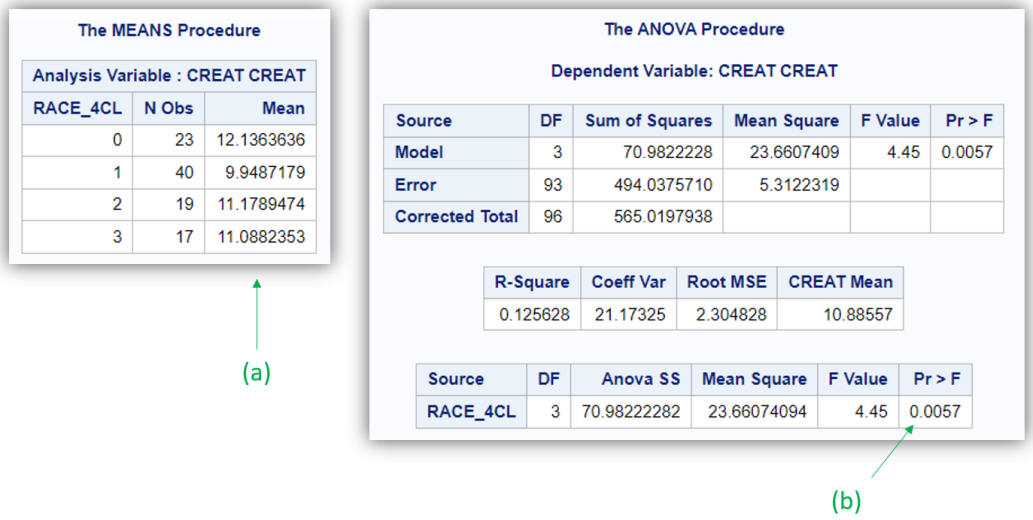
Je vais prendre comme 2ème exemple les variables RACE_4CL et CREAT. La procédure PROC ANOVA est une des procédures permettant de réaliser un test de l'ANOVA. Il se trouve que la procédure PROC ANOVA ne fournit pas les moyennes qui sont comparées puis testées. Ainsi, la procédure PROC MEANS doit être préalablement utilisée pour fournir les moyennes qui vont être ensuite testées.

Les lignes de programme ci-dessous permettent d'obtenir puis de tester les moyennes de la concentration en créatinine au sein de chacune des quatre races de chiens.

```
PROC MEANS DATA = Donnees_pour_guide MEAN;
CLASS RACE_4CL;
VAR CREAT;
RUN;

PROC ANOVA DATA = Donnees_pour_guide;
CLASS RACE_4CL;
MODEL CREAT = RACE_4CL;
RUN;
```

Les résultats des lignes de programme ci-dessus se trouvent sur la Figure 23.



The MEANS Procedure

Analysis Variable : CREAT CREAT		
RACE_4CL	N Obs	Mean
0	23	12.1363636
1	40	9.9487179
2	19	11.1789474
3	17	11.0882353

(a)

The ANOVA Procedure

Dependent Variable: CREAT CREAT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	70.9822228	23.6607409	4.45	0.0057
Error	93	494.0375710	5.3122319		
Corrected Total	96	565.0197938			

R-Square	Coeff Var	Root MSE	CREAT Mean
0.125628	21.17325	2.304828	10.88557

Source	DF	Anova SS	Mean Square	F Value	Pr > F
RACE_4CL	3	70.98222282	23.66074094	4.45	0.0057

(b)

Figure 23

Ainsi, les valeurs des moyennes de la concentration en créatinine parmi chacune des quatre races étaient de 12,14, 9,49, 11,18, et 11,09 mg/L respectivement pour les races « 0 », « 1 », « 2 », et « 3 » (cf. Figure 23.a). La valeur du degré de signification du test de l'ANOVA est celle pointée par la flèche (b) sur la Figure 23 : p = 0,006 (en arrondissant). Ainsi, les quatre moyennes de concentration en créatinine étaient significativement différentes.

5. Comparaison de trois médianes ou plus

Je vais continuer d'utiliser les variables RACE_4CL et CREAT. La procédure PROC NPAR1WAY est une procédure permettant de réaliser un test de Kruskal-Wallis. Comme vous l'avez vu avec la comparaison de deux médianes, la procédure PROC NAPR1WAY ne fournit pas les médianes qui sont comparées puis testées. Ainsi, la procédure PROC MEANS doit être préalablement utilisée pour fournir les médianes qui vont être ensuite testées.

Les lignes de programme ci-dessous permettent d'obtenir puis de tester les médianes de la concentration en créatinine au sein de chacune des quatre races de chiens.

```
PROC MEANS DATA = Donnees_pour_guide MEDIAN;
CLASS RACE_4CL;
VAR CREAT;
RUN;

PROC NPAR1WAY DATA = Donnees_pour_guide WILCOXON;
CLASS RACE_4CL;
VAR CREAT;
RUN;
```

Les résultats des lignes de programme ci-dessus se trouvent sur la Figure 24.

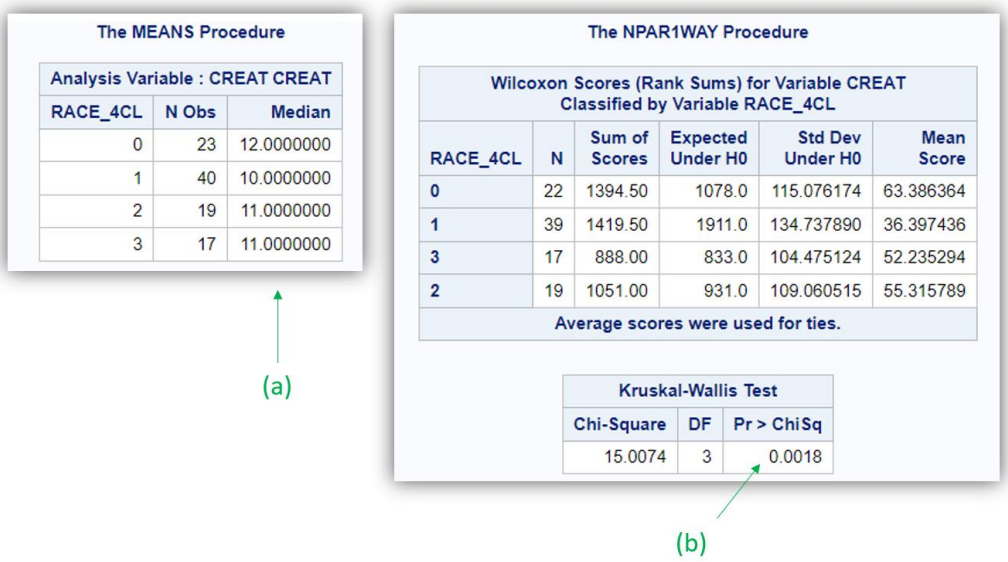


Figure 24

Ainsi, les valeurs des médianes de la concentration en créatinine parmi chacune des quatre races étaient de 12,00, 10,00, 11,00, et 11,00 mg/L respectivement pour les races « 0 », « 1 », « 2 », et « 3 » (cf. Figure 24.a). La valeur du degré de signification du test de Kruskal-Wallis est celle pointée par la flèche (b) sur la Figure 24 : p = 0,002 (en arrondissant). Ainsi, les quatre médianes de concentration en créatinine étaient significativement différentes.

C. Association entre deux variables quantitatives

1. Introduction

Pour savoir si deux variables quantitatives sont associées, on peut calculer un coefficient de corrélation. Ce sera celui de Pearson si la distribution des deux variables quantitatives peut être considérée comme normale, ou le coefficient de Spearman si au moins l'une des deux distributions ne peut pas être considérée comme normale. La procédure PROC CORR est celle qui peut être utilisée pour calculer ces coefficients de corrélation. Elle fournit aussi le test statistique permettant de tester si le coefficient de corrélation estimé est, ou non, significativement différent de 0 (la valeur 0 pour un coefficient de corrélation représentant une absence de corrélation entre les deux variables quantitatives).

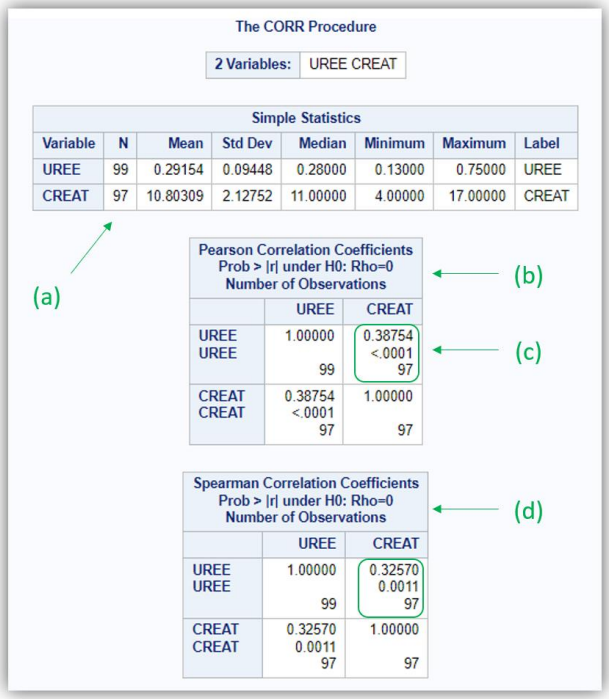
Avant de calculer un coefficient de corrélation, je ne peux que vous conseiller de représenter graphiquement l'association entre les deux variables quantitatives concernées à l'aide d'un nuage de points (cf. page 21).

2. Coefficient de corrélation

Je vais prendre comme exemple les variables UREE et CREAT. Les lignes de programme ci-dessous permettent d'obtenir les coefficients de corrélation de Pearson et de Spearman quantifiant la corrélation entre les variables UREE et CREAT (dont la représentation graphique de l'association est présentée sur la Figure 18).

```
PROC CORR DATA = Donnees_pour_guide PEARSON SPEARMAN;
VAR UREE CREAT;
RUN;
```

Les résultats des lignes de programme ci-dessus se trouvent sur la Figure 25.



The CORR Procedure

2 Variables: UREE CREAT

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum	Label
UREE	99	0.29154	0.09448	0.28000	0.13000	0.75000	UREE
CREAT	97	10.80309	2.12752	11.00000	4.00000	17.00000	CREAT

(a)

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

(b)

	UREE	CREAT
UREE UREE	1.00000 99	0.38754 <.0001 97
CREAT CREAT	0.38754 <.0001 97	1.00000 97

(c)

Spearman Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

(d)

	UREE	CREAT
UREE UREE	1.00000 99	0.32570 0.0011 97
CREAT CREAT	0.32570 0.0011 97	1.00000 97

Figure 25

La flèche (a) sur la Figure 25 indique que parmi les 99 chiens de l'étude, 2 n'ont pas de données pour la variable CREAT (97 données dans l'échantillon des 99 chiens). Ainsi, les coefficients de corrélation ne seront calculés que parmi les 97 chiens qui ont des données pour les deux variables CREAT et UREE. Dans le tableau pointé par la flèche (b) sur la Figure 25, on peut lire l'ordre des informations qui sont fournies dans le tableau : d'abord que le coefficient de corrélation de Pearson (r = 0,39), ensuite la valeur du degré de signification ($p < 0,01$), et enfin le nombre d'observations utilisées pour le calcul (97 chiens). Dans le tableau pointé par la flèche (d) sur la Figure 25, on peut lire que le coefficient de Spearman est égal à 0,33, avec un degré de signification $p < 0,01$.

VI. Réaliser des tests statistiques dans un sous échantillon

A. Introduction

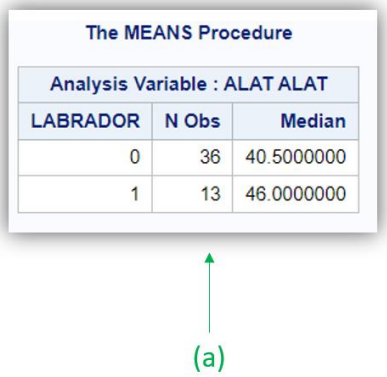
Il peut être parfois intéressant de réaliser des analyses statistiques sur un sous-échantillon de l'échantillon initial, en réalisant la sélection des individus de ce sous-échantillon sur la valeur d'une ou de plusieurs variables, qu'elles soient binaires, qualitatives, ou bien quantitatives. Pour cela, il faut utiliser, dans n'importe quelle procédure, l'instruction « WHERE ». L'instruction « AND » permettra de sélectionner des individus selon les valeurs de plusieurs variables. Vous allez voir ci-dessous les lignes de programme SAS® selon le type de la variable, puis vous verrez la situation d'une sélection d'individus selon les valeurs de plusieurs variables.

B. Sélection des individus sur une variable binaire

Supposons que l'on veuille estimer les médianes de la concentration en ALAT selon la race des chiens (Labradors *versus* autre race – utilisation de la variable LABRADOR), seulement parmi les chiens femelles de l'échantillon. Les lignes de programme ci-dessous permettent de réaliser l'analyse souhaitée.

```
PROC MEANS DATA = Donnees_pour_guide MEDIAN;
WHERE FEMELLE = 1;
CLASS LABRADOR;
VAR ALAT;
RUN;
```

Le résultat des lignes de programme ci-dessus se trouve sur la Figure 26.



The MEANS Procedure

Analysis Variable : ALAT ALAT		
LABRADOR	N Obs	Median
0	36	40.5000000
1	13	46.0000000

(a)

Figure 26

Comme vous pouvez le voir sur la Figure 26, rien n'indique que les médianes calculées ne l'ont été que parmi les chiens femelles de l'échantillon. Vous pouvez uniquement voir que la somme des deux nombres de chiens qui ont été utilisés (36 et 13 ; cf. Figure 26.a) n'est pas égale à 99 chiens (la taille totale de l'échantillon).

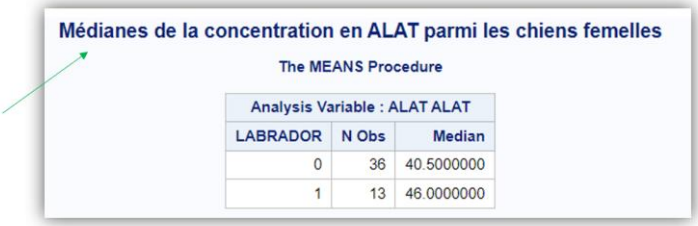
Pour faciliter l'interprétation d'un résultat SAS®, un titre peut être ajouté. Pour cela, il faut utiliser l'instruction « TITLE », que vous pouvez taper de façon isolée dans l'éditeur de programme (c'est-à-dire, pas nécessairement au sein d'une procédure). La ligne de programme ci-dessous permet d'ajouter le titre que vous souhaitez, à mettre entre guillemets.

```
TITLE "Le titre que vous souhaitez";
```

Les lignes de programme ci-dessous permettent d'obtenir le résultat de la Figure 26, mais avec le titre souhaité.

```
TITLE "Médianes de la concentration en ALAT parmi les chiens femelles";
PROC MEANS DATA = Donnees_pour_guide MEDIAN;
WHERE FEMELLE = 1;
CLASS LABRADOR;
VAR ALAT;
RUN;
```

Le résultat des lignes de programme ci-dessus se trouve sur la Figure 27.



Médianes de la concentration en ALAT parmi les chiens femelles

The MEANS Procedure

Analysis Variable : ALAT ALAT		
LABRADOR	N Obs	Median
0	36	40.5000000
1	13	46.0000000

Figure 27

Attention, dès que vous utilisez l'instruction « TITLE », le titre sera placé lors de l'exécution de *toutes* les procédures suivantes ! Pour supprimer l'affichage d'un titre, il suffit de taper l'instruction « TITLE ; » dans l'éditeur de programme, de façon isolée, de la sélectionner, puis de l'exécuter.

C. Sélection des individus sur une variable qualitative

Supposons que l'on veuille estimer les médianes de la concentration en ALAT selon l'obésité des chiens (variable OBESE), seulement parmi les chiens de race croisée Golden/Labrador (RACE_4CL = 2). Les lignes de programme ci-dessous permettent de réaliser l'analyse souhaitée, avec l'ajout d'un titre.

```
TITLE "Médianes de la concentration en ALAT parmi les chiens de race croisée
Labrador/Golden";
PROC MEANS DATA = Donnees_pour_guide MEDIAN;
WHERE RACE_4CL = 2;
CLASS OBESE;
VAR ALAT;
RUN;
```

Supposons maintenant que l'on veuille estimer les médianes de la concentration en ALAT selon l'obésité des chiens (variable OBESE), parmi les chiens de race Golden (RACE_4CL = 0) ou de race croisée Golden/Labrador (RACE_4CL = 2). Pour cela, il faut utiliser l'instruction « IN () », en mettant entre parenthèses toutes les valeurs souhaitées pour la sélection, séparées par une virgule. Les lignes de programme ci-dessous permettent de réaliser l'analyse souhaitée.

```
TITLE "Médianes de la concentration en ALAT parmi les chiens de race Golden et parmi
ceux de race croisée Labrador/Golden";
PROC MEANS DATA = Donnees_pour_guide MEDIAN;
WHERE RACE_4CL IN (0,2);
CLASS OBESE;
VAR ALAT;
RUN;
```

D. Sélection des individus sur une variable quantitative

Dès que l'on souhaite sélectionner des individus selon les valeurs d'une variable quantitative, il faut faire très attention aux données manquantes de cette variable quantitative. Comme je l'ai écrit ci-dessus (page 14), dans SAS®, la donnée manquante est matérialisée par un « . » (cf. Figure 8.b et Figure 8.c) et elle vaut -∞.

Ainsi, en supposant que l'on veuille estimer la médiane de la concentration en urée des chiens dont la concentration en ALAT est inférieure à 50 UI/L, l'exécution des lignes de programme ci-dessous fournira la médiane de la concentration en urée parmi tous les chiens dont la concentration en ALAT est inférieure à 50 UI/L *ainsi que* parmi les chiens dont la donnée manque sur la concentration en ALAT[4] (ce qui n'est pas du tout souhaité !).

```
PROC MEANS DATA = Donnees_pour_guide MEDIAN N;
WHERE ALAT < 50;
VAR UREE;
RUN;
```

Le résultat des lignes de programme ci-dessus est présenté sur la Figure 28.a.

De façon générale, pour exclure de la sélection les individus qui ont une donnée manquante sur la variable quantitative sur laquelle porte la sélection, et lorsque l'on souhaite sélectionner les individus selon une valeur inférieure à un seuil, il faut ajouter dans l'instruction « > . » (puisque le « . » pour une variable quantitative vaut -∞). Les lignes de programme ci-dessous permettent d'estimer la médiane de la concentration en urée des chiens dont la concentration en ALAT est non manquante et inférieure à 50 UI/L.

[4] Il se trouve que dans le fichier de données utilisé pour ce guide, il n'y a qu'un seul chien pour lequel la donnée manque sur la concentration en créatinine.

```
PROC MEANS DATA = Donnees_pour_guide MEDIAN N;
WHERE (. < ALAT < 50);
VAR UREE;
RUN;
```

Le résultat des lignes de programme ci-dessus est présenté sur la Figure 28.b.

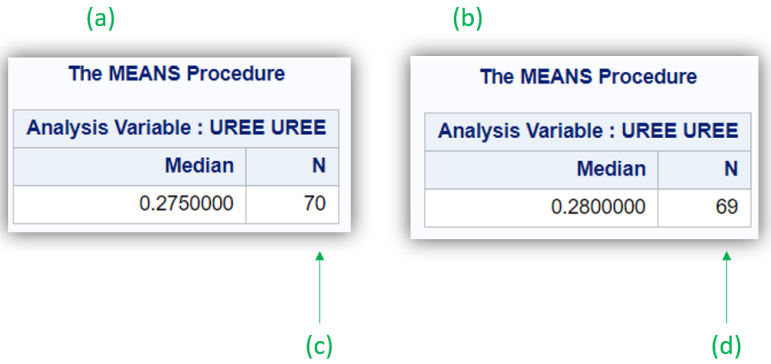


Figure 28

On peut remarquer que la différence du nombre de valeurs de concentration en urée utilisées pour calculer la médiane de la concentration en urée est bien de 1 (70 valeurs avec l'instruction « WHERE ALAT < 50; » qui inclut la donnée manquante sur la concentration en ALAT contre 69 avec l'instruction « WHERE (. < ALAT < 50); », cf. Figure 28.c et Figure 28.d, respectivement).

Bien entendu, si l'on souhaite sélectionner les individus sur une variable quantitative *à partir* d'une valeur de cette variable, et non *jusqu'à* une valeur de cette variable, la présence de données manquantes n'est plus un problème. Les lignes de programme ci-dessous permettent d'estimer la médiane de la concentration en urée parmi les chiens dont la concentration en ALAT est supérieure ou égale à 50 UI/L, et le chien qui a une donnée manquante sur la concentration en ALAT ne sera pas inclus dans la sélection.

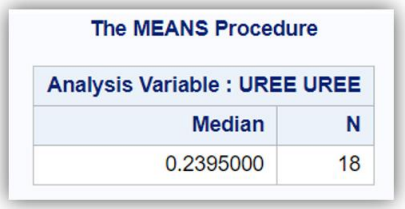
```
PROC MEANS DATA = Donnees_pour_guide MEDIAN;
WHERE ALAT >= 50;
VAR UREE;
RUN;
```

E. Sélection des individus sur plusieurs variables

Supposons que l'on veuille estimer la médiane de la concentration en urée parmi les chiens dont la concentration en ALAT est inférieure ou égale à 50 UI/L, mâles, et de race Golden (RACE_4CL = 0) ou de race croisée Golden/Labrador (RACE_4CL = 2). Pour cela, il faut utiliser l'instruction « AND » dans l'instruction « WHERE ». Les lignes de programme ci-dessous permettent de réaliser l'analyse souhaitée (les parenthèses dans l'instruction « WHERE » ci-dessous ne sont pas indispensables, je les ai mises pour faciliter la lecture des lignes de programme).

```
PROC MEANS DATA = Donnees_pour_guide MEDIAN N;
WHERE (. < ALAT <= 50) AND (FEMELLE = 0) AND (RACE_4CL IN (0,2));
VAR UREE;
RUN;
```

Le résultat des lignes de programme ci-dessus est présenté sur la Figure 29.



The MEANS Procedure

Analysis Variable : UREE UREE	
Median	N
0.2395000	18

Figure 29

Ainsi, parmi les 18 chiens dont la concentration en ALAT est inférieure ou égale à 50 UI/L (mais non manquante), mâles, et de race Golden ou de race croisée Golden/Labrador, la médiane de la concentration en urée était égale à 0,24 g/L.

VII. Analyse de survie à l'aide des courbes de Kaplan-Meier

A. Introduction

Vous devez avoir acquis les connaissances de base en analyse de survie (cf. page 8), et avoir lu et compris tout ce qui précède dans ce guide avant de poursuivre.

Pour dresser une ou plusieurs courbes de Kaplan-Meier, je vous suggère d'utiliser la procédure PROC LIFETEST. Dans tous les exemples qui vont suivre dans cette partie « Analyse de survie à l'aide des courbes de Kaplan-Meier », je vais utiliser comme événement d'intérêt la survenue d'un décès au cours du temps (variable DECES, qui vaut « 1 » pour les chiens décédés au cours de l'étude, et « 0 » pour les chiens censurés). Le temps de survie a été créé dans le fichier de données Excel®, et la variable correspondante est SURVIE.

B. Réalisation d'une seule courbe de Kaplan-Meier dans l'ensemble de l'échantillon

Supposons que l'on veuille dresser la courbe de Kaplan-Meier représentant l'incidence *globale* d'un décès parmi les 99 chiens de l'échantillon. Les lignes de programme ci-dessous permettent de dresser une telle courbe.

```
PROC LIFETEST DATA = Donnees_pour_guide;
TIME SURVIE * DECES(0);
RUN;
```

La variable qui suit « TIME » et qui précède le signe « * » correspond à la variable relative au temps de survie. La variable qui suit le signe « * » correspond à la variable relative à l'événement. Le chiffre entre parenthèses doit être celui correspondant aux individus censurés. Dans l'exemple, ce chiffre est « 0 » car il s'agit de la valeur qui a été attribuée aux chiens censurés dans l'étude (DECES = 0).

Les résultats des lignes de programme ci-dessus sont présentés sur la Figure 30 et sur la Figure 31.

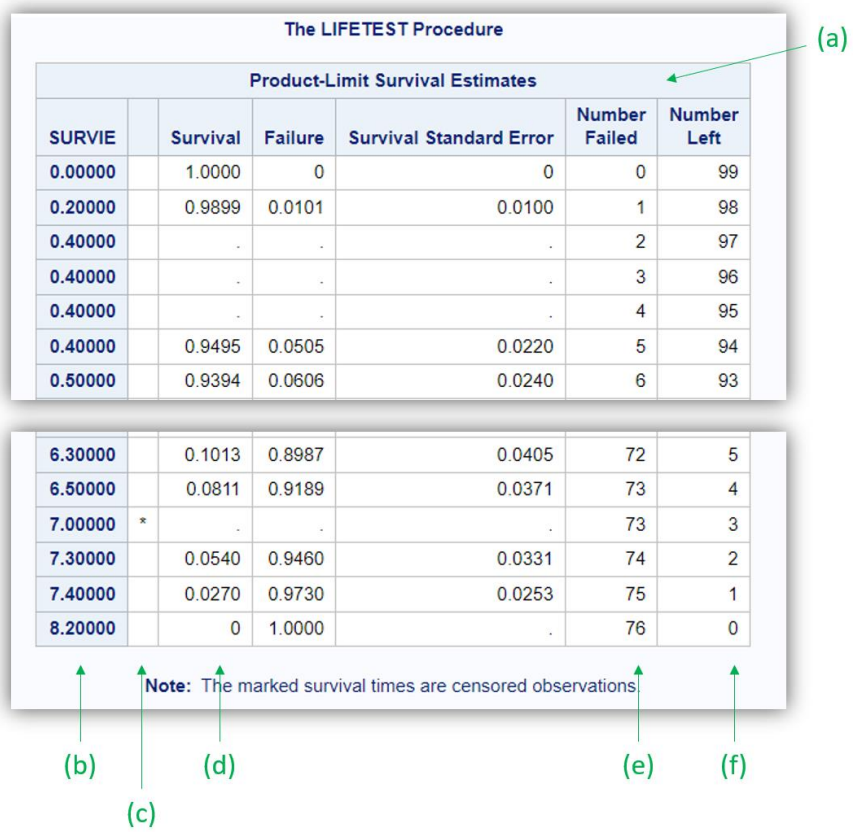


Figure 30

Le tableau « Product-Limit Survival Estimates » (cf. Figure 30.a, dont il s'agit d'un extrait) comprend de nombreuses informations que je vais détailler ici. La colonne « SURVIE » qui prend le nom de la variable correspondant au temps de survie dans les analyses (cf. Figure 30.b) liste tous les temps de survie du fichier de données, triés du plus petit (en haut) au plus grand (en bas). La colonne pointée par la flèche (c) sur la Figure 30 comprend une « * » si le temps de survie correspond à une censure. La colonne « Survival » (cf. Figure 30.d) comprend la valeur de S(t) estimée par la méthode de Kaplan-Meier. La colonne « Number Failed » (cf. Figure 30.e) comprend le nombre cumulé, à chaque temps de survie, d'événements survenus jusqu'au temps de survie considéré. La colonne « Number Left » (cf. Figure 30.f) comprend le nombre cumulé, à chaque temps de survie, d'individus encore à risque d'événements *juste après* le temps de survie considéré.

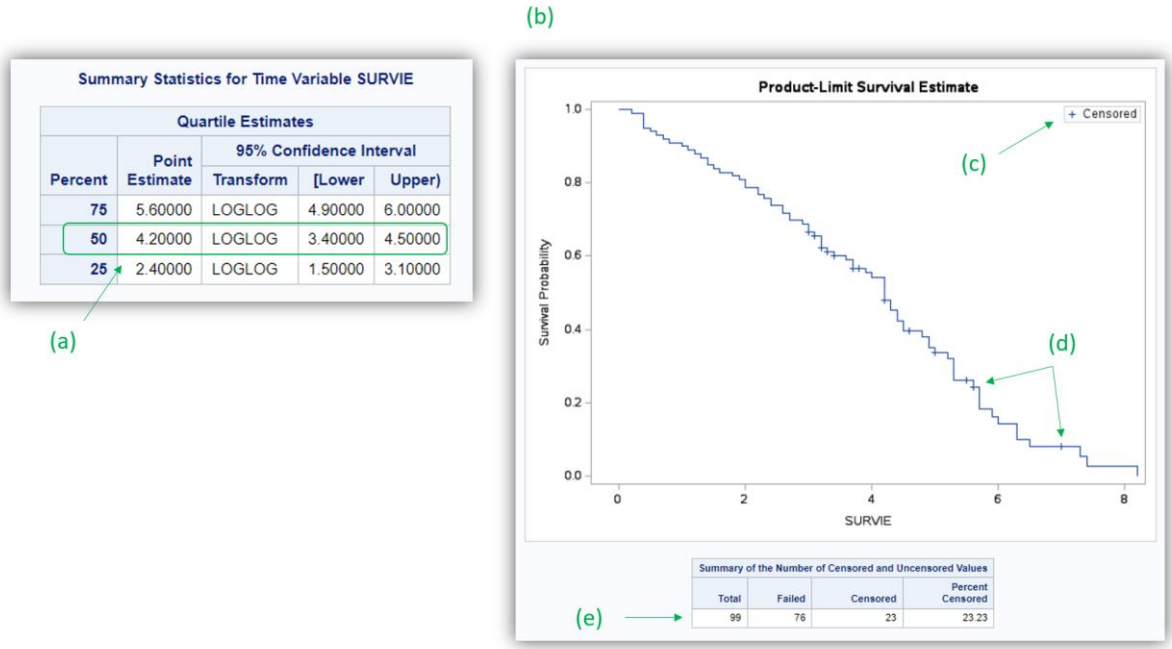


Figure 31

La flèche (a) sur la Figure 31 indique le temps de survie médian (ici, 4,2 ans) et son intervalle de confiance à 95% ($IC_{95\%}$), ici 3,4 – 4,5 ans. La Figure 31.b représente la courbe de Kaplan-Meier, sur laquelle sont représentées les censures à l'aide d'un signe « + » (cf. Figure 31.c), comme par exemple les deux signes « + » pointés par la flèche (d) sur la Figure 31. Le tableau sous la courbe de Kaplan-Meier pointé par la flèche (e) sur la Figure 31 rappelle le nombre de chiens utilisés pour l'analyse de survie (ici, 99 chiens), le nombre d'événements survenus au cours du temps (ici, 76 décès), et le nombre de censures (ici 23 chiens censurés).

C. Réalisation de plusieurs courbes de Kaplan-Meier

1. Introduction

Lorsque l'on souhaite savoir si une variable est associée à la survenue d'un événement au cours du temps, il est tout à fait possible d'étudier cette association en comparant (puis en testant) deux ou plusieurs courbes de survie de Kaplan-Meier. Pour étudier une telle association en comparant des courbes de Kaplan-Meier, la variable doit être binaire ou qualitative. Si elle est quantitative, il faudra la rendre binaire ou qualitative. Nous verrons plus loin que dans ce dernier cas de figure, il n'est pas nécessaire de créer une variable binaire ou qualitative à partir de la variable quantitative au préalable dans le fichier de données (SAS® le fait tout seul).

Pour dresser plusieurs courbes de Kaplan-Meier puis pour les tester avec le test du Log-rank, il faut utiliser l'instruction « STRATA » au sein de la procédure PROC LIFETEST.

2. Situation d'une variable binaire

Supposons que l'on veuille savoir si la présence d'une démarche anormale observée à J0 (variable binaire DEMARCHE_ANORMALE) est associée à la survenue d'un décès chez les chiens de l'étude. Les lignes de programme ci-dessous permettent de réaliser l'analyse souhaitée.

```
PROC LIFETEST DATA = Donnees_pour_guide;
TIME SURVIE * DECES(0);
STRATA DEMARCHE_ANORMALE;
RUN;
```

Les résultats (extrait) des lignes de programme ci-dessous sont présentés sur la Figure 32 et sur la Figure 33.

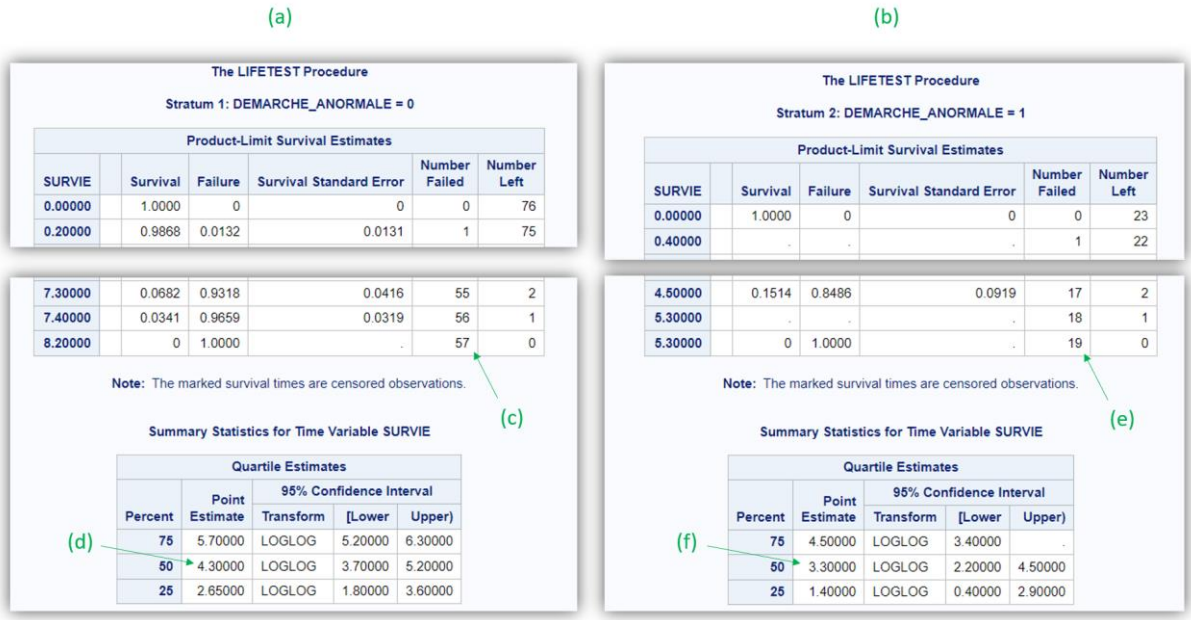


Figure 32

La Figure 32.a fournit les temps de survie et le temps de survie médian (entre autres) des chiens qui ne présentaient pas de démarche anormale (DEMARCHE_ANORMALE = 0). Notamment, le temps de survie médian de ces 57 chiens (cf. Figure 32.c) est égal à 4,3 ans (cf. Figure 32.d). La Figure 32.b fournit les temps de survie et le temps de survie médian (entre autres) des chiens qui présentaient une démarche anormale (DEMARCHE_ANORMALE = 1). Notamment, le temps de survie médian de ces 19 chiens (cf. Figure 32.e) est égal à 3,3 ans (cf. Figure 32.f).

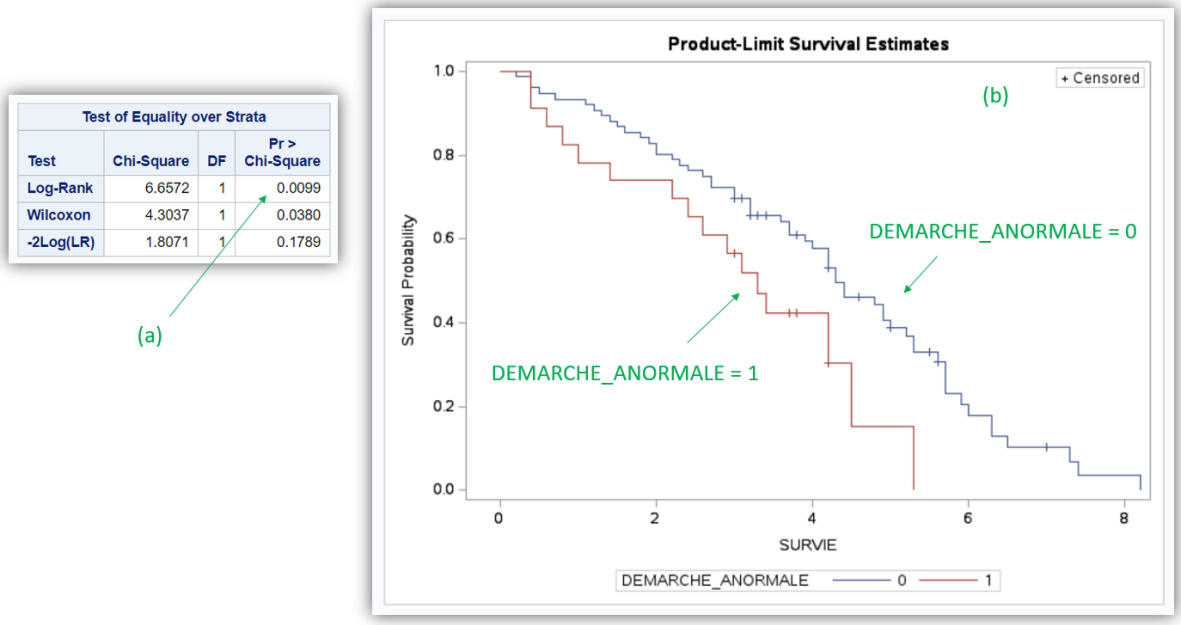


Figure 33

Le degré de signification du test du Log-rank de valeur inférieure à 0,05 (cf. Figure 33.a) permet de dire que les deux temps de survie médians cités ci-dessus (4,3 et 3,3 ans) étaient significativement différents. Sur la Figure 33.b, on observe que les chiens avec une démarche anormale (courbe rouge, cf. légende de la figure) étaient décédés plus rapidement que les chiens sans démarche normale (courbe bleue, qui se trouve toujours au-dessus de la courbe rouge). Dans la mesure où le test du Log-rank est significatif, on peut aussi dire que les deux courbes de Kaplan-Meier étaient significativement différentes.

3. Situation d'une variable qualitative

L'utilisation d'une variable qualitative dans l'instruction « STRATA » fournit exactement les mêmes informations que ce que je viens de présenter ci-dessus pour une variable binaire.

Supposons que l'on veuille savoir si l'âge du chien (en utilisant la variable AGE_4CL, variable qualitative ordinale en quatre classes) est associé à la survenue d'un décès chez les chiens de l'étude. Les lignes de programme ci-dessous permettent de réaliser l'analyse souhaitée.

```
PROC LIFETEST DATA = Donnees_pour_guide;
TIME SURVIE * DECES(0);
STRATA AGE_4CL;
RUN;
```

Les courbes de Kaplan-Meier produites par les lignes de programme ci-dessus sont présentées sur la Figure 34.a.

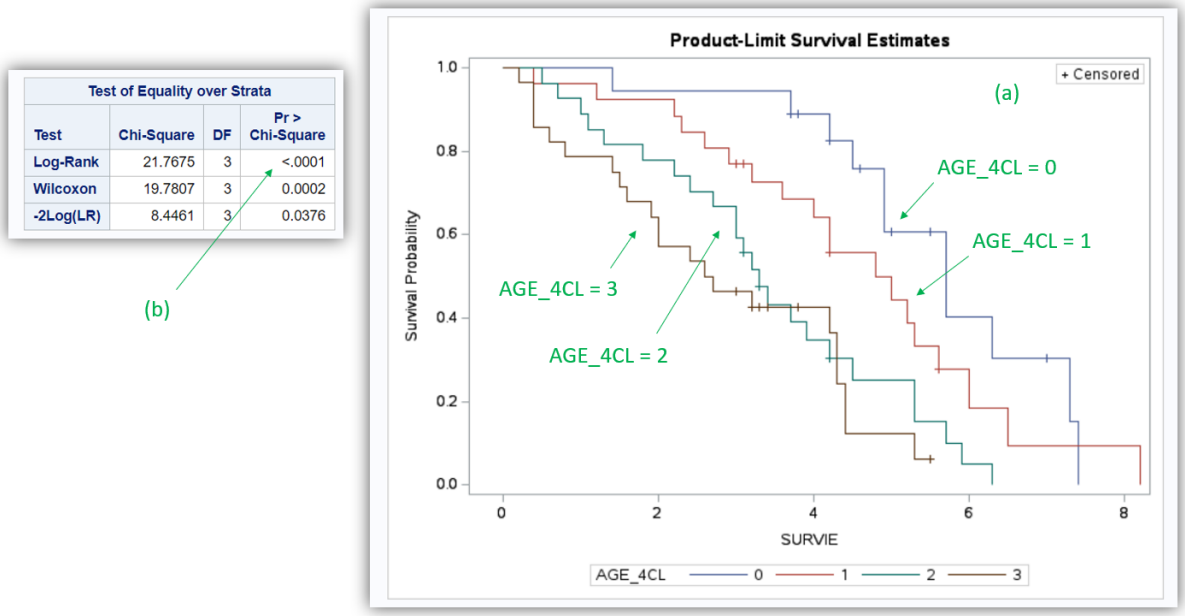


Figure 34

On y observe que les chiens les plus jeunes (AGE_4CL = 0) décèdent moins rapidement que les chiens un peu plus vieux (AGE_4CL = 1), eux-mêmes décédant moins rapidement que les chiens encore un peu plus vieux (AGE_4CL = 2), eux-mêmes décédant moins rapidement que les chiens les plus vieux (AGE_4CL = 3). Lorsque la variable qualitative est ordinale (comme c'est le cas ici), attention à ne pas sur-interpréter le fait que le test du Log-rank soit significatif (cf. Figure 34.b). Bien que l'on *observe* une relation « dose-effet » avec l'âge (plus la classe d'âge augmente, et plus la survenue de décès était rapide), le test du log-rank ne teste pas de relation « dose-effet ». Ainsi, il n'est pas question de dire que la survenue de décès était *significativement* plus rapide lorsque l'âge du chien augmentait. Par exemple, si les courbes bleues (AGE_4CL = 0) et marron (AGE_4CL = 3) avaient été interverties, on

n'aurait plus du tout observé de relation « dose-effet » avec l'âge, alors que la valeur du degré de signification issu du test du Log-rank eût été *a priori* identique.

4. Situation d'une variable quantitative

Supposons que l'on veuille savoir si la concentration en créatinine est associée à la survenue d'un décès dans l'échantillon. Pour rendre binaire ou qualitative une variable quantitative, pour ensuite dresser les courbes de survies correspondant à chacune des classes de cette variable binaire ou qualitative, il faut préciser, entre parenthèses dans l'instruction « STRATA », le ou les seuils de la variable quantitative utilisés pour la rendre binaire ou qualitative.

Les lignes de programme ci-dessous permettent de dresser les courbes de Kaplan-Meier selon que la concentration en créatinine est supérieure ou inférieure à 11 mg/L (ainsi, deux courbes de Kaplan-Meier seront dressées).

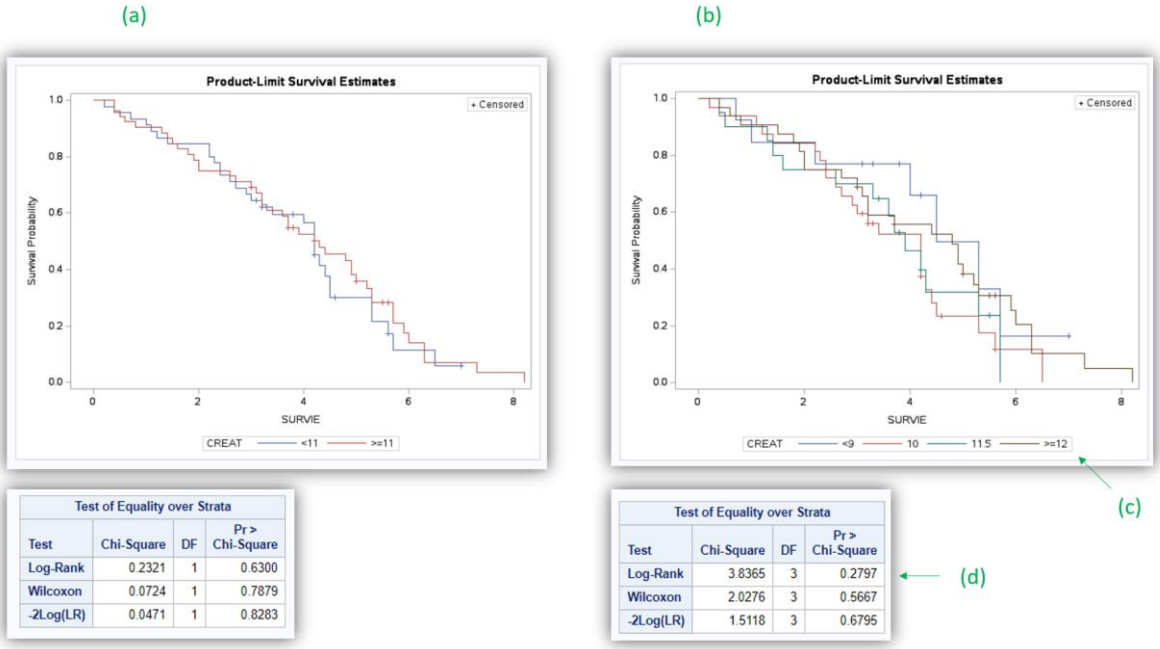
```
PROC LIFETEST DATA = Donnees_pour_guide;
TIME SURVIE * DECES(0);
STRATA CREAT (11);
RUN;
```

Les courbes de Kaplan-Meier produites par les lignes de programme ci-dessus sont présentées sur la Figure 35.a.

Les lignes de programme ci-dessous permettent de dresser les courbes de Kaplan-Meier selon que la concentration en créatinine est inférieure à 9, comprise entre 9 et 11 (exclus), comprise entre 11 et 12 (exclus), et supérieure ou égale à 12 mg/L (ainsi, quatre courbes de Kaplan-Meier seront dressées).

```
PROC LIFETEST DATA = Donnees_pour_guide;
TIME SURVIE * DECES(0);
STRATA CREAT (9 11 12);
RUN;
```

Les courbes de Kaplan-Meier produites par les lignes de programme ci-dessus sont présentées sur la Figure 35.b.



Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.2321	1	0.6300
Wilcoxon	0.0724	1	0.7879
-2Log(LR)	0.0471	1	0.8283

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	3.8365	3	0.2797
Wilcoxon	2.0276	3	0.5667
-2Log(LR)	1.5118	3	0.6795

Figure 35

La légende de la Figure 35.b (cf. flèche (c) sur la Figure 35.b) correspond au centre « mathématique » des classes : la classe « 10 » correspond à la classe « entre 9 et 11 mg/L », et le chiffre de 10 correspond

(cc) BY-NC-ND Licence Creative Commons – Guide d'utilisation de SAS® Studio – Loïc Desquilbet – Version v1.6

au milieu de l'intervalle 9-11. De même la classe « 11.5 » correspond à la classe « entre 11 et 12 mg/L », et le chiffre de 11,5 correspond au milieu de l'intervalle 11-12.

Bien entendu, comme déjà écrit précédemment, le test du Log-rank pointé par la flèche (d) sur la Figure 35.b ne teste de pas de relation « dose-effet » entre la concentration en créatinine et la survenue d'un décès.

VIII. Modèles de régression

A. Théorie des modèles de régression

Vous devez avoir acquis les connaissances de base en épidémiologie (analytique) (cf. page 8), et avoir lu et compris tout ce qui précède dans ce guide avant de poursuivre (sauf la partie précédente sur l'analyse de survie si vous ne comptez pas utiliser un modèle de Cox).

1. Vérification d'hypothèses sur lesquelles repose un modèle de régression

Un modèle de régression, utilisé sur des données médicales, repose très souvent sur une ou plusieurs hypothèses. Ces hypothèses *doivent* être acceptables d'un point de vue biologique ou physiopathologique. Ce n'est pas la statistique qui dit ce qu'il se passe d'un point de vue biologique ou physiopathologique. C'est la clinique ou la médecine qui dicte d'abord sa loi, et le modèle de régression, lui, s'exécute. Il s'exécutera mal si les hypothèses sur lesquelles repose le modèle ne sont pas correctes biologiquement ou physiopathologiquement parlant. De plus, je parle d'hypothèses biologiques ou physiopathologiques au sein de la population cible, dans « la vraie vie ». Certes, nous n'avons toujours que les données d'un échantillon pour vérifier des hypothèses, mais vous devez *toujours* avoir conscience que la fluctuation d'échantillonnage pourra conduire à observer des données dans l'échantillon éloignées de celles de la population. Par conséquent, ce n'est pas parce qu'une hypothèse est vérifiée dans l'échantillon qu'elle l'est dans la population cible, et ce n'est pas parce qu'elle ne l'est pas dans l'échantillon qu'elle ne l'est pas non plus dans la population cible. Notamment, si la vérification d'une hypothèse se base sur la valeur d'un degré de signification (qui proviendra forcément d'un test statistique dont l'hypothèse nulle H_0 est « l'hypothèse que l'on souhaite vérifier est vraie dans la population cible »), ayez en tête qu'un degré de signification supérieur à 0,05 (acceptation de H_0) ne veut pas dire que H_0 est vraie (donc cela ne veut pas dire que l'hypothèse à vérifier est vraie dans la population cible). Cela indique simplement que l'on n'a pas de preuves fortes qu'elle est fausse. De plus, lorsqu'un degré de signification est inférieur à 0,05, on a d'autant *moins* de preuves que H_0 est fausse que (1) la taille de l'échantillon est faible et que (2) il y a *a priori* peu de chances que H_0 soit fausse (Desquilbet, 2020). Ainsi, lorsque l'on a de bonnes raisons (médicales, physiologiques, etc.) de penser qu'une hypothèse est vérifiée dans la population cible, une valeur de degré de signification, issu d'un test statistique testant l'hypothèse à vérifier, inférieure à 0,05 n'apporte pas de preuves fortes qu'elle n'est pas vérifiée dans la population cible (ces preuves seront « légères »). Par conséquent, dans certaines situations, il pourra ne pas être nécessaire de vérifier une hypothèse à partir des données de l'échantillon si l'on a de fortes raisons de penser que cette hypothèse est vérifiée dans la population cible.

Bref, tout ça pour dire que les données de l'échantillon donnent *uniquement* une *indication* sur le fait qu'une hypothèse est, ou n'est pas, vérifiée dans la population cible, mais la réflexion et les connaissances médicales restent le « Gold Standard ».

2. Ecriture d'un modèle de régression

Un modèle de régression met en relation le CdJ, quantifié par Y, et une ou plusieurs variables (⇔ expositions) E_i. Tout d'abord, il est fortement recommandé que chacune des variables incluses dans un modèle soit une variable numérique (cf. page 9). Ainsi, un modèle de régression comprenant N variables E_i ($i \in \{1, \ldots, N\}$) s'écrit de façon générale :

$$\bar{Y}_{/E_1,E_2,\ldots,E_N} = \alpha + \sum_{i=1}^{N} \beta_i . E_i$$

En français, ce « $\bar{Y}_{/E_1,E_2,\ldots,E_N}$ » se lit « la valeur de l'espérance de Y sachant les valeurs des variables E_1, E_2, …, E_N ». Le mot « espérance », qui est un terme mathématique[5], peut être compris comme « la valeur estimée de Y à partir des données de l'échantillon pour des valeurs fixées de chacune des variables incluses dans le modèle ».

Par exemple, soit E_1 la variable OBESE et E_2 la variable AGE. Supposons que l'on fasse tourner le modèle suivant à partir des données de l'échantillon : $Y = \alpha + \beta_1.OBESE + \beta_2.AGE$. Supposons que SAS® estime, à partir des données de l'échantillon, que $\alpha = 6$, $\beta_1 = 2$, et $\beta_2 = -0{,}5$. Le modèle s'écrit alors : $Y = 6 + 2.OBESE - 0{,}5.AGE$. Par conséquent, le logiciel SAS® estime que la valeur de Y pour un chien obèse (OBESE = 1) âgé de 8 ans (AGE = 8) est égale à : $6 + 2x1 - 0{,}5x8 = 4$.

Un modèle de régression permet par conséquent d'estimer la valeur de Y en fonction des valeurs des différentes variables incluses dans le modèle. Néanmoins, un modèle de régression est rarement utilisé à cette fin en recherche clinique (l'étude de Darnis et al. utilise, pour le coup, des modèles de régression à cette fin-là (Darnis et al., 2018)). Il est bien davantage utilisé pour quantifier puis tester l'association entre une des expositions E_i incluses dans le modèle et Y. Et c'est ce sur quoi je vais me focaliser dans la suite de cette partie sur les modèles de régression.

Si N = 1, on dira que le modèle de régression est « univarié » (il ne contient qu'une seule variable), et si $N \geq 2$, alors on dira que le modèle de régression est multivarié. Dans le modèle, α et les β_i ($i \in \{1, \ldots, N\}$) sont appelés « coefficients » du modèle.

3. Choix d'un modèle de régression et écriture mathématique du modèle

Ce qui guide le choix d'un modèle de régression est le type de la variable relative au CdJ.

Si le CdJ est quantitatif (par exemple, la concentration en ALAT), le modèle de régression est la régression linéaire et Y est directement la variable relative au CdJ. Notez que pour utiliser un modèle de régression linéaire, le CdJ quantitatif Y doit suivre à peu près une loi normale. Si tel n'est pas le cas, je vous recommande alors de transformer la variable quantitative correspondant au CdJ en une variable binaire, en utilisant un seuil qui a un sens clinique (et vous utiliserez alors un modèle de régression logistique – cf. ci-dessous). Le modèle de régression linéaire s'écrit :

$\overline{CdJ}_{/E_1,E_2,\ldots,E_N} = \alpha + \sum_{i=1}^{N} \beta_i . E_i$, où « $\overline{CdJ}_{/E_1,E_2,\ldots,E_N}$ » est l'espérance de la valeur du CdJ quantitatif en fonction des valeurs des variables E_i incluses dans le modèle. De façon générale, β_i quantifie l'association entre le CdJ (quantitatif) et E_i en tant différence moyenne de valeurs du CdJ quantitatif.

[5] https://fr.wikipedia.org/wiki/Esp%C3%A9rance_math%C3%A9matique

Si le CdJ est binaire, et non assorti d'un temps de survenue (par exemple, dans une étude cas-témoins ou transversale), alors le modèle de régression est la régression logistique. Le modèle de régression logistique s'écrit :

$Logit\left(\bar{P}_{/E_1,E_2,\ldots,E_N}\right) = Ln\left(\frac{\bar{P}_{/E_1,E_2,\ldots,E_N}}{1-\bar{P}_{/E_1,E_2,\ldots,E_N}}\right) = \alpha + \sum_{i=1}^{N} \beta_i . E_i$, où « $\bar{P}_{/E_1,E_2,\ldots,E_N}$ » est l'espérance de la probabilité de présenter le CdJ en fonction des valeurs des variables E_i incluses dans le modèle. De façon générale, β_i quantifie l'association entre la *présence* du CdJ (binaire) et E_i en tant que valeur du $Ln\left(OR_{E_i}\right)$, où OR_{E_i} est l'Odds Ratio quantifiant l'association entre la présence du CdJ et la variable E_i.

Si le CdJ est binaire et assorti d'un temps de survenue (par exemple, dans une étude de cohorte), alors le modèle de régression est le modèle de Cox (Cox, 1972). Le modèle de Cox s'écrit :

$Ln\left(\overline{\lambda(t)}_{/E_1,E_2,\ldots,E_N}\right) = Ln\left(\lambda_0(t)\right) + \sum_{i=1}^{N} \beta_i . E_i$, où « $\overline{\lambda(t)}_{/E_1,E_2,\ldots,E_N}$ » est l'espérance de l'incidence instantanée du CdJ en fonction de la valeur des variables E_i incluses dans le modèle. De façon générale, β_i quantifie l'association entre la *survenue* du CdJ (binaire) et E_i en tant que valeur du $Ln\left(HR_{E_i}\right)$, où HR_{E_i} est le Risque Relatif (« Hazard Ratio » pour un modèle de Cox, qui est un « rapport des incidences instantanées ») quantifiant l'association entre la survenue du CdJ et la variable E_i.

4. Test statistique des coefficients d'un modèle de régression

Chaque coefficient d'un modèle de régression peut être testé par un test statistique. Le test statistique de Wald est un des tests statistiques proposés par les logiciels de statistique pour tester un coefficient d'un modèle. Vous devez savoir que l'hypothèse nulle H_0 d'un test statistique testant l'association entre une variable E_i et un CdJ est « il n'y a pas d'association entre E_i et le CdJ dans la population cible ». Or, s'il n'y a pas d'association entre E_i et le CdJ, cela signifie que $\beta_i = 0$ dans la population cible. Ainsi, l'hypothèse nulle H_0 du test statistique de Wald testant un coefficient β_i est : « $\beta_i = 0$ dans la population cible ». Si la valeur du degré de signification du test de Wald testant un coefficient β_i est inférieure à 0,05, cela veut dire que le coefficient β_i est significativement différent de 0, ce qui veut dire que, dans l'échantillon, il existait une association significative entre la variable E_i et le CdJ.

5. Problématique des données manquantes

Ce point est très important et il est souvent omis par les utilisateurs de modèles de régression. Les individus de l'échantillon qui sont utilisés pour estimer les coefficients du modèle $\bar{Y}_{/E_1,E_2,\ldots,E_N} = \alpha + \sum_{i=1}^{N} \beta_i . E_i$ sont les individus de l'échantillon tels qu'aucune de leurs *N* variables E_i n'a de donnée manquante – et bien entendu, ces individus ne doivent pas non plus avoir de donnée manquante sur Y.

La Figure 36 présente un exemple fictif d'un fichier de données de six individus, pour la variable Y correspondant au CdJ et pour les trois variables E_1, E_2, et E_3. Dans ce fichier de données, l'individu #1 a une donnée manquante pour la variable E_2, l'individu #2 a une donnée manquante pour la variable E_3, l'individu #3 a une donnée manquante pour la variable Y, l'individu #4 a une donnée manquante pour la variable E_1, et les individus #5 et #6 n'ont aucune donnée manquante pour les quatre variables.

(cc) BY-NC-ND Licence Creative Commons – Guide d'utilisation de SAS® Studio – Loïc Desquilbet – Version v1.6

ID	Y	E1	E2	E3
1	5	56		55
2	6	34	11	
3		89	24	87
4	2		21	90
5	4	77	9	65
6	3	54	27	50

Figure 36

Le modèle de régression (A) $\bar{Y}_{/E_1,E_2,E_3} = \alpha + \beta_1.E_1 + \beta_2.E_2 + \beta_3.E_3$ ne tournera que sur les individus #5 et #6, car ce sont uniquement ces deux individus pour lesquels aucune donnée ne manque sur les variables E_1, E_2, E_3, et Y. Le modèle de régression (B) $\bar{Y}_{/E_1,E_2} = \alpha + \gamma_1.E_1 + \gamma_2.E_2$ tournera quant à lui sur les individus #2, #5, et #6.

Il y a deux conséquences de cela très importantes. La première, c'est qu'un modèle tourne parfois sur beaucoup moins d'individus qu'attendus, même si, individuellement, chaque individu a très peu de données manquantes. Parfois, il faudra admettre de ne pas inclure une variable dans un modèle si elle est manquante pour beaucoup d'individus (ce qui est très embêtant si cette variable est un facteur de confusion). La seconde est la suivante. En reprenant l'exemple des modèles (A) et (B) ci-dessus, si vous souhaitez comparer les valeurs de β_2 et γ_2 (notamment pour savoir si la variable E_3 a joué un rôle de confusion dans l'étude de l'association entre E_2 et le CdJ), ces valeurs de β_2 et γ_2 ne pourront être comparées que si les deux modèles (A) et (B) tournent sur les *mêmes* individus – ce qui n'est ici pas le cas sans intervention de votre part (puisque l'individu #2 a été utilisé pour estimer les valeurs de γ_1 et γ_2 mais cet individu #2 n'a pas été utilisé pour estimer les valeurs de β_1, de β_2, et de β_3). Pour imposer le fait que ces deux modèles (A) et (B) tournent sur les mêmes individus (les individus #5 et #6 uniquement), il faudra utiliser l'instruction « WHERE » que nous avons vue précédemment et qui permet de réaliser les analyses sur un sous-échantillon (cf. page 30) : en l'occurrence ici le sous-échantillon sera constitué d'individus dont aucune donnée ne manque sur les variables d'intérêt. La ligne de programme ci-dessous permet de réaliser les analyses statistiques sur les individus dont aucune donnée ne manque sur les variables E_1, E_2, et E_3.

```
WHERE E1 NE . AND E2 NE . AND E3 NE .;
```

Dans la ligne de programme ci-dessus, « NE » (précédé *et* suivi d'un espace) signifie « not equivalent » en anglais, soit « non égal à » en français, et vous devez vous souvenir que le signe « . » signifie que la donnée est manquante pour une variable numérique. Il est inutile de préciser que la variable Y relative au CdJ ne doit pas manquer puisque de toute façon, les modèles (A) et (B) n'utilisent pas l'individu #3 pour lequel la variable Y est manquante.

B. La régression linéaire

1. Introduction

Même si vous ne prévoyez de n'utiliser que la régression logistique ou que le modèle de Cox, il est néanmoins indispensable de lire toute cette partie sur la régression linéaire. En effet, bien que l'interprétation des résultats d'une régression logistique ou d'un modèle de Cox ne soit pas la même que celle d'une régression linéaire, la démarche d'interprétation est quant à elle identique aux trois modèles. (Et je ne vais pas répéter les choses dans les parties dédiées à la régression logistique et au modèle de Cox que j'ai déjà écrites dans cette partie sur la régression linéaire.)

Pour illustrer l'interprétation des résultats d'une régression linéaire, je vais prendre comme variable relative au CdJ la concentration en créatinine, dont on suppose qu'elle suit une loi normale (variable quantitative CREAT).

Pour réaliser une régression linéaire sous SAS®, je vous suggère d'utiliser la procédure « PROC GLM ».

2. Interprétation des résultats d'une régression linéaire univariée

a) Cas général

Le modèle de régression linéaire univarié s'écrit : $\bar{Y}_{/E} = \alpha + \beta.E$, avec E une variable quelconque (binaire, qualitative, ou quantitative). Pour interpréter la valeur de β, je vais écrire ce modèle pour deux groupes d'animaux : un groupe au sein duquel les animaux ont une valeur égale à e_1 pour E, et un second groupe au sein duquel les animaux ont une valeur égale à e_2 pour E.

$\bar{Y}_{/E=e_1} = \alpha + \beta.e_1$

$\bar{Y}_{/E=e_2} = \alpha + \beta.e_2$

Ainsi, le modèle estime que la valeur de Y chez les animaux dont E vaut e_1 est $\alpha + \beta.e_1$, et que la valeur de Y chez les animaux dont E vaut e_2 est $\alpha + \beta.e_2$. Maintenant, je fais la soustraction entre les deux estimations :

$\bar{Y}_{/E=e_2} - \bar{Y}_{/E=e_1} = (\alpha + \beta.e_2) - (\alpha + \beta.e_1) = \beta.(e_2 - e_1)$

Ainsi, lorsque l'écart sur la valeur de la variable E entre deux groupes d'animaux vaut +1 ($\Leftrightarrow e_2 - e_1 = +1$), alors $\bar{Y}_{/E=e_2} - \bar{Y}_{/E=e_1} = \beta$.

Par conséquent, β s'interprète de la façon suivante dans un modèle de régression linéaire univarié : β est la différence moyenne, estimée à partir des données de l'échantillon, des valeurs de Y entre deux groupes d'animaux différant de +1 unité pour leur variable E, quelles que soient les valeurs de leur variable E.

Cette interprétation ci-dessus est fondamentale. Nous allons voir les conséquences d'une telle interprétation en fonction des différents types de variable (variable binaire, qualitative, et quantitative).

Par ailleurs, si β = 0, cela signifie que, en moyenne, il n'existe aucune différence de valeurs de Y entre deux groupes d'animaux différant de +1 unité pour leur variable E. Ainsi, β = 0 traduit bien une absence d'association entre le CdJ (quantifié par Y) et E. Si β ≤ 0,05, alors il existait une association significative entre le CdJ (quantifié par Y) et E dans l'échantillon.

b) Modèle de régression linéaire univarié avec une variable binaire

Supposons le modèle de régression linéaire suivant, incluant une seule variable binaire E :

$\bar{Y}_{/E} = \alpha + \beta.E$

Je recommande fortement le codage d'une variable binaire dans le fichier de données de telle façon à ce que les animaux *exposés* à E aient une valeur pour E égale à 1, et à ce que les animaux *non exposés* à E aient une valeur pour E égale à 0[6]. Ainsi, β est la différence moyenne, estimée à partir des données de l'échantillon, des valeurs de Y entre les animaux exposés et les animaux non exposés (car avec le codage fortement recommandé, $e_2 - e_1 = +1$).

[6] Le choix des catégories « exposé » et « non exposé » vous revient totalement. C'est un choix dicté par la clinique, et non pas par la statistique.

Par exemple, je vais faire tourner le modèle suivant :

$$\overline{CREAT}_{/FEMELLE} = \alpha + \beta.FEMELLE$$

Les lignes de programme ci-dessous permettent de faire tourner ce modèle dans SAS®.

```
PROC GLM DATA = Donnees_pour_guide;
MODEL CREAT = FEMELLE / SOLUTION CLPARM;
RUN;
```

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 37.

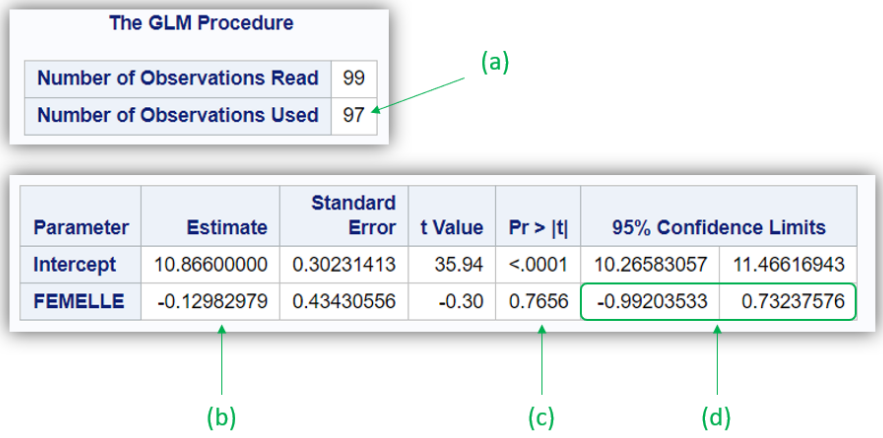


Figure 37

La flèche (a) sur la Figure 37 indique que le modèle a tourné sur 97 chiens (en raison des données manquantes sur les variables CREAT et/ou FEMELLE).

A partir des résultats présentés dans la colonne « Estimate » (cf. flèche (b) sur la Figure 37), le modèle de régression linéaire estimé par SAS® reliant la concentration en créatinine au sexe des chiens s'écrit de la façon suivante :

$$\overline{CREAT}_{/FEMELLE} = 10{,}87 - 0{,}13.FEMELLE$$

La valeur de β de « -0,13 » (en arrondissant) s'interprète de la façon suivante : la différence moyenne des valeurs de concentration en créatinine entre les chiens femelles (FEMELLE = 1) et les chiens mâles (FEMELLE = 0) vaut -0,13 g/L. Autrement dit, dans l'échantillon, les chiens femelles avaient, en moyenne, une concentration en créatinine inférieure de 0,13 g/L à celles des chiens mâles. Cette différence de -0,13 n'était pas significativement différente de 0 (p = 0,77 ; cf. Figure 37.c). Ainsi, il n'existait pas d'association significative dans l'échantillon entre la concentration en créatinine et le sexe des chiens. L'$IC_{95\%}$ de cette différence moyenne estimée est pointé par la flèche (d) sur la Figure 37 : $[-0{,}99 ; +0{,}73]_{95\%}$. Cet $IC_{95\%}$ comprenant « 0 », on retrouve le fait que la différence de -0,13 estimée n'était pas significativement différente de 0. N'oubliez pas que les résultats de cette régression linéaire ne sont valides que si la distribution de Y (ici, la concentration en créatinine) peut être considérée comme normale.

c) *Modèle de régression linéaire univarié avec une variable quantitative*

Supposons le modèle de régression linéaire suivant, incluant une seule variable quantitative E :

$$\bar{Y}_{/E} = \alpha + \beta.E$$

Vous allez voir ci-dessous que ce modèle repose sur une hypothèse que l'on appelle « l'hypothèse de la linéarité de l'association entre le CdJ (quantifié par Y) et la variable E ». Si cette hypothèse est vérifiée, le modèle est valide, et l'interprétation du coefficient β est possible. Si cette hypothèse n'est pas vérifiée, alors ce modèle ne doit pas être utilisé, car l'estimation de β ne sera pas interprétable.

Pour illustrer cette hypothèse de la linéarité de l'association entre Y et E, je vais faire tourner le modèle suivant à partir des données de l'échantillon :

$$\overline{CREAT}_{/AGE} = \alpha + \beta.AGE$$

Les lignes de programme ci-dessous permettent de faire tourner ce modèle dans SAS®.

```
PROC GLM DATA = Donnees_pour_guide;
MODEL CREAT = AGE / SOLUTION CLPARM;
RUN;
```

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 38.

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	9.856888642	0.71444750	13.80	<.0001	8.438531017	11.275246266
AGE	0.107978590	0.07775248	1.39	0.1682	-0.046379595	0.262336776

(a)

Figure 38

A partir du résultat de la régression linéaire présenté sur la Figure 38 (cf. Figure 38.a), le modèle de régression linéaire estimé par SAS® reliant la concentration en créatinine à l'âge des chiens s'écrit de la façon suivante :

$$\overline{CREAT}_{/AGE} = 9{,}86 + 0{,}11.AGE$$

Cela signifie qu'à partir des données de l'échantillon, le modèle estime que la différence moyenne de concentration en créatinine entre deux groupes de chiens différant de +1 année d'âge est égale à +0,11 g/L. Pourquoi « différant de +1 année » et pas « différant de +1 mois » ou « différant de +1 jour » d'âge ? Parce que la variable AGE est exprimée en années (l'unité de la variable AGE est donc l'année), et non pas en mois ou en jour. Et souvenez-vous de l'interprétation générale du coefficient β : « β est la différence moyenne des valeurs de Y entre deux groupes d'animaux différant de +1 *unité* pour leur variable E, quelles que soient les valeurs de leur variable E. ». Ainsi, le modèle estime qu'en moyenne, et dans l'échantillon, la différence de concentration en créatinine entre des chiens de (par exemple) 5 ans et des chiens de 4 ans est égale à +0,11 g/L. Le modèle estime aussi que la différence de concentration en créatinine entre des chiens de (par exemple) 14 ans et des chiens de 13 ans est *aussi* de +0,11 g/L. Ce modèle fait donc l'*hypothèse* qu'une augmentation de +1 année d'âge, quelle que soit la valeur de l'âge, se traduit par la *même* différence moyenne sur la concentration en créatinine : c'est ce que l'on appelle l'hypothèse de la linéarité de l'association entre Y et E (ici entre la concentration en créatinine et l'âge). Si dans la population cible, la différence moyenne de concentration en créatinine n'est pas la même, pour une même augmentation +1 année d'âge, entre des valeurs faibles de l'âge et des valeurs de l'âge plus élevées, alors l'hypothèse de la linéarité de l'association n'est pas vérifiée, et le modèle fournira une estimation de β ininterprétable.

(Si ce que j'ai écrit ci-dessus est du charabia, relisez une seconde puis éventuellement une troisième fois. Si ce que j'ai écrit *reste* du charabia après ces trois (voire plus) relectures, alors je vous recommande fortement de n'utiliser *que* des variables binaires dans vos analyses, car les interprétations sont en effet beaucoup plus faciles. Utiliser des variables quantitatives dans un modèle de régression sans comprendre l'hypothèse d'une telle utilisation vous expose à dire de sacrées belles bêtises.)

Cette hypothèse de la linéarité de l'association entre Y et E doit être vérifiée *avant* d'inclure une variable quantitative E dans un modèle de régression (linéaire) si l'on n'a pas de fortes raisons de penser qu'elle l'est dans la population cible. La vérification de cette hypothèse fait l'objet de la sous-partie C de cette partie VIII. « Modèles de régression ».

Attention (mais je vais écrire, autrement, quelque chose que j'ai déjà écrit dans la sous-partie « Vérification d'hypothèses sur lesquelles repose un modèle de régression » ci-dessus, page 40), si le modèle estime une valeur de coefficient β qui est ininterprétable parce que le modèle est incorrect, ce n'est pas à cause du logiciel ou de la statistique, c'est à cause de celle ou celui qui a choisi de faire tourner un tel modèle ! C'est pour cela que conduire des analyses statistiques nécessite de savoir toutes les conditions ou hypothèses sur lesquelles reposent ces analyses. N'attendez pas qu'un logiciel vous dise « attention, vous ne devriez pas faire tourner ce modèle, car il repose sur une hypothèse qui n'a pas l'air de tenir la route, biologiquement ou physiopathologiquement parlant » !

d) *Modèle de régression linéaire univarié avec une variable qualitative ordinale*

Supposons le modèle de régression linéaire suivant, incluant une seule variable qualitative ordinale E :

$$\bar{Y}_{/E} = \alpha + \beta . E$$

Vous allez voir ci-dessous que ce modèle, lui aussi, repose sur l'hypothèse de la linéarité de l'association entre le CdJ (quantifié par Y) et E. Là encore bien entendu, si cette hypothèse n'est pas vérifiée, alors ce modèle ne devra pas être utilisé, car l'estimation de β ne sera pas interprétable. Je vais illustrer cela à partir du modèle suivant que l'on va faire tourner à partir des données de l'échantillon :

$$\overline{CREAT}_{/CHOLES\_3CL} = \alpha + \beta . CHOLES\_3CL$$

Les lignes de programme ci-dessous permettent de faire tourner ce modèle dans SAS®.

```
PROC GLM DATA = Donnees_pour_guide;
MODEL CREAT = CHOLES_3CL / SOLUTION CLPARM;
RUN;
```

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 39.

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	10.88101293	0.36735923	29.62	<.0001	10.15171264	11.61031322
CHOLES_3CL	-0.08127155	0.30911248	-0.26	0.7932	-0.69493742	0.53239432

(a) (b)

Figure 39

A partir du résultat de la régression linéaire présenté sur la Figure 39 (cf. Figure 39.a), le modèle de régression linéaire estimé par SAS® reliant la concentration en créatinine à la cholestérolémie (en trois classes) s'écrit de la façon suivante :

$$\overline{CREAT}_{/CHOLES\_3CL} = 10{,}88 - 0{,}08 . CHOLES\_3CL$$

Cela signifie qu'à partir des données de l'échantillon, le modèle estime que la différence moyenne de concentration en créatinine entre deux groupes d'animaux différant de +1 unité pour la variable CHOLES_3CL est égale à -0,08 UI/L. Ainsi, le modèle estime que la différence moyenne de concentration en créatinine entre des chiens pour lesquels CHOLES_3CL = 2 (⇔ chiens avec une hypercholestérolémie) et des chiens pour lesquels CHOLES_3CL = 1 (⇔ chiens avec une normocholestérolémie) est égale à -0,08 UI/L. Ce modèle estime de la même façon que la différence

moyenne de concentration en créatinine entre des chiens pour lesquels CHOLES_3CL = 1 (⇔ chiens avec une normocholestérolémie) et des chiens pour lesquels CHOLES_3CL = 0 (⇔ chiens avec une hypocholestérolémie) est *là encore* de -0,08 UI/L. Notez que cette différence moyenne de la concentration en créatinine entre une classe de cholestérolémie et une autre consécutive n'est pas significative (p = 0,79 ; Figure 39.b).

De la même façon que pour une variable quantitative, inclure une variable qualitative ordinale fait l'hypothèse de la linéarité de l'association entre le CdJ et cette variable qualitative ordinale. Comme pour une variable quantitative, il faudra vérifier cette hypothèse avant d'inclure une variable qualitative ordinale dans un modèle de régression (linéaire) si l'on n'a pas de fortes raisons de penser que cette hypothèse est vérifiée dans la population cible.

e) *Modèle de régression linéaire univarié avec une variable qualitative nominale*

(1) *Problématique*

Supposons le modèle de régression linéaire suivant, incluant une seule variable qualitative nominale E :

$$\bar{Y}_{/E} = \alpha + \beta . E$$

Vous allez voir ci-dessous que ce modèle est tout simplement incorrect, car il fournit une estimation du coefficient β *systématiquement* ininterprétable. Je vais illustrer cela à partir du modèle suivant que l'on va faire tourner à partir des données de l'échantillon :

$$\overline{CREAT}_{/RACE\_4CL} = \alpha + \beta . RACE\_4CL$$

Les lignes de programme ci-dessous permettent de faire tourner ce modèle dans SAS®.

```
PROC GLM DATA = Donnees_pour_guide;
MODEL CREAT = RACE_4CL / SOLUTION CLPARM;
RUN;
```

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 40.

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	10.86296296	0.35703120	30.43	<.0001	10.15416641	11.57175952
RACE_4CL	-0.04537037	0.21479524	-0.21	0.8332	-0.47179283	0.38105209

(a)

Figure 40

A partir du résultat de la régression linéaire présenté sur la Figure 40 (cf. Figure 40.a), le modèle de régression linéaire estimé par SAS® reliant la concentration en créatinine à la race (en quatre classes) s'écrit de la façon suivante :

$$\overline{CREAT}_{/RACE\_4CL} = 10{,}86 - 0{,}05 . RACE\_4CL$$

Cela signifie qu'à partir des données de l'échantillon, le modèle estime que la différence moyenne de concentration en créatinine entre deux groupes d'animaux différant de +1 unité pour la variable RACE_4CL est égale à -0,05 g/L. Ainsi, le modèle estime que la différence moyenne de concentration en créatinine entre des chiens pour lesquels RACE_4CL = 3 (⇔ autre race) et des chiens pour lesquels RACE_4CL = 2 (⇔ race croisée Golden/Labrador), que la différence moyenne de concentration en créatinine entre des chiens pour lesquels RACE_4CL = 2 (⇔ race croisée Golden/Labrador) et des chiens pour lesquels RACE_4CL = 1 (⇔ race Labrador), et que la différence moyenne de concentration

en créatinine entre des chiens pour lesquels RACE_4CL = 1 (⇔ race Labrador) et des chiens pour lesquels RACE_4CL = 0 (⇔ race Golden), valent *toutes* -0,05 g/L. Cette estimation de -0,05 g/L n'a aucun sens, car elle repose sur l'hypothèse, n'ayant *aucun* fondement biologique ou physiologique, d'une *même* différence moyenne de concentration en créatinine lorsque la variable RACE_4CL augmente de +1 unité.

Il est par conséquent *interdit* d'inclure *telle quelle* une variable qualitative nominale dans un modèle de régression (quel que soit le modèle de régression), car ce modèle reposerait sur le fait que l'augmentation de +1 unité de cette variable qualitative *nominale* a un sens, alors qu'il n'en a aucun. En effet, le choix du chiffre attribué à une classe d'une variable qualitative nominale est purement arbitraire, et n'est en aucun cas fondé sur un ordre quelconque (« 1 », plus petit que « 2 », lui-même plus petit que « 3 », etc.).

(2) Création de « variables indicatrices » et interprétation théorique

Pour étudier l'association entre un CdJ et une variable qualitative nominale à *K* classes à l'aide d'un modèle de régression (quel que soit le modèle de régression), il faut créer *K* variables binaires (appelées par la suite « variables indicatrices ») à partir des valeurs de la variable qualitative nominale, puis inclure *K*-1 de ces variables indicatrices dans le modèle. Ce que je viens d'écrire peut être écrit autrement : parmi les *K* variables indicatrices créées, il faut en exclure une du modèle et inclure toutes les autres. La variable indicatrice que l'on *choisit* de ne pas inclure dans le modèle est considérée comme la *classe de référence* de la variable qualitative nominale initiale.

Le tableau ci-dessous présente la façon d'attribuer les valeurs 0 et 1 à ces variables indicatrices binaires, nommées VAR_IND_i, à partir d'une variable qualitative nommée VAR_QUAL : les « 1 » sont sur la diagonale, et les « 0 » sont partout ailleurs.

VAR_QUAL	VAR_IND_1	VAR_IND_2	…	VAR_IND_i	…	VAR_IND_K
1	1	0	…	0	…	0
2	0	1	…	0	…	0
…	…	…	…	…	…	…
i	0	0	…	1	…	0
…	…	…	…	…	…	…
K	0	0	…	0	…	1

Reprenons l'exemple de la variable RACE_4CL. Dans la mesure où cette variable comprend quatre classes, il y aura quatre variables indicatrices, que je vais choisir de nommer : GOLDEN, LABRADOR, CROISEE, AUTRE_RACE (ces variables indicatrices ont été créées dans le fichier de données Excel®). L'attribution des valeurs 0 et 1 pour chacune de ces quatre variables est décrite ci-dessous.

RACE_4CL	GOLDEN	LABRADOR	CROISEE	AUTRE_RACE
0	1	0	0	0
1	0	1	0	0
2	0	0	1	0
3	0	0	0	1

Ainsi, un chien de race Labrador (RACE_4CL = 1) se verra attribuer les valeurs de 0, 1, 0, et 0, respectivement pour les variables binaires GOLDEN, LABRADOR, CROISEE, et AUTRE_RACE.

J'ai écrit ci-dessus qu'il fallait ensuite inclure *K*-1 variables indicatrices parmi les *K* créées. Dans l'exemple, il faut donc inclure trois parmi les quatre variables indicatrices créées (GOLDEN, LABRADOR, CROISEE, et AUTRE_RACE). Comme je l'ai écrit, le choix de la variable indicatrice qui ne sera pas incluse dans le modèle vous revient totalement. Supposons que l'on choisisse de ne pas inclure la variable GOLDEN dans le modèle. Le modèle est donc celui-ci :

$$\overline{CREAT}_{/LABRADOR,CROISEE,AUTRE\_RACE} = \alpha + \beta_{LAB}.LABRADOR + \beta_{CROISEE}.CROISEE + \beta_{AUTRE\_R}.AUTRE\_RACE$$

Pour interpréter chacun des trois coefficients du modèle (β_{LAB}, $\beta_{CROISEE}$, β_{AUTRE_R},), je vais écrire le modèle pour chaque race de chien.

Pour les chiens de race Golden, les variables LABRADOR, CROISEE, et AUTRE_RACE valent toutes les trois 0. Ainsi, pour les chiens de race Golden, le modèle s'écrit :

$$\overline{CREAT}_{/LABRADOR=0,CROISEE=0,AUTRE\_RACE=0} = \alpha + \beta_{LAB} \times 0 + \beta_{CROISEE} \times 0 + \beta_{AUTRE\_R} \times 0 = \alpha$$

Pour les chiens de race Labrador, le modèle s'écrit :

$$\overline{CREAT}_{LABRADOR=1,CROISEE=0,AUTRE\_RACE=0} = \alpha + \beta_{LAB} \times 1 + \beta_{CROISEE} \times 0 + \beta_{AUTRE\_R} \times 0 = \alpha + \beta_{LAB}$$

Pour les chiens de race croisée Golden/Labrador, le modèle s'écrit :

$$\overline{CREAT}_{LABRADOR=0,CROISEE=1,AUTRE\_RACE=0} = \alpha + \beta_{LAB} \times 0 + \beta_{CROISEE} \times 1 + \beta_{AUTRE\_R} \times 0 = \alpha + \beta_{CROISEE}$$

Pour les chiens d'autre race, le modèle s'écrit :

$$\overline{CREAT}_{LABRADOR=0,CROISEE=0,AUTRE\_RACE=1} = \alpha + \beta_{LAB} \times 0 + \beta_{CROISEE} \times 0 + \beta_{AUTRE\_R} \times 1 = \alpha + \beta_{AUTRE\_R}$$

Comme précédemment, je vais faire la soustraction de deux modèles pour interpréter chaque coefficient β.

$$\overline{CREAT}_{/LABRADOR=1,CROISEE=0,AUTRE\_RACE=0} - \overline{CREAT}_{/LABRADOR=0,CROISEE=0,AUTRE\_RACE=0} = \beta_{LAB}$$

$$\overline{CREAT}_{/LABRADOR=0,CROISEE=1,AUTRE\_RACE=0} - \overline{CREAT}_{/LABRADOR=0,CROISEE=0,AUTRE\_RACE=0} = \beta_{CROISEE}$$

$$\overline{CREAT}_{/LABRADOR=0,CROISEE=0,AUTRE\_RACE=1} - \overline{CREAT}_{/LABRADOR=0,CROISEE=0,AUTRE\_RACE=0} = \beta_{AUTRE\_R}$$

Ainsi, β_{LAB} est la différence moyenne de concentration en créatinine entre les chiens de race Labrador et les chiens de race Golden, $\beta_{CROISEE}$ est la différence moyenne de concentration en créatinine entre les chiens de race croisée et les chiens de race Golden, et β_{AUTRE_R} est la différence moyenne de concentration en créatinine entre les chiens d'autre race et les chiens de race Golden.

Vous venez de voir que chacun des coefficients du modèle compare une classe de la variable RACE_4CL à la *classe de référence* qui correspond à la variable indicatrice qui n'a pas été incluse dans le modèle, à savoir ici la classe « Golden ». Si l'on avait choisi de ne pas inclure la variable indicatrice LABRADOR, alors chacun des trois coefficients du modèle aurait quantifié la différence moyenne de concentration en créatinine entre chacune des classes de la variable RACE_4CL et la classe de référence correspondant à la race Labrador.

(3) Mise en pratique avec SAS®

Vous allez voir désormais comment inclure des variables indicatrices dans un modèle de régression (ici, linéaire, mais la démarche sera exactement la même pour la régression logistique et le modèle de Cox), pour étudier l'association entre un CdJ (ici quantitatif, la concentration en créatinine) et une variable qualitative nominale (ici la variable RACE_4CL en quatre classes correspondant à la race des chiens), avec SAS®.

Il se trouve que les *K* variables indicatrices peuvent être créées par SAS®, sans avoir besoin de les créer au préalable dans le fichier Excel®.

Les lignes de programme ci-dessous permettent d'inclure la variable RACE_4CL sous forme de variables indicatrices, en choisissant la classe « 0 » pour RACE_4CL comme classe de référence, ce qui correspond à la race Golden. C'est-à-dire que SAS® va, en quelque sorte, créer les variables indicatrices équivalentes aux variables LABRADOR, CROISEE, et AUTRE_RACE que j'ai mentionnées ci-dessus.

```
PROC GLM DATA = Donnees_pour_guide;
CLASS RACE_4CL (REF = '0');
MODEL CREAT = RACE_4CL / SOLUTION CLPARM;
RUN;
```

Les résultats (extrait) des lignes de programme ci-dessus sont présentés sur la Figure 41.

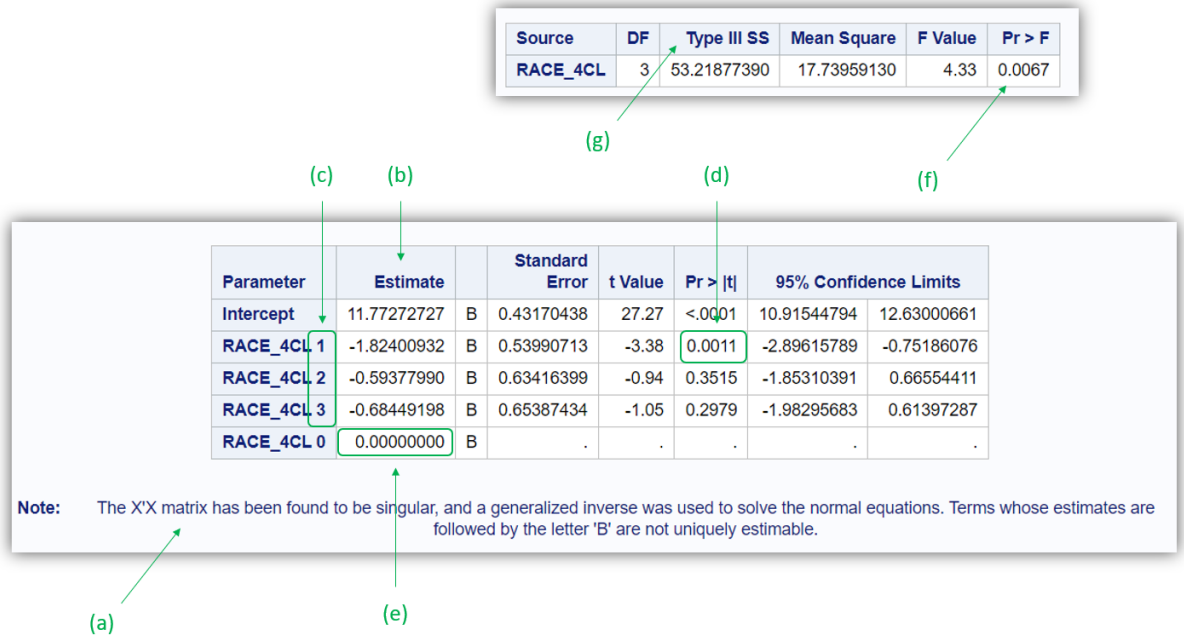


Figure 41

Le message pointé par la flèche (a) sur la Figure 41 ne doit pas vous effrayer. Il sera toujours présent lorsque vous utiliserez la procédure PROC GLM et que vous inclurez dans le modèle de régression linéaire une variable qualitative sous forme de variables indicatrices en utilisant l'instruction « SOLUTION » sur la ligne « MODEL ». Ce n'est pas une erreur, et cela ne veut pas dire que les estimations des coefficients ne sont pas correctes.

Pour la suite de ce guide, et dans la mesure où je vais souvent inclure dans un modèle de régression une variable qualitative (nominale ou ordinale) sous forme de variables indicatrices, voici ci-dessous comment je vais choisir d'écrire le modèle qui correspond aux lignes de programme ci-dessus (je mets entre parenthèses la variable qualitative incluse dans le modèle sous forme de variables indicatrices, en mettant en indice et entre guillemets la classe de référence choisie) :

$$\overline{CREAT}_{/(RACE\_4CL)_{"0"}} = \alpha + \beta_1.RACE\_4CL(1) + \beta_2.RACE\_4CL(2) + \beta_3.RACE\_4CL(3)$$

A partir des résultats présentés dans la colonne « Estimate » (cf. flèche (b) sur la Figure 41), le modèle de régression linéaire estimé par SAS® reliant la concentration en créatinine à la race des chiens s'écrit de la façon suivante (j'ai mis les chiffres « 1 », « 2 », et « 3 » entre parenthèses ci-dessous, pour davantage de clarté – dans SAS®, ces parenthèses ne sont pas présentes, cf. Figure 41.c) :

$$\overline{CREAT}_{/(RACE\_4CL)_{"0"}} = 11{,}77 - 1{,}82.RACE\_4CL(1) - 0{,}59.RACE\_4CL(2) - 0{,}68.RACE\_4CL(3)$$

Là, il va être fondamental de bien comprendre ces résultats. Car sinon, vous pourriez vraiment dire de belles bêtises à partir des degrés de signification présentés sur la Figure 41. Notamment, il faut comprendre ces « 1 », « 2 », et « 3 » à droite de « RACE_4CL » (cf. Figure 41.c).

Dans la mesure où la variable qualitative nominale RACE_4CL a été incluse sous forme de variables indicatrices (et c'est obligatoire de faire cela car il s'agit d'une variable qualitative *nominale*), chaque coefficient β_i du modèle (ici, $i \in \{1,2,3\}$) quantifie la différence moyenne de concentration en créatinine entre une des classes de la variable qualitative et la *classe de référence*. La classe de référence choisie est celle correspondant à la valeur « 0 » pour la variable RACE_4CL (instruction « CLASS RACE_4CL (REF = '0') » dans la ligne de programme ci-dessus).

Ainsi, la valeur de β_1 de -1,82 correspond à la différence moyenne de concentration en créatinine entre les chiens pour lesquels la variable RACE_4CL vaut 1 (⇔ les chiens de race Labrador) et les chiens pour lesquels la variable RACE_4CL vaut 0 (⇔ les chiens de race Golden, la classe de référence). Et comme

il est indiqué « 1 » devant RACE_4CL (cf. Figure 41.c), c'est bien une comparaison des chiens de race Labrador *par rapport* aux chiens de race Golden et non les chiens de race Golden *par rapport* aux chiens de race Labrador. Cette différence étant négative, la concentration en créatinine était en moyenne moins élevée parmi les chiens de race Labrador que parmi les chiens de race Golden. Cette différence négative de -1,82 g/L était d'ailleurs significative : la valeur du degré de signification testant la différence moyenne de concentration en créatinine entre les deux races (Labrador *versus* Golden) était de 0,0011 (cf. Figure 41.d), donc inférieure à 0,05. Par conséquent, on peut aussi interpréter cette valeur de -1,82, significativement différente de 0, de la façon suivante : les chiens de race Labrador avaient, en moyenne, une concentration en créatinine significativement inférieure à celle des chiens de race Golden (Δ = -1,82 g/L ; p < 0,01).

De même, la valeur de β_2 de -0,54 correspond à la différence moyenne de concentration en créatinine entre les chiens pour lesquels la variable RACE_4CL vaut 2 ($\Leftrightarrow$ les chiens de race croisée Golden/Labrador) et les chiens pour lesquels la variable RACE_4CL vaut 0 ($\Leftrightarrow$ les chiens de race Golden, la classe de référence). Cette différence étant là encore négative, mais non significative (p = 0,35), la concentration en créatinine était, en moyenne, moins élevée parmi les chiens de race croisée Golden/Labrador que parmi les chiens de race Golden (Δ = -0,59 g/L ; p = 0,35).

Enfin, la valeur de β_3 de -0,68 correspond à la différence moyenne de concentration en créatinine entre les chiens pour lesquels la variable RACE_4CL vaut 3 (les chiens d'autre race) et les chiens pour lesquels la variable RACE_4CL vaut 0 (les chiens de race Golden, la classe de référence). Cette différence étant là encore négative et toujours non significative, la concentration en créatinine était, en moyenne, moins élevée parmi les chiens d'autre race que parmi les chiens de race Golden (Δ = -0,68 g/L ; p = 0,30).

Vous comprenez maintenant pourquoi la colonne « ESTIMATE » comprend la valeur 0,0000 sur la ligne « RACE_4CL 0 » (cf. Figure 41.e) : la différence moyenne de concentration en créatinine entre les chiens pour lesquels la variable RACE_4CL vaut 0 et les chiens pour lesquels la variable RACE_4CL vaut 0 est évidemment nulle ! (Et c'est ce « 0,0000 » qui ne plait pas à SAS® et qui conduit au message pointé par la flèche (a) sur la Figure 41.)

Pour information, les lignes de programme ci-dessous utilisent les variables indicatrices LABRADOR, CROISEE, et AUTRE_RACE qui avaient été (inutilement) créées dans le fichier Excel® pour étudier l'association entre la concentration en créatinine et la race des chiens.

```
PROC GLM DATA = Donnees_pour_guide;
MODEL CREAT = LABRADOR CROISEE AUTRE_RACE / SOLUTION CLPARM;
RUN;
```

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 43.

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	11.77272727	0.43170438	27.27	<.0001	10.91544794	12.63000661
LABRADOR	-1.82400932	0.53990713	-3.38	0.0011	-2.89615789	-0.75186076
CROISEE	-0.59377990	0.63416399	-0.94	0.3515	-1.85310391	0.66554411
AUTRE_RACE	-0.68449198	0.65387434	-1.05	0.2979	-1.98295683	0.61397287

(a)

Figure 42

On retrouve bien évidemment que les valeurs des coefficients du modèle pointés par la flèche (a) sur la Figure 43 sont identiques à celles pointées par la flèche (b) sur la Figure 41.

Les tests statistiques testant individuellement chaque coefficient β_i (par exemple celui pointé par la flèche (d) sur la Figure 41 pour le coefficient β_1 associé à la variable indicatrice RACE_4CL(1)) ne permettent pas de savoir si, *globalement*, la race est, ou n'est pas, significativement associée à la concentration en créatinine. Pour cela, il faut lire la valeur du degré de signification pointée par la flèche (f) sur la Figure 41. Attention, il faut lire cette valeur de degré de signification dans le tableau qui comprend « Type III SS » sur la 1ère ligne (cf. Figure 41.g).

Vous vous êtes normalement rendu compte que l'interprétation des résultats fournis par SAS® lorsqu'une variable qualitative nominale est incluse dans un modèle (sous forme de variables indicatrices) dépend totalement du choix de la classe de référence. Par exemple, si on avait voulu que ce soit les chiens de race croisée (RACE_4CL = 2) qui soient considérés comme « classe de référence », il aurait fallu faire tourner les lignes de programme ci-dessous.

```
PROC GLM DATA = Donnees_pour_guide;
CLASS RACE_4CL (REF = '2');
MODEL CREAT = RACE_4CL / SOLUTION CLPARM;
RUN;
```

3. Interprétation des résultats d'une régression linéaire multivariée

a) *Interprétation générale*

Si, maintenant, le modèle inclut deux variables E_1 et E_2, il devient alors :

$$\bar{Y}_{/E_1,E_2} = \alpha + \beta_1.E_1 + \beta_2.E_2$$

L'interprétation de β_1 est la suivante : « β_1 est la différence moyenne ajustée sur E_2, estimée à partir des données de l'échantillon, des valeurs de Y entre deux groupes d'animaux différant de +1 unité pour leur variable E_1, quelles que soient les valeurs de leur variable E_1. » L'expression « ajustée sur E_2 » est équivalente à « indépendamment de E_2 » ou « après avoir pris en compte E_2 ». Je ne vais pas faire, dans ce guide, la démonstration qui prouve que le fait d'inclure la variable E_2 dans le modèle conduit à ce que β_1 quantifie l'association entre E_1 et Y *ajustée* sur E_2.

Si, enfin, le modèle inclut N variables E_1, E_2, …, E_N, il devient alors :

$$\bar{Y}_{/E_1,E_2,\ldots,E_N} = \alpha + \sum_{i=1}^{N} \beta_i.E_i$$

L'interprétation de β_i devient : « β_i est la différence moyenne ajustée sur toutes les autres variables incluses dans le modèle, estimée à partir des données de l'échantillon, des valeurs de Y entre deux groupes d'animaux différant de +1 unité pour leur variable E_i, quelles que soient les valeurs de leur variable E_i. »

L'ordre des variables incluses dans un modèle multivarié n'a aucune importance. Ainsi, que l'on souhaite étudier l'association entre Y et E_1, ajustée sur E_2 et E_3, ou bien que l'on souhaite étudier l'association entre Y et E_2, ajustée sur E_1 et E_3, dans les deux cas le modèle de régression (linéaire) sera celui-ci :

$$\bar{Y}_{/E_1,E_2,E_3} = \alpha + \beta_1.E_1 + \beta_2.E_2 + \beta_3.E_3$$

b) *En pratique avec SAS®*

Supposons que l'on veuille étudier l'association entre la concentration en créatinine et le sexe des chiens, ajustée sur l'âge et la race des chiens. Si l'on choisit la race Golden (RACE_4CL = 0) comme classe de référence, le modèle de régression linéaire multivarié correspondant est donc celui-ci :

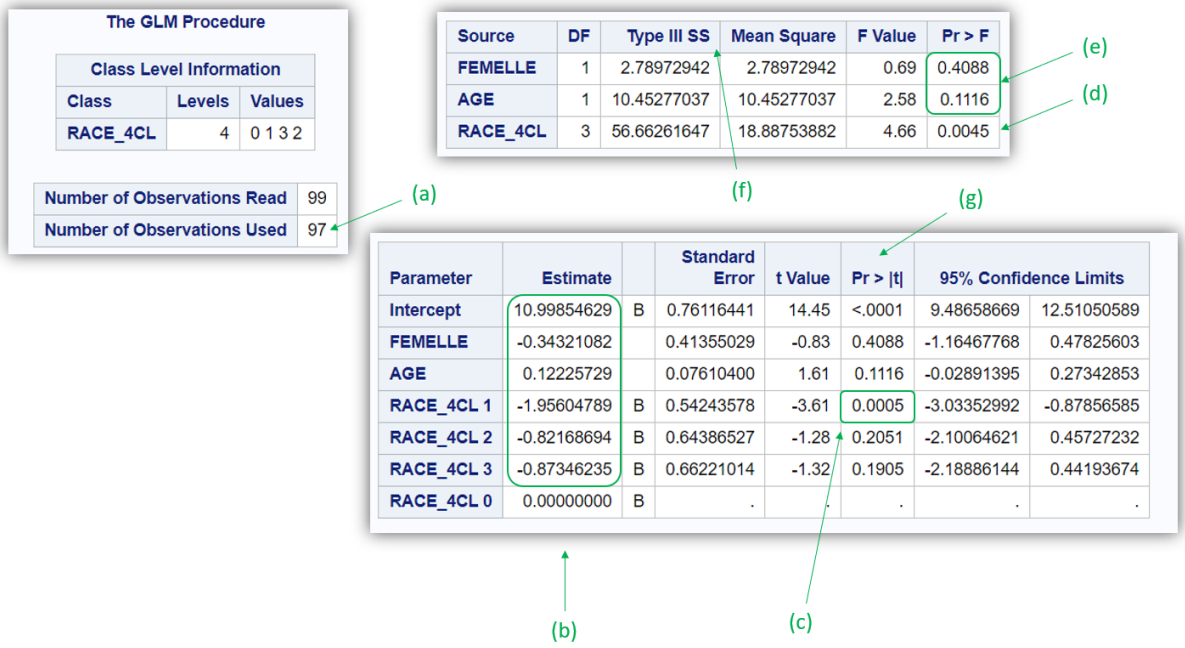
$$\overline{CREAT}_{/FEMELLE,AGE,(RACE\_4CL)_{"0"}} = \alpha + \beta.FEMELLE + \gamma.AGE + \delta_1.RACE\_4CL(1) + \delta_2.RACE\_4CL(2) + \delta_3.RACE\_4CL(3)$$

Je rappelle vivement deux choses. Tout d'abord, dans le modèle ci-dessus, la variable qualitative RACE_4CL doit obligatoirement être incluse dans le modèle sous forme de variables indicatrices parce qu'elle est *nominale*. Ensuite, *tous* les coefficients du modèle ci-dessus (β, γ, δ_1, δ_2, et δ_3) ne sont interprétables que si le modèle repose sur les deux hypothèses suivantes vérifiées dans la population cible : (1) la concentration en créatinine suit une loi normale, et (2) l'association entre l'âge et la concentration en créatinine est linéaire.

Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus.

```
PROC GLM DATA = Donnees_pour_guide;
CLASS RACE_4CL (REF = '0');
MODEL CREAT = FEMELLE AGE RACE_4CL / SOLUTION CLPARM;
RUN;
```

Les résultats (extrait) des lignes de programme ci-dessus sont présentés sur la Figure 43.



The GLM Procedure

Class Level Information

Class	Levels	Values
RACE_4CL	4	0 1 3 2

Number of Observations Read	99
Number of Observations Used	97

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FEMELLE	1	2.78972942	2.78972942	0.69	0.4088
AGE	1	10.45277037	10.45277037	2.58	0.1116
RACE_4CL	3	56.66261647	18.88753882	4.66	0.0045

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	10.99854629	B	0.76116441	14.45	<.0001	9.48658669	12.51050589
FEMELLE	-0.34321082		0.41355029	-0.83	0.4088	-1.16467768	0.47825603
AGE	0.12225729		0.07610400	1.61	0.1116	-0.02891395	0.27342853
RACE_4CL 1	-1.95604789	B	0.54243578	-3.61	0.0005	-3.03352992	-0.87856585
RACE_4CL 2	-0.82168694	B	0.64386527	-1.28	0.2051	-2.10064621	0.45727232
RACE_4CL 3	-0.87346235	B	0.66221014	-1.32	0.1905	-2.18886144	0.44193674
RACE_4CL 0	0.00000000	B	.	.	.	.	.

Figure 43

A partir des résultats de la régression linéaire présentés sur la Figure 43, le modèle de régression linéaire estimé par SAS® à partir des 97 chiens pour lesquels aucune donnée ne manquait sur la concentration en créatinine et sur les trois variables incluses dans le modèle (cf. Figure 43.a), et reliant la concentration en créatinine au sexe, à l'âge, et à la race, s'écrit de la façon suivante (à partir de la colonne « Estimate », cf. Figure 43.b) :

$$\overline{CREAT}_{/FEMELLE,AGE,(RACE\_4CL)_{"0"}} = 11{,}00 - 0{,}34.FEMELLE + 0{,}12.AGE - 1{,}96.RACE\_4CL(1) - 0{,}82.RACE\_4CL(2) - 0{,}87.RACE\_4CL(3)$$

Je vais faire l'hypothèse que l'association entre l'âge et la concentration en créatinine est linéaire. Interprétons, maintenant que l'on peut le faire, chacun des coefficients associés aux variables incluses

dans le modèle multivarié (cf. Figure 43.b). La valeur de -0,34 devant la variable binaire FEMELLE signifie qu'indépendamment de l'âge et de la race, dans l'échantillon, les chiens femelles avaient une concentration en créatinine en moyenne inférieure de 0,34 g/L à celle des chiens mâles. La valeur de 0,12 devant la variable quantitative AGE signifie qu'indépendamment du sexe et de la race, dans l'échantillon, une augmentation de +1 année d'âge est associée à une augmentation de la concentration en créatinine de, en moyenne, 0,12 g/L. La valeur de -1,95 devant la variable indicatrice RACE_4CL(1) signifie qu'indépendamment du sexe et de l'âge, dans l'échantillon, les chiens de race Labrador (RACE_4CL = 1) avaient une concentration en créatinine en moyenne inférieure de 1,95 g/L à celle des chiens de race Golden (RACE_4CL = 0, qui est la classe de référence ; cette différence de 1,95 g/L entre les deux races était d'ailleurs significative, cf. Figure 43.c). La valeur de -0,82 devant la variable indicatrice RACE_4CL(2) signifie qu'indépendamment du sexe et de l'âge, dans l'échantillon, les chiens de race croisée Golden/Labrador (RACE_4CL = 2) avaient une concentration en créatinine en moyenne inférieure de 0,82 g/L à celle des chiens de race Golden. La valeur de -0,87 devant la variable indicatrice RACE_4CL(3) signifie qu'indépendamment du sexe et de l'âge, dans l'échantillon, les chiens d'autre race (RACE_4CL = 3) avaient une concentration en créatinine en moyenne inférieure de 0,87 g/L à celle des chiens de race Golden.

Globalement, indépendamment du sexe et de l'âge, la race était significativement associée à la concentration en créatinine ($p < 0,01$; cf. Figure 43.d). On peut d'ailleurs remarquer que les valeurs des deux degrés de signification pointés par la flèche (e) sur la Figure 43, fournis dans le tableau « Type III SS » (cf. Figure 43.f) pour les variables FEMELLE et AGE sont identiques à celles dans la colonne « Pr > |t| » (cf. Figure 43.g), ce qui est tout à fait attendu. En effet, lorsque la variable est incluse telle quelle dans un modèle de régression (c'est-à-dire sans que ce soit sous forme de variables indicatrices), il n'y a qu'un seul coefficient β pour la variable : le test global (cf. Figure 43.e) correspond au test local (cf. Figure 43.g).

C. Vérification de l'hypothèse de la linéarité de l'association

1. Introduction

La première question à se poser avant de vérifier l'hypothèse de la linéarité de l'association entre une variable qualitative ordinale ou quantitative et le CdJ (quel qu'en soit le type : quantitatif ou binaire, assorti ou non d'un temps de survie) est de savoir s'il y a des raisons biologiques ou physiopathologiques qui pourraient laisser penser que cette association n'est pas linéaire dans la population cible. Et il est fondamental de tenter de répondre à cette question *avant* la vérification de l'hypothèse de la linéarité de l'association. En effet, comme je l'ai déjà écrit dans la sous-partie « Vérification d'hypothèses sur lesquelles repose un modèle de régression » (page 40), il faut garder en tête que ce que l'on peut observer dans un échantillon peut être éloigné de ce qu'il se passe dans la population cible, entre autres à cause de la fluctuation d'échantillonnage. Par conséquent, il ne faut pas *trop* attendre des données de l'échantillon qu'elles nous disent si l'association est, ou n'est pas, linéaire dans la population cible.

La démarche de vérification de l'hypothèse de la linéarité de l'association entre l'état de santé Y et une variable qualitative ordinale ou quantitative peut être considérée comme fastidieuse. Elle est néanmoins indispensable à réaliser si l'on souhaite inclure dans un modèle de régression (quel qu'il soit) une telle variable. Inclure une variable qualitative ordinale ou quantitative dans un modèle et interpréter les résultats de ce modèle sans avoir vérifié au préalable cette hypothèse de la linéarité de l'association expose l'auteur.e de ce modèle à des interprétations fausses, et par conséquent à des erreurs de communication scientifique.

Si, après avoir lu ce qui suit, vous trouvez effectivement que la démarche est *trop* fastidieuse, alors vous n'avez plus qu'une seule solution : utiliser des variables uniquement binaires. Cela signifie que si

vos variables d'intérêt et/ou vos facteurs de confusion ne sont pas des variables binaires, vous *devrez* les recoder en variables binaires, selon un seuil déjà décrit dans la littérature, ou bien selon un seuil qui a cliniquement du sens, ou enfin selon la médiane ou selon le premier ou troisième quartile. L'inconvénient de rendre binaire une variable initialement qualitative ordinale ou quantitative est entre autres décrit dans les articles suivants (Altman and Royston, 2006; Brenner and Blettner, 1997; Royston et al., 2006).

Je vais présenter ci-dessous une démarche pour vérifier l'hypothèse de la linéarité de l'association entre une exposition qualitative ordinale ou quantitative et un CdJ quantitatif, en utilisant la régression linéaire. Mais cette démarche est absolument identique quel que soit le modèle de régression (et donc avec un CdJ binaire, assorti ou non d'un temps de survenue).

2. Cas d'une variable qualitative ordinale

a) Aspect théorique

Soit VAR_QUAL_K_CL une variable qualitative ordinale à *K* classes, codée 0, 1, …, *K*-1 dans le fichier de données. Pour vérifier que l'association entre le CdJ quantitatif (quantifié par Y) et la variable VAR_QUAL_K_CL est linéaire (c'est-à-dire, pour vérifier qu'une augmentation de +1 unité de VAR_QUAL_K_CL se traduit par une même augmentation sur Y, quelle que soit la valeur de VAR_QUAL_K_CL), je vous recommande de suivre la démarche suivante :

1) Inclure VAR_QUAL_K_CL sous forme de variables indicatrices en choisissant comme classe de référence la plus petite valeur de VAR_QUAL_K_CL ;

2) Noter les valeurs des coefficients β_i, ainsi que la Standard Error de β_i ($SE_{\beta i}$) de chacune des *K*-1 variables indicatrices ;

3) Placer sur un graphique *K*-1 points, chacun ayant pour ordonnée la valeur du coefficient β_i d'une variable indicatrice et pour abscisse la valeur représentant la classe concernée de la variable VAR_QUAL_K_CL, puis placez verticalement autour de chacun de ces points l'$IC_{95\%}$ de β_i ;

4) Placer en plus le point d'ordonnée 0 et d'abscisse la valeur représentant la 1ère classe de variable VAR_QUAL_K_CL (la classe de référence) ;

5) Vérifier visuellement que tous les points du graphique (incluant le point d'ordonnée 0) sont relativement bien alignés sur une droite.

Si l'association entre le CdJ (quantifié par Y) et VAR_QUAL_K_CL était *parfaitement* linéaire, la Figure 44 présente ce que l'on devrait obtenir comme valeurs des *K*-1 coefficients β_i ($i \in \{1, \ldots, K\text{-}1\}$) : ils devraient être tous alignés sur une droite (en passant par le point d'abscisse la valeur de la 1ère classe, ici 0, et d'ordonnée 0, comme l'étape 4 ci-dessus le demande).

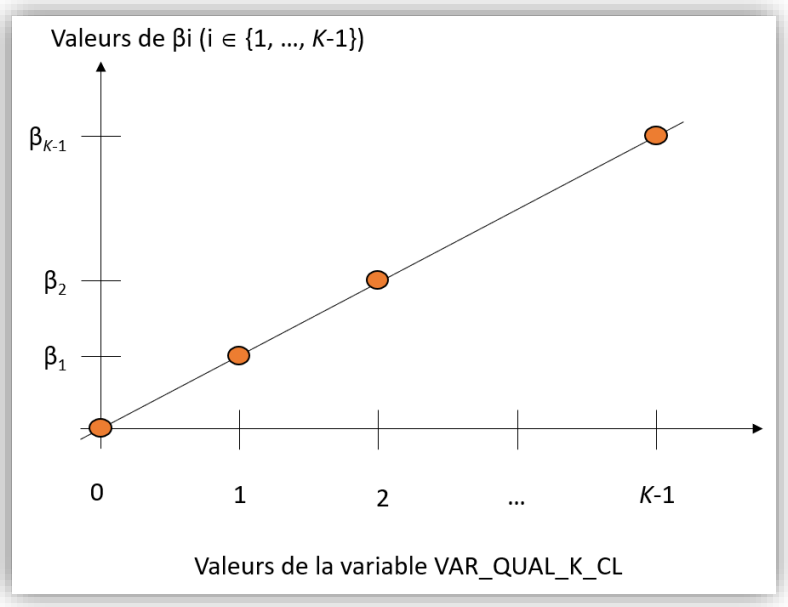


Figure 44

Je vais passer un peu de temps sur l'étape 5 ci-dessus. Tout d'abord, vous pouvez tout à fait vous aider des $IC_{95\%}$ des coefficients β_i qui donnent une indication de la précision avec laquelle ces coefficients β_i ont été estimés. En effet, la fluctuation d'échantillonnage peut conduire à des estimations des coefficients β_i éloignées des valeurs réelles, et ce d'autant plus que chaque classe de la variable qualitative ordinale est composée de peu d'individus. Ainsi, les coefficients β_i peuvent ne pas être alignés non pas parce que l'association n'est réellement pas linéaire, mais tout simplement à cause d'une imprécision des estimations des coefficients β_i à partir des données de l'échantillon. La Figure 45 présente deux situations, avec des valeurs des coefficients β_i identiques, mais avec des $SE_{\beta i}$ (donc des $IC_{95\%}$, représentés sur la figure) différents.

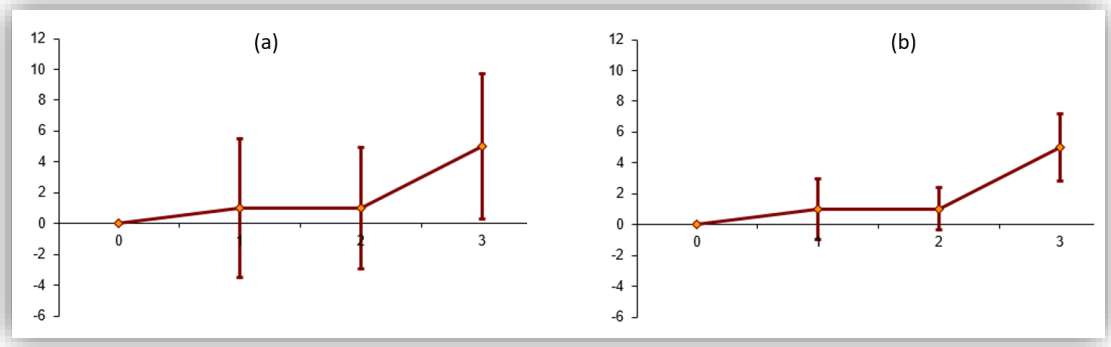


Figure 45

La situation (a) de la Figure 45 est celle où l'on pourrait accepter l'hypothèse de la linéarité de l'association dans le cas de figure où l'on n'aurait aucune raison de penser qu'en vrai, l'association n'est pas linéaire. En effet, bien que les quatre points ne soient pas vraiment alignés, la largeur des $IC_{95\%}$ laisse penser que ce non alignement semble davantage dû à une imprécision des trois estimations des coefficients qu'à une non linéarité réelle. La situation (b) laisse en revanche penser que l'association n'est pas linéaire dans la population cible.

Si l'hypothèse de la linéarité de l'association est vérifiée pour VAR_QUAL_K_CL, alors il est possible de faire tourner le modèle de régression (linéaire) incluant cette variable telle quelle : le coefficient β (unique) associé à VAR_QUAL_K_CL sera interprétable.

Si l'hypothèse de la linéarité de l'association n'est pas vérifiée, alors il faut soit inclure la variable VAR_QUAL_K_CL sous forme de variables indicatrices, soit rendre binaire la variable VAR_QUAL_K_CL, en regroupant les classes entre elles. Ce regroupement doit d'abord être guidé par la « clinique ». Ce regroupement doit en effet, et avant tout, être cliniquement pertinent. S'il est guidé par la représentation graphique des points, c'est dangereux. En effet, recoder des variables doit se faire *a priori*, et non pas *après* avoir vu les résultats. La démonstration serait trop longue ici, et sort de toute façon du cadre de ce guide.

Sur la Figure 44, j'ai utilisé comme abscisses des points les valeurs des classes de VAR_QUAL_K_CL. Dans le cas où VAR_QUAL_K_CL a été codée à partir d'une variable quantitative (comme c'est le cas pour la variable UREE_4CL du fichier de données), vous verrez plus loin dans ce guide qu'au lieu d'utiliser les valeurs des classes de VAR_QUAL_K_CL, je vous recommande d'utiliser le « centre » des classes de VAR_QUAL_K_CL. Ce « centre » d'une classe i de la variable VAR_QUAL_K_CL pourra être la médiane de la variable quantitative (dont est issue VAR_QUAL_K_CL) calculée parmi tous les individus pour lesquels VAR_QUAL_K_CL = i.

b) *En pratique avec SAS®*

Prenons l'exemple de l'association entre la concentration en créatinine et la cholestérolémie (variable qualitative ordinale CHOLES_3CL en trois classes : hypocholestérolémie, normocholestérolémie, et hypercholestérolémie). En incluant la variable CHOLES_3CL telle quelle dans le modèle, celui-ci s'écrit ainsi :

$$\overline{CREAT}_{/CHOLES\_3CL} = \alpha + \beta.CHOLES\_3CL$$

Comme vous l'avez vu ci-dessus (cf. Figure 39), l'estimation de β de valeur égale à -0,08 n'a de sens que si l'association entre la concentration en créatinine et la variable CHOLES_3CL est biologiquement ou physiopathologiquement linéaire : une augmentation de +1 unité pour la variable CHOLES_3CL se traduit *a priori* par une même augmentation de la concentration en créatinine. Autrement dit, et de façon plus médicale, cette hypothèse est celle d'une même augmentation de concentration en créatinine entre une hypocholestérolémie (CHOLES_3CL = 0) et une normocholestérolémie (CHOLES_3CL = 1), et entre une normocholestérolémie (CHOLES_3CL = 1) et une hypercholestérolémie (CHOLES_3CL = 2). Je vais désormais vérifier cette hypothèse de linéarité de l'association, en faisant tourner le modèle suivant dans SAS® :

$$\overline{CREAT}_{/(CHOLES\_3CL)"0"} = \alpha + \beta_1.CHOLES\_3CL(1) + \beta_2.CHOLES\_3CL(2)$$

Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus.

```
PROC GLM DATA = Donnees_pour_guide;
CLASS CHOLES_3CL (REF = '0');
MODEL CREAT = CHOLES_3CL / SOLUTION CLPARM;
RUN;
```

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 46.

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	10.86923077	B	0.42149565	25.79	<.0001	10.03234127	11.70612027
CHOLES_3CL 1	-0.05698587	B	0.52146563	-0.11	0.9132	-1.09236811	0.97839637
CHOLES_3CL 2	-0.16468531	B	0.62259034	-0.26	0.7920	-1.40085303	1.07148240
CHOLES_3CL 0	0.00000000	B	.	.	.	.	.

(a) (c) (b)

Figure 46

Le modèle estimé par SAS® s'écrit donc de la façon suivante (à partir de la valeur des coefficients pointés par la flèche (a) sur la Figure 46) :

$$\overline{CREAT}_{/(CHOLES\_3CL)"0"} = 10{,}87 - 0{,}06.CHOLES\_3CL(1) - 0{,}16.CHOLES\_3CL(2)$$

Pour vérifier la linéarité de l'association entre la concentration en créatinine et la variable CHOLES_3CL, il faut dresser le graphique de la Figure 44 en plaçant trois points (car la variable CHOLES_3CL a trois classes) : le point d'ordonnée 0, le point d'ordonnée β_1, et le point d'ordonnée β_2 (les valeurs de β_1 et β_2 sont pointées par la flèche (a) sur la Figure 46). Les abscisses respectives sont les valeurs des trois classes de la variable CHOLES_3CL, à savoir 0, 1, et 2. De plus, comme indiqué dans la partie théorique ci-dessus, je vous recommande de placer les $IC_{95\%}$ des coefficients β_1 et β_2 (cf. Figure 46.b). Pour ma part, j'utilise un fichier Excel® pour dresser le graphique souhaité (que je peux vous envoyer par email) en utilisant la SE des coefficients (pointée par la flèche (c) sur la Figure 46), cf. Figure 47.

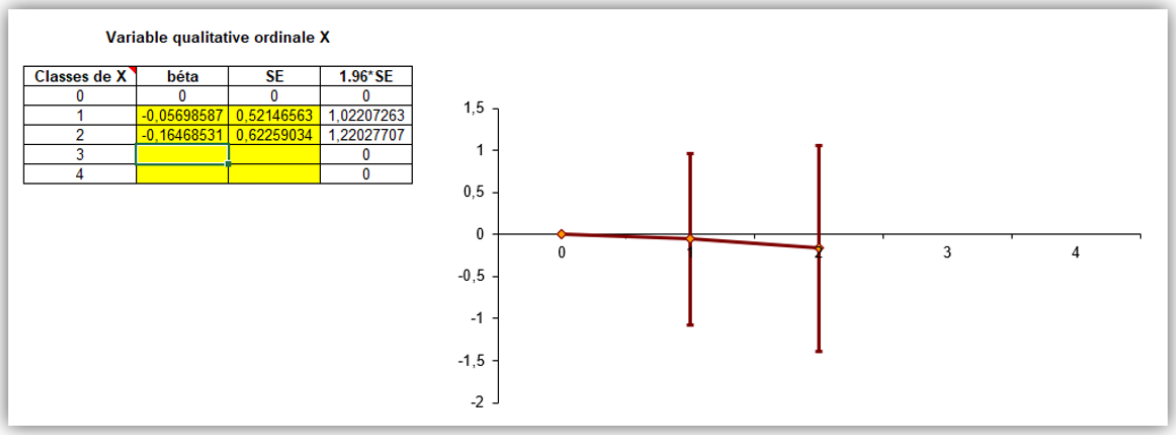


Figure 47

Sur la Figure 47, on peut voir que les trois points sont alignés. Ainsi, on peut considérer à partir des données de l'échantillon, que l'association entre la concentration en créatinine et la variable CHOLES_3CL semble linéaire (ou bien qu'il n'y a pas du tout d'association ces deux variables). Ainsi, sous cette hypothèse, il serait possible de faire tourner le modèle $\overline{CREAT}_{/CHOLES\_3CL} = \alpha + \beta.CHOLES\_3CL$, en incluant donc telle quelle la variable CHOLES_3CL. Les résultats ont déjà été présentés (cf. Figure 39) et interprétés (sous la Figure 39). Et le test statistique du coefficient β (dont le degré de signification vaut p = 0,79 ; cf. Figure 39.c) teste l'association entre la concentration en créatinine et la cholestérolémie (en trois classes). Lorsque l'hypothèse de la linéarité de l'association avec une variable qualitative ordinale (ou quantitative) est acceptée, le test statistique du coefficient de cette variable (incluse telle quelle dans le modèle) teste la relation « dose-effet », dans sa partie linéaire, entre cette variable et le CdJ.

Si en revanche les coefficients du modèle et leur SE respective avaient été tels que le graphique eût été celui de la Figure 48, et s'il y avait des raisons de penser que l'association entre la concentration en créatinine et CHOLES_3CL n'était pas linéaire, physiopathologiquement parlant, alors l'hypothèse de la linéarité de l'association n'aurait pas pu être considérée comme vérifiée.

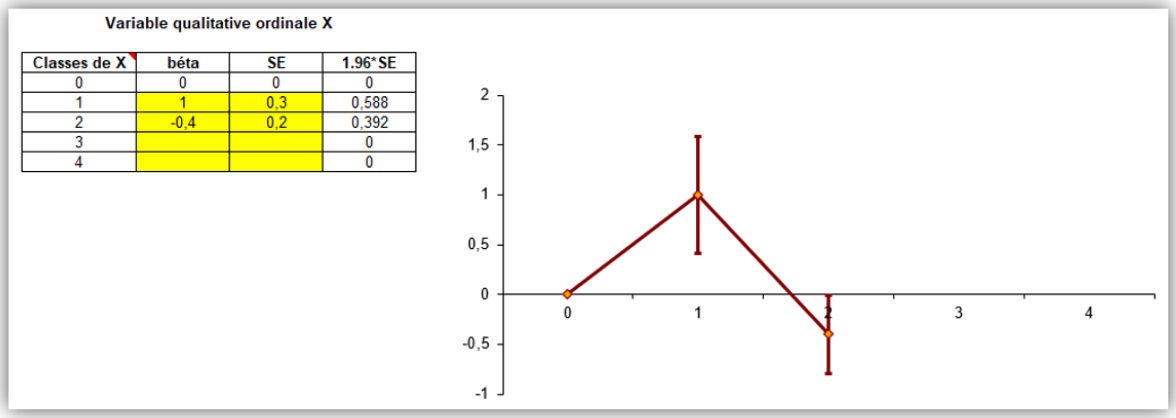


Figure 48

Dans une telle situation, une première solution consiste à laisser la variable CHOLES_3CL sous forme de variables indicatrices dans le modèle, et maintenir les lignes de programme qui avaient permis d'obtenir les résultats présentés sur la Figure 46 ci-dessus. Une deuxième solution consiste à rendre binaire la variable CHOLES_3CL en regroupant de façon pertinente deux de ses trois classes. Par exemple, on aurait pu regrouper les classes « hypocholestérolémie » et « normocholestérolémie ». La variable HYPER_CHOLES du fichier de données correspond à ce regroupement. Les lignes de programme auraient été alors celles-ci-dessous.

```
PROC GLM DATA = Donnees_pour_guide;
MODEL CREAT = HYPER_CHOLES / SOLUTION CLPARM;
RUN;
```

3. Cas d'une variable quantitative

a) Aspect théorique

Pour vérifier l'hypothèse de la linéarité de l'association avec une variable quantitative, deux étapes sont nécessaires : (1) créer une variable qualitative ordinale à partir de la variable quantitative, et (2) vérifier l'hypothèse de la linéarité de l'association entre le CdJ et cette variable qualitative ordinale créée (ce que l'on vient de faire ci-dessus).

Si l'hypothèse de la linéarité de l'association avec la variable qualitative ordinale créée pour l'occasion est vérifiée, alors on fera l'hypothèse que la linéarité de l'association est *aussi* vérifiée pour la variable quantitative en question. La linéarité de l'association avec une variable quantitative peut être montrée directement sur la variable quantitative, sans passer par la création de la variable qualitative ordinale, mais cela demande plus de travail. Ce travail peut être réalisé sous SAS® à l'aide d'une macro SAS® (Desquilbet and Mariotti, 2010).

Si l'hypothèse de la linéarité de l'association avec la variable qualitative ordinale créée pour l'occasion n'est pas vérifiée, alors on fera l'hypothèse que cette linéarité de l'association n'est pas non plus vérifiée pour la variable quantitative. Dans ce cas-là, il faudra inclure dans le modèle la variable quantitative sous forme de variables indicatrices, ou bien après l'avoir rendue binaire.

Pour vérifier l'hypothèse de la linéarité de l'association avec une variable quantitative, je vous recommande de créer la variable qualitative ordinale à partir des quartiles de la variable quantitative :

Valeur de la variable quantitative	Codage de la variable qualitative ordinale à créer
≤ 1er quartile	0
]1er quartile ; médiane]	1
]Médiane ; 3ème quartile]	2
> 3ème quartile	3

Ensuite, pour vérifier que les coefficients β_i de chacune des variables indicatrices issues de la variable qualitative ordinale sont alignés, je vous recommande fortement d'attribuer la valeur de la médiane de la variable quantitative pour chaque classe de la variable qualitative ordinale créée pour l'occasion, sur l'axe des abscisses du graphique de vérification de la linéarité de l'association.

b) *En pratique avec SAS®*

Supposons que l'on veuille étudier l'association entre la concentration en créatinine et la concentration en urée (variables quantitatives CREAT et UREE). Je vous rappelle que le modèle $\overline{CREAT}_{/UREE} = \alpha + \beta.UREE$ ne fournit une valeur de β interprétable que si l'association entre les variables UREE et CREAT est linéaire. Je vais vérifier cela ci-dessous. Pour cela, il faut créer une variable qualitative en quatre classes selon les quartiles de la variable UREE. Il se trouve que le fichier de données contient déjà cette variable, qui se nomme UREE_4CL. Il s'agit donc de vérifier la linéarité de l'association entre la variable CREAT et la variable UREE_4CL. Je vais donc faire tourner dans SAS® le modèle ci-dessous :

$$\overline{CREAT}_{/(UREE\_4CL)_{"0"}} = \alpha + \beta_1.UREE\_4CL(1) + \beta_2.UREE\_4CL(2) + \beta_3.UREE\_4CL(3)$$

Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus.

```
PROC GLM DATA = Donnees_pour_guide;
CLASS UREE_4CL (REF = '0');
MODEL CREAT = UREE_4CL / SOLUTION CLPARM;
RUN;
```

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 49.

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	10.16363636	B	0.43266304	23.49	<.0001	9.30445331	11.02281942
UREE_4CL 1	0.40436364	B	0.59323768	0.68	0.4972	-0.77368882	1.58241609
UREE_4CL 2	0.25559441	B	0.58787331	0.43	0.6647	-0.91180548	1.42299430
UREE_4CL 3	1.88636364	B	0.59899486	3.15	0.0022	0.69687856	3.07584872
UREE_4CL 0	0.00000000	B	.	.	.	.	.

(a) (b)

Figure 49

Le modèle estimé s'écrit par conséquent (à partir des valeurs des coefficients pointées par la flèche (a) sur la Figure 49) :

$$\overline{CREAT}_{/(UREE\_4CL)_{"0"}} = 10{,}16 + 0{,}40.UREE\_4CL(1) + 0{,}26.UREE\_4CL(2) + 1{,}89.UREE\_4CL(3)$$

Pour vérifier la linéarité de l'association entre la concentration en créatinine et la variable UREE_4CL, il faut dresser le graphique de la Figure 44 en plaçant quatre points (car UREE_4CL a quatre classes) : le point d'ordonnée 0, le point d'ordonnée β_1, le point d'ordonnée β_2, et le point d'ordonnée β_3. Dans cette situation où la variable qualitative ordinale a été créée à partir de variable quantitative dont on cherche à vérifier la linéarité de l'association, je vous recommande de placer ces quatre points avec comme abscisse la médiane de la variable quantitative (ici, la médiane de variable UREE) pour chacune des quatre classes de la variable UREE_4CL. Pour cela, il suffit de faire tourner les lignes de programme ci-dessous, qui font référence à ce que l'on a vu dans la sous-partie « Description selon les classes d'une variable binaire ou qualitative », page 18.

```
PROC MEANS DATA = Donnees_pour_guide MEDIAN;
CLASS UREE_4CL;
VAR UREE;
RUN;
```

Le résultat des lignes de programme ci-dessus est présenté sur la Figure 50.

The MEANS Procedure

Analysis Variable : UREE UREE		
UREE_4CL	N Obs	Median
0	23	0.2000000
1	25	0.2600000
2	26	0.2900000
3	25	0.3600000

Figure 50

Le centre des classes de la variable UREE_4CL que l'on va donc utiliser pour dresser le graphique de la Figure 44 sont donc de 0,20, 0,26, 0,29, et 0,36 respectivement pour les 1ère, 2ème, 3ème, et 4ème classes de la variable UREE_4CL. Le graphique théorique de la Figure 44 devient en pratique celui de la Figure 51 (en utilisant les valeurs des SE des coefficients pointées par la flèche (b) sur la Figure 49).

Variable quantitative X

Valeurs de X	béta	SE	1.96*SE
0,2	0	0	0
0,26	0,40436364	0,59323768	1,16274585
0,29	0,25559441	0,58787331	1,15223169
0,36	1,88636364	0,59899486	1,17402993
			0

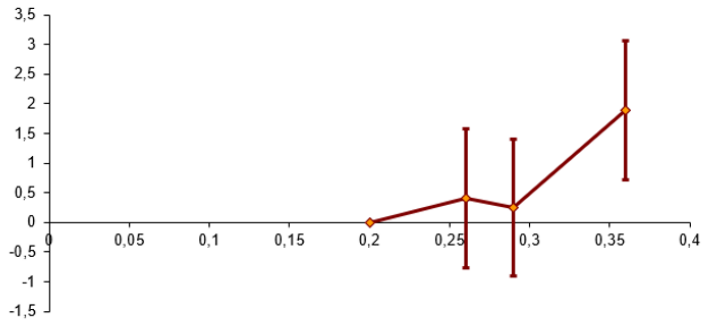


Figure 51

Si l'on n'a aucune raison de penser que l'association entre la concentration en urée et celle en créatinine n'est pas linéaire, alors la Figure 51 laisserait accepter l'hypothèse de la linéarité de l'association, et le modèle $\overline{CREAT}_{/UREE} = \alpha + \beta.UREE$ fournirait une valeur de β interprétable (cf. sous-partie « Modèle de régression linéaire univarié avec une variable quantitative », page 45). Dans le cas contraire, alors vous ne devez pas faire tourner le modèle $\overline{CREAT}_{/UREE} = \alpha + \beta.UREE$, et vous devez inclure la variable UREE_4CL sous forme de variables indicatrices, ou bien inclure une variable binaire créée à partir de la variable UREE (selon un seuil clinique pertinent, la médiane, ou l'un des quartiles).

D. La régression logistique

1. Introduction

Je vous rappelle qu'une régression logistique s'utilise lorsque le CdJ est binaire et non assorti d'un temps de survenue (par exemple, dans une étude cas-témoins ou transversale). Je vous recommande fortement de coder votre CdJ de la façon suivante : « 0 » pour les individus n'ayant pas présenté le CdJ, et « 1 » ceux qui l'ont présenté. Toutes les lignes de programme qui vont être présentées dans cette partie sur la régression logistique reposent sur le fait que le CdJ est codé selon la recommandation ci-dessus.

Dans la suite de ce guide, toutes les interprétations des coefficients issus d'un modèle de régression logistique vont s'appuyer sur celles des coefficients issus du modèle de régression linéaire vus précédemment, avec comme seule différence la suivante : tandis que β représentait la différence moyenne des valeurs du CdJ entre deux groupes d'animaux différant de +1 unité pour leur variable E, quelles que soient les valeurs de leur variable E, dans une régression logistique, β représente $Ln(OR_E)$ où l'OR_E quantifie l'association entre E et la présence du CdJ binaire en comparant des individus différant de +1 unité sur E, quelles que soient les valeurs de leur variable E. Par conséquent, vous ne pouvez pas lire la suite de ce guide si vous n'avez pas lu *tout* ce qui précède sur la régression linéaire !

Pour l'ensemble des exemples ci-dessous, je vais utiliser comme variable relative au CdJ la variable binaire DECES_3_ANS, valant « 0 » si le chien était toujours en vie 3 ans après J0, et « 1 » s'il était décédé dans les 3 ans après J0 (je rappelle que tous les chiens avaient été suivis au moins 3 ans, sauf s'ils décédaient avant, et il n'y avait aucun perdu de vue dans l'étude).

2. Interprétation des résultats d'une régression logistique univariée

a) *Modèle de régression logistique avec une variable binaire*

Supposons que l'on veuille savoir s'il existe une association entre la présence d'un décès dans les 3 ans et le sexe des chiens (variable binaire FEMELLE). Pour répondre à la question, je vais faire tourner un modèle de régression logistique, qui s'écrit de la façon suivante :

$Logit(\bar{P}_{/FEMELLE}) = \alpha + \beta.FEMELLE$, avec $\bar{P}_{/FEMELLE}$ l'espérance de la probabilité d'être décédé dans les 3 ans selon que le chien est un mâle ou une femelle.

Les lignes de programme ci-dessous permettent de faire tourner dans SAS® le modèle ci-dessus.

```
PROC LOGISTIC DATA = Donnees_pour_guide;
CLASS DECES_3_ANS (DESC);
MODEL DECES_3_ANS = FEMELLE;
RUN;
```

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 52.

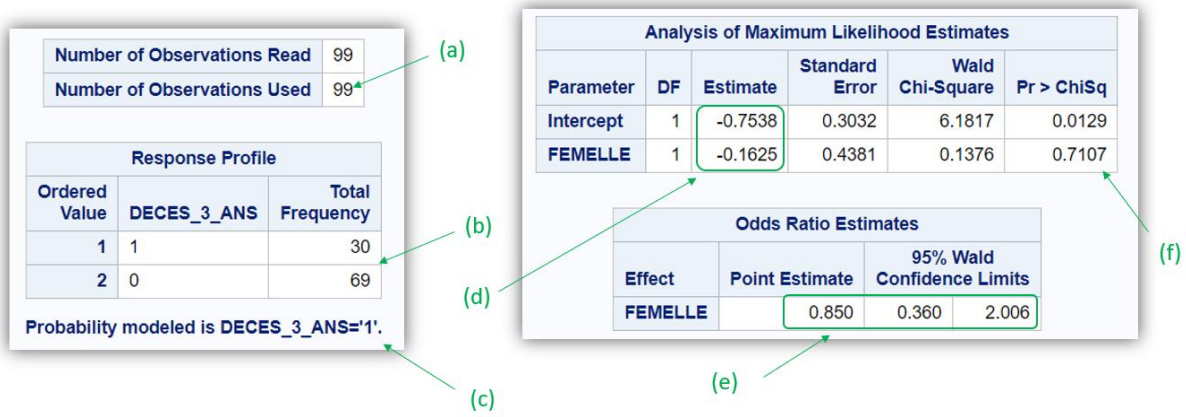


Figure 52

La flèche (a) sur la Figure 52 indique que le modèle a tourné sur 99 chiens, dont 30 sont décédés dans les 3 ans et dont 60 ne sont pas décédés dans les 3 ans (cf. Figure 52.b). Sur la 2ème ligne de programme ci-dessus, l'instruction « (DESC) » est indispensable pour que SAS® modélise l'espérance de la probabilité de présenter le CdJ en fonction des valeurs des variables incluses dans le modèle (sans « (DESC) » sur la 2ème ligne de programme, SAS® modélise la probabilité de *ne pas* présenter le CdJ, ce qui change complètement les résultats). Ainsi, c'est bien la valeur « 1 » qui doit apparaître à l'endroit que pointe la flèche (c) sur la Figure 52. La colonne « Estimate » (cf. Figure 52.d) fait référence aux

coefficients du modèle de régression logistique, comme dans la régression linéaire. Ainsi, le modèle s'écrit :

$$Logit\left(\overline{P(DECES\_3\_ANS = 1)}_{/FEMELLE}\right) = Logit\left(\bar{P}_{/FEMELLE}\right) = -0{,}75 - 0{,}16.FEMELLE.$$

Comme indiqué ci-dessus dans l'introduction, le coefficient β d'une régression logistique quantifie l'association entre la présence du CdJ (binaire) et E en tant que valeur du $Ln(OR_E)$ où OR_E est l'Odds Ratio quantifiant l'association entre le CdJ et la variable E, pour une augmentation de +1 unité de E, quelles que soient les valeurs de leur variable E. Ici, la variable FEMELLE est binaire, et elle est codée en 0/1, avec « 1 » pour les chiens femelles et « 0 » pour les chiens mâles. Puisque $Ln(OR_{FEMELLE}) = 0{,}16$, alors l'$OR_{\text{Femelles } versus \text{ mâles}} = e^{-0{,}16} = 0{,}85$, valeur que l'on retrouve dans la sortie SAS® (cf. Figure 52.e). Dans la mesure où cet OR est < 1, on peut donc dire que, dans l'échantillon, le décès dans les 3 ans était survenu *moins* fréquemment parmi les chiens femelles que parmi les chiens mâles. L'$IC_{95\%}$ de cet OR est $[0{,}36\ ;\ 2{,}01]_{95\%}$ (cf. Figure 52.e). Cet $IC_{95\%}$ comprend la valeur « 1 », donc il n'est pas significativement différent de 1, ce que l'on retrouve avec le degré de signification de β, de valeur 0,71 (cf. Figure 52.f), supérieur à 0,05. Ainsi, dans l'échantillon, il n'existait pas d'association significative entre le fait de décéder dans les 3 ans et le sexe des chiens.

b) *Modèle de régression logistique univarié avec une variable quantitative*

Supposons que l'on veuille savoir s'il existe une association entre la présence d'un décès dans les 3 ans et l'âge des chiens (variable quantitative AGE, exprimée en années), à l'aide de la régression logistique. Pour répondre à la question, je vais faire tourner un modèle de régression logistique, qui s'écrit de la façon suivante :

$$Logit\left(\bar{P}_{/AGE}\right) = \alpha + \beta.AGE$$

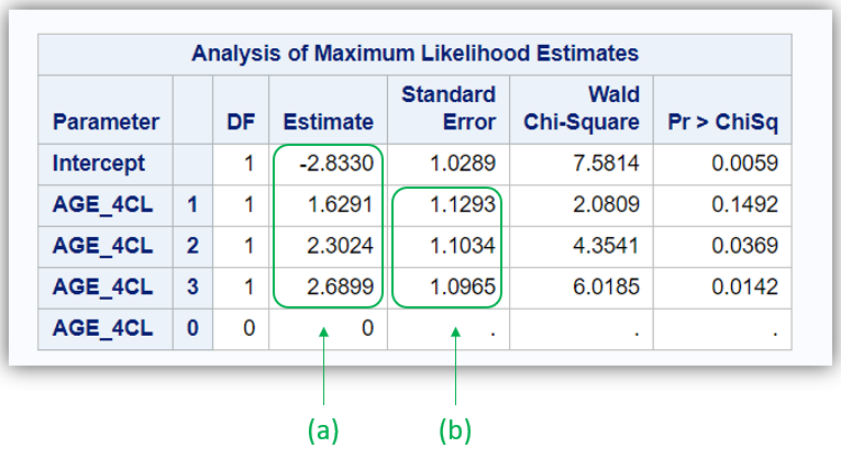
Je vous rappelle (vivement) que faire tourner le modèle ci-dessus nécessite d'avoir vérifié que l'association entre l'âge et la présence d'un décès à 3 ans est linéaire. Vous devez vous souvenir que pour vérifier la linéarité de l'association avec une variable quantitative (comme ici, la variable AGE), il faut créer une variable qualitative ordinale à partir de la variable quantitative. Il se trouve que pour la variable AGE, cette variable qualitative ordinale est déjà présente dans le fichier de données : c'est la variable AGE_4CL, dont les classes correspondent aux quartiles de la variable AGE (comme je vous recommande de le faire pour vérifier la linéarité d'une association). Je vais donc faire tourner le modèle ci-dessous :

$$Logit\left(\bar{P}_{/(AGE\_4CL)_{"0"}}\right) = \alpha + \beta_1.AGE\_4CL(1) + \beta_2.AGE\_4CL(2) + \beta_3.AGE\_4CL(3)$$

Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus.

```
PROC LOGISTIC DATA = Donnees_pour_guide;
CLASS DECES_3_ANS (DESC);
CLASS AGE_4CL (REF = '0') / PARAM = GLM;
MODEL DECES_3_ANS = AGE_4CL;
RUN;
```

Notez que dans les lignes de programme ci-dessus, il y a deux instructions « CLASS », une pour le CdJ, et une pour la variable qualitative que l'on inclut sous forme de variables indicatrices. Notez aussi qu'il est indispensable de taper l'instruction « / PARAM = GLM » sur la 3ème ligne de programme. Je vous conseille de taper autant d'instructions « CLASS » (contenant entre autres « / PARAM = GLM » sur la ligne) que de variables qualitatives à inclure sous forme de variables indicatrices. Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 53.



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.8330	1.0289	7.5814	0.0059
AGE_4CL	1	1	1.6291	1.1293	2.0809	0.1492
AGE_4CL	2	1	2.3024	1.1034	4.3541	0.0369
AGE_4CL	3	1	2.6899	1.0965	6.0185	0.0142
AGE_4CL	0	0	0	.	.	.

Figure 53

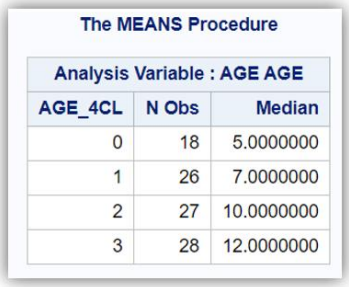
Le modèle estimé par SAS® s'écrit donc de la façon suivante (à partir de la valeur des coefficients pointés par la flèche (a) sur la Figure 53) :

$$Logit\left(\bar{P}_{/(AGE\_4CL)"0"}\right) = -2{,}83 + 1{,}63.AGE\_4CL(1) + 2{,}30.AGE\_4CL(2) + 2{,}69.AGE\_4CL(3)$$

Pour vérifier la linéarité de l'association entre la présence d'un décès à 3 ans et la variable AGE_4CL, il faut dresser le graphique de la Figure 44 en plaçant quatre points (car AGE_4CL a quatre classes) : le point d'ordonnée 0, le point d'ordonnée β_1, le point d'ordonnée β_2, et le point d'ordonnée β_3. Comme indiqué dans la sous-partie « Cas d'une variable quantitative - Aspect théorique », page 60), je vous recommande de placer ces quatre points avec comme abscisse la médiane de la variable quantitative (ici, la médiane de la variable AGE) pour chacune des quatre classes de la variable AGE_4CL. Pour cela, je vais utiliser la procédure PROC MEANS, dont les lignes de programme sont ci-dessous.

```
PROC MEANS DATA = Donnees_pour_guide MEDIAN;
CLASS AGE_4CL;
VAR AGE;
RUN;
```

Le résultat des lignes de programme ci-dessus est présenté sur la Figure 54.



The MEANS Procedure

Analysis Variable : AGE AGE

AGE_4CL	N Obs	Median
0	18	5.0000000
1	26	7.0000000
2	27	10.0000000
3	28	12.0000000

Figure 54

Le centre des classes de la variable AGE_4CL que je vais donc utiliser pour dresser le graphique de la Figure 44 sont donc de 5, 7, 10, et 12 ans respectivement pour les 1ère, 2ème, 3ème, et 4ème classes de la variable AGE_4CL. Le graphique théorique de la Figure 44 devient en pratique celui de la Figure 55, en utilisant les valeurs des SE pointées par la flèche (b) sur la Figure 53, et le fichier Excel® précédemment mentionné.

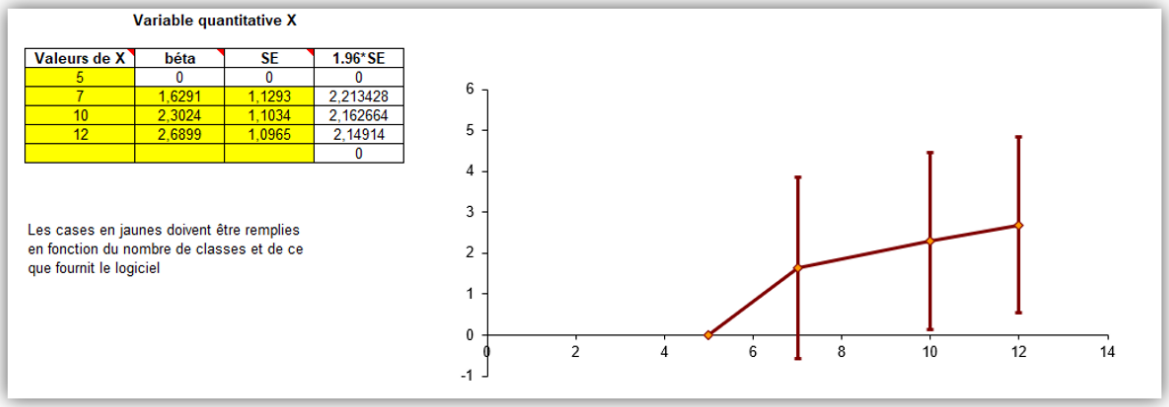


Figure 55

La Figure 55 nous permet d'accepter la linéarité de l'association entre la présence d'un décès à 3 ans et la variable AGE_4CL. Ainsi, on peut aussi accepter la linéarité de l'association entre la présence d'un décès à 3 ans et la variable quantitative AGE. Ainsi, je peux faire tourner le modèle ci-dessous, et interpréter la valeur du coefficient β.

$$Logit\left(\bar{P}_{/AGE}\right) = \alpha + \beta . AGE$$

Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus.

```
PROC LOGISTIC DATA = Donnees_pour_guide;
CLASS DECES_3_ANS (DESC);
MODEL DECES_3_ANS = AGE;
RUN;
```

Les résultats (extrait) des lignes de programme ci-dessus sont présentés sur la Figure 56.

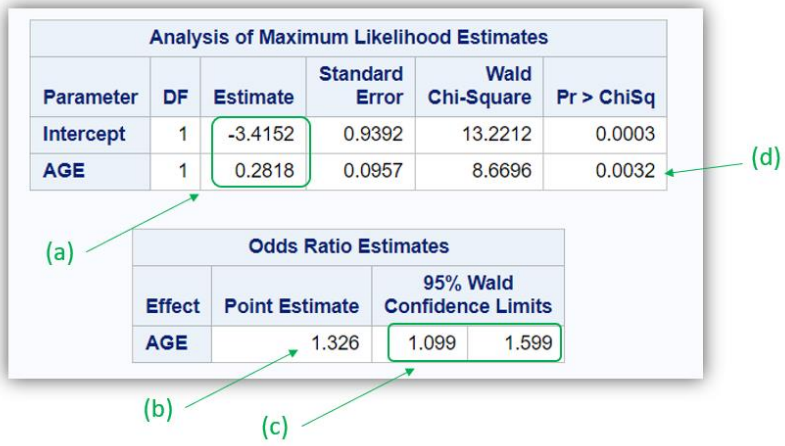


Figure 56

Le modèle estimé par SAS® s'écrit donc de la façon suivante (à partir de la valeur des coefficients pointés par la flèche (a) sur la Figure 56) :

$$Logit\left(\bar{P}_{/AGE}\right) = -3{,}41 + 0{,}28 . AGE$$

Ainsi, l'OR_{AGE} = $e^{0,28}$ = 1,33 (cf. Figure 56.b), avec comme $IC_{95\%}$: $[1{,}10\ ;\ 1{,}60]_{95\%}$ (cf. Figure 56.c), significativement différent de 1 (cf. Figure 56.d). Vous avez vu précédemment que le coefficient β quantifie l'association entre la présence du CdJ (binaire) et E en tant que valeur du $Ln(OR_E)$, où OR_E est l'Odds Ratio quantifiant l'association entre le CdJ et la variable E, pour une augmentation de +1 unité de E, quelles que soient les valeurs de leur variable E. Ainsi, l'OR_{AGE} de valeur 1,33 s'interprète de la façon suivante : une augmentation de +1 année d'âge pour les chiens de l'échantillon, quel que soit

l'âge des chiens, se traduisait par un OR [$IC_{95\%}$] de présenter un décès dans les 3 ans de 1,33 [1,10 ; 1,60]$_{95\%}$. Cet OR étant significativement différent de 1, il existait une association significative entre l'âge des chiens et le fait de présenter un décès dans les 3 ans. Ainsi, voici ce que l'on écrirait dans un article : « il existait une association significative entre l'âge et le fait de présenter un décès dans les 3 ans (OR [$IC_{95\%}$] pour une augmentation de +1 année d'âge de 1,33 [1,10 ; 1,60], $p < 0,01$) ».

Puisque l'hypothèse de la linéarité de l'association avec l'âge est acceptée, le test statistique du coefficient β de la variable AGE (dont le degré de signification est inférieur à 0,01 ; cf. Figure 56.d) teste la relation « dose-effet », dans sa partie linéaire, entre la présence d'un décès dans les 3 ans et l'âge. Ainsi, la fréquence d'un décès dans les 3 ans était significativement d'autant plus importante que l'âge des chiens augmentait.

c) *Modèle de régression logistique univarié avec une variable qualitative ordinale*

Supposons que l'on veuille savoir s'il existe une association entre la présence d'un décès dans les 3 ans et la cholestérolémie des chiens (variable qualitative ordinale CHOLES_3CL en trois classes : hypocholestérolémie, normocholestérolémie, et hypercholestérolémie), à l'aide de la régression logistique. Le modèle ci-dessous pourrait permettre de réaliser l'analyse statistique souhaitée :

$$Logit\left(\bar{P}_{/CHOLES\_3CL}\right) = \alpha + \beta.CHOLES\_3CL$$

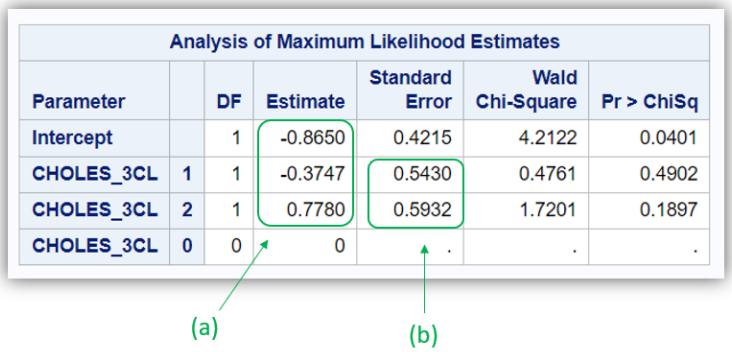
Mais je vous rappelle (encore et toujours) que faire tourner le modèle ci-dessus nécessite d'avoir vérifié que l'association entre la cholestérolémie (en trois classes) et la présence d'un décès à 3 ans est linéaire. C'est ce que je vais désormais vérifier, en faisant tourner le modèle ci-dessous :

$$Logit\left(\bar{P}_{/(CHOLES\_3CL)"0"}\right) = \alpha + \beta_1.CHOLES\_3CL(1) + \beta_2.CHOLES\_3CL(2)$$

Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus.

```
PROC LOGISTIC DATA = Donnees_pour_guide;
CLASS DECES_3_ANS (DESC);
CLASS CHOLES_3CL (REF = '0') / PARAM = GLM;
MODEL DECES_3_ANS = CHOLES_3CL;
RUN;
```

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 57.



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.8650	0.4215	4.2122	0.0401
CHOLES_3CL	1	1	-0.3747	0.5430	0.4761	0.4902
CHOLES_3CL	2	1	0.7780	0.5932	1.7201	0.1897
CHOLES_3CL	0	0	0	.	.	.

Figure 57

Le modèle estimé par SAS® s'écrit donc de la façon suivante (à partir de la valeur des coefficients pointés par la flèche (a) sur la Figure 57) :

$$Logit\left(\bar{P}_{/(CHOLES\_3CL)"0"}\right) = -0,87 - 0,37.CHOLES\_3CL(1) + 0,78.CHOLES\_3CL(2)$$

Pour vérifier la linéarité de l'association entre la présence d'un décès à 3 ans et la variable CHOLES_3CL, il faut dresser le graphique de la Figure 44 en plaçant trois points (car CHOLES_3CL a trois classes) : le point d'ordonnée 0, le point d'ordonnée β_1, et le point d'ordonnée β_2 (cf. valeurs pointées par la flèche (a) sur la Figure 57) et comme abscisse les valeurs de 0, 1, et 2 qui sont les valeurs des trois classes de

CHOLES_3CL. A partir des valeurs des SE des coefficients β_i (cf. Figure 57.b), et grâce au fichier Excel®, on dresse le graphique présenté sur la Figure 58.

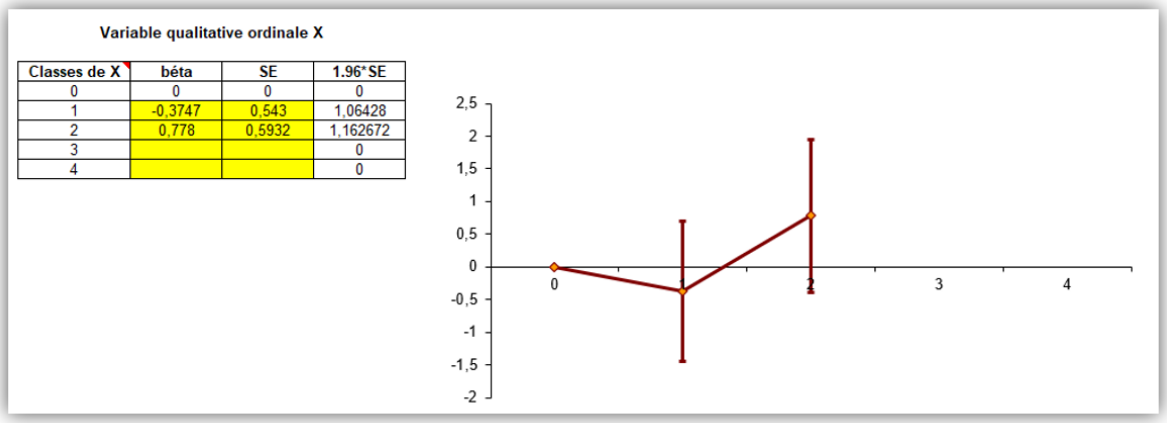


Figure 58

A partir d'un tel graphique, tout dépend des hypothèses *a priori* de la forme de l'association entre la cholestérolémie et le décès à 3 ans. S'il n'y a pas de fortes raisons de penser que, dans la population, l'association n'est pas linéaire, alors le graphique de la Figure 58 permet d'accepter l'hypothèse de la linéarité de l'association (les $IC_{95\%}$ sont suffisamment larges pour le faire). Le modèle correspondant serait alors celui-ci-dessous.

$$Logit\left(\bar{P}_{/CHOLES\_3CL}\right) = \alpha + \beta.CHOLES\_3CL$$

Dans le modèle ci-dessus, la variable CHOLES_3CL est donc incluse telle quelle dans le modèle, c'est-à-dire sans qu'elle ne soit incluse sous forme de variables indicatrices. Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus.

```
PROC LOGISTIC DATA = Donnees_pour_guide;
CLASS DECES_3_ANS (DESC);
MODEL DECES_3_ANS = CHOLES_3CL;
RUN;
```

Les résultats (extrait) des lignes de programme ci-dessus sont présentés sur la Figure 59.

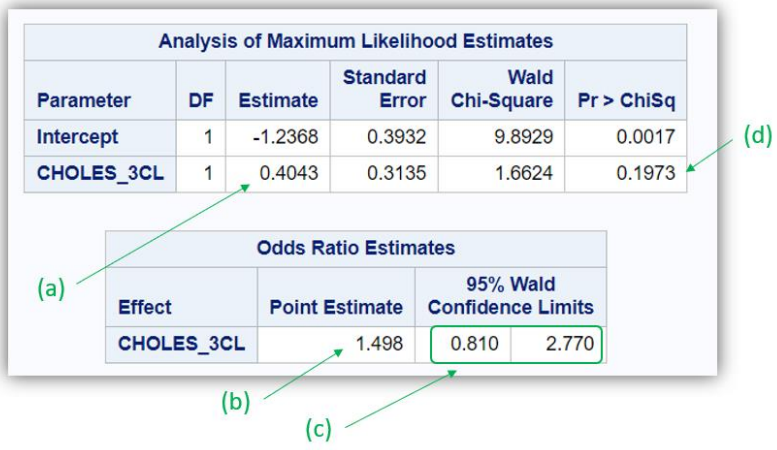


Figure 59

La valeur du coefficient β est égale à 0,40 (cf. Figure 59.a), ce qui conduit à un OR de valeur de 1,50 (cf. Figure 59.b), avec un $IC_{95\%}$ de $[0,81 ; 2,77]_{95\%}$ (cf. Figure 59.c). Cet OR n'est pas significativement différent de 1 (p = 0,20 ; cf. Figure 59.d). Dans la mesure où la linéarité de l'association entre la présence d'un décès à 3 ans et la cholestérolémie a été acceptée, l'OR de valeur 1,50 est interprétable, et son interprétation est la suivante : une augmentation d'une classe de la variable relative à la

cholestérolémie (c'est-à-dire, le passage d'une hypocholestérolémie à une normocholestérolémie, ou bien le passage d'une normocholestérolémie à une hypercholestérolémie) pour les chiens de l'échantillon se traduisait par un OR [$IC_{95\%}$] de présenter un décès dans les 3 ans de 1,50 [0,81 ; 2,77]$_{95\%}$. Vous vous rendez compte que l'interprétation d'un OR pour une variable qualitative ordinale dont la linéarité de l'association est vérifiée est quand même un peu ardue… !

Notez que ce n'est pas parce que l'hypothèse de la linéarité d'une association avec une variable qualitative ordinale est vérifiée qu'il est obligatoire de l'inclure telle quelle dans le modèle ! Si l'interprétation telle que celle ci-dessus est un peu trop « ardue », alors incluez la variable qualitative ordinale sous forme de variables indicatrices d'emblée, sans chercher à vérifier l'hypothèse de la linéarité de l'association avec le CdJ.

Si maintenant il y a des raisons de penser que, dans la population, l'association entre la présence d'un décès à 3 ans et la cholestérolémie n'est pas linéaire, alors le graphique de la Figure 58 semble confirmer cette hypothèse. (Vous vous rendez compte qu'il est fondamental d'avoir des idées ou des hypothèses *a priori*, et ne pas attendre que les données de l'échantillon vous *disent* ce qu'il se passe dans la population !) Dans cette situation de non acceptation de la linéarité de l'association entre la présence d'un décès à 3 ans et la cholestérolémie, alors il y a deux possibilités (comme on l'a vu pour la régression linéaire) : inclure la variable CHOLES_3CL sous forme de variables indicatrices, ou bien recoder cette variable de façon binaire (par exemple « hypocholestérolémie en oui/non », ou bien « hypercholestérolémie en oui/non »). Si l'on ne souhaite pas regrouper deux classes de cholestérolémie parce que cela n'est pas cliniquement pertinent, alors incluons cette variable sous forme de variables indicatrices. Mais pour cette variable CHOLES_3CL, il semble que la classe de référence la plus pertinente d'un point de vue clinique serait la classe correspondant à une normocholestérolémie (CHOLES_3CL = 1). Je vais donc faire tourner le modèle ci-dessous :

$$Logit\left(\bar{P}_{/(CHOLES\_3CL)"1"}\right) = \alpha + \beta_0.CHOLES\_3CL(0) + \beta_2.CHOLES\_3CL(2)$$

Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus (« REF = '1' » à la 3ème ligne de programme ci-dessous).

```
PROC LOGISTIC DATA = Donnees_pour_guide;
CLASS DECES_3_ANS (DESC);
CLASS CHOLES_3CL (REF = '1') / PARAM = GLM;
MODEL DECES_3_ANS = CHOLES_3CL;
RUN;
```

Les résultats (extrait) des lignes de programme ci-dessus sont présentés sur la Figure 60.

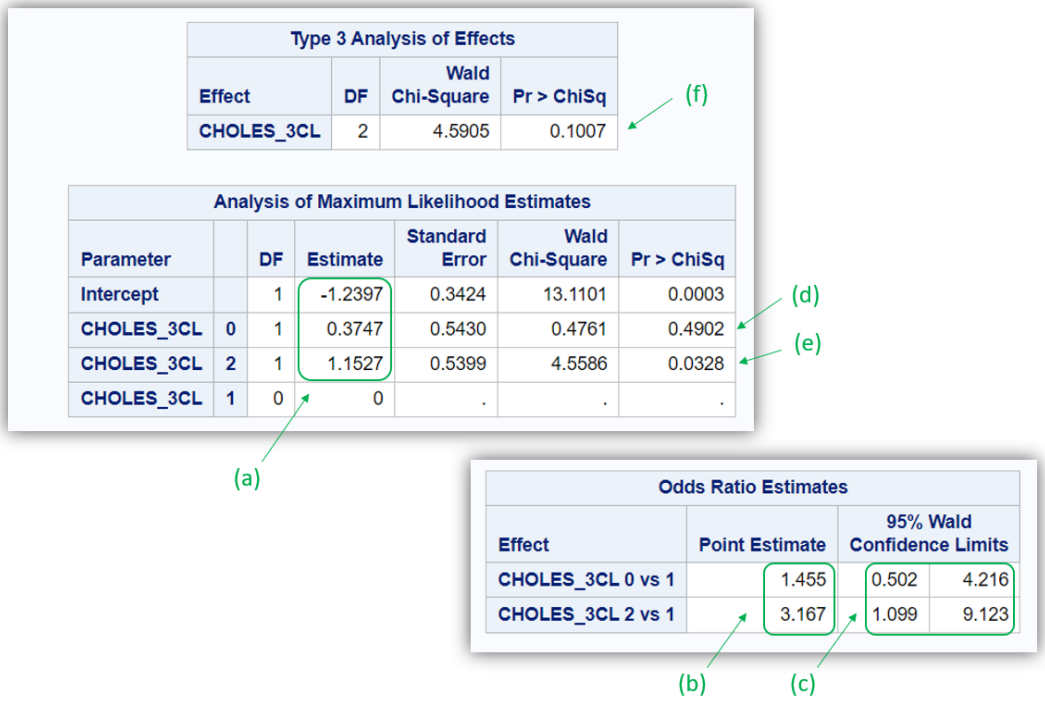


Figure 60

Le modèle estimé par SAS® s'écrit donc de la façon suivante (à partir de la valeur des coefficients pointés par la flèche (a) sur la Figure 60) :

$$Logit\left(\bar{P}_{/(CHOLES\_3CL)"1"}\right) = -1{,}24 + 0{,}37.CHOLES\_3CL(0) + 1{,}15.CHOLES\_3CL(2)$$

Les coefficients β_0 et β_2, respectivement de 0,37 et 1,15, conduisent aux valeurs d'OR de 1,46 et 3,17 (cf. Figure 60.b), avec leur $IC_{95\%}$ pointé par la flèche (c) sur la Figure 60. La classe de référence étant celle correspondant à CHOLES_3CL = 1 (chiens en normocholestérolémie), les interprétations dans l'échantillon de ces OR sont les suivantes. L'$OR_{\text{hypo } versus \text{ normocholestérolémie}}$ étant supérieur à 1 (OR = 1,46), les chiens avec une hypocholestérolémie étaient décédés dans les 3 ans de façon plus fréquente que les chiens avec une normocholestérolémie. Cet OR n'était pas significativement différent de 1 (p = 0,49 ; cf. Figure 60.d). L'$OR_{\text{hyper } versus \text{ normocholestérolémie}}$ étant lui aussi supérieur à 1 (OR = 3,17), les chiens avec une hypercholestérolémie étaient décédés dans les 3 ans de façon plus fréquente que les chiens avec une normocholestérolémie. Cet OR était quant à lui significativement différent de 1 (p = 0,03 ; cf. Figure 60.e). Pour information, globalement, la variable CHOLES_3CL, sous forme de variables indicatrices, n'était pas significativement associée à la présence d'un décès dans les 3 ans (valeur du degré de signification de 0,10 dans le tableau « Type 3 Analysis of Effects » ; cf. Figure 60.f).

d) Modèle de régression logistique univarié avec une variable qualitative nominale

Supposons que l'on veuille savoir s'il existe une association entre la présence d'un décès dans les 3 ans et la race des chiens (variable qualitative nominale RACE_4CL en quatre classes), à l'aide de la régression logistique.

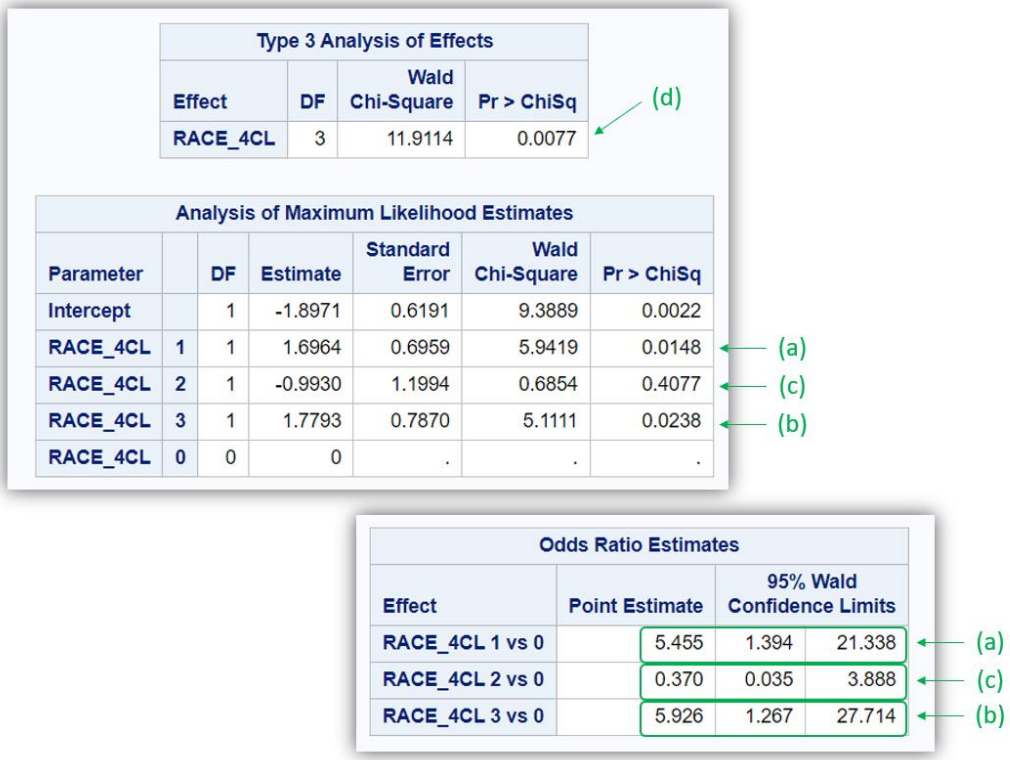
Nous allons reprendre exactement la même démarche que celle avec la régression linéaire, à l'époque où nous voulions savoir s'il existait une association entre la concentration en créatinine et la race des chiens : la variable RACE_4CL doit obligatoirement être incluse dans le modèle sous forme de variables indicatrices. Je vais de plus choisir comme classe de référence les chiens de race Golden (RACE_4CL = 0). Je vais donc faire tourner le modèle ci-dessous :

$$Logit\left(\bar{P}_{/(RACE\_4CL)"0"}\right) = \alpha + \beta_1.RACE\_4CL(1) + \beta_2.RACE\_4CL(2) + \beta_3.RACE\_4CL(3)$$

Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus.

```
PROC LOGISTIC DATA = Donnees_pour_guide;
CLASS DECES_3_ANS (DESC);
CLASS RACE_4CL (REF = '0') / PARAM = GLM;
MODEL DECES_3_ANS = RACE_4CL;
RUN;
```

Les résultats (extrait) des lignes de programme ci-dessus sont présentés sur la Figure 61.



Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
RACE_4CL	3	11.9114	0.0077

(d)

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.8971	0.6191	9.3889	0.0022
RACE_4CL	1	1	1.6964	0.6959	5.9419	0.0148
RACE_4CL	2	1	-0.9930	1.1994	0.6854	0.4077
RACE_4CL	3	1	1.7793	0.7870	5.1111	0.0238
RACE_4CL	0	0	0	.	.	.

(a) (c) (b)

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
RACE_4CL 1 vs 0	5.455	1.394	21.338
RACE_4CL 2 vs 0	0.370	0.035	3.888
RACE_4CL 3 vs 0	5.926	1.267	27.714

(a) (c) (b)

Figure 61

Ainsi, le décès à 3 ans était significativement plus fréquent parmi les chiens de race Labrador (RACE_4CL = 1) que parmi les chiens de race Golden (RACE_4CL = 0) : $OR_{\text{Labrador } versus \text{ Golden}} = 5,45\ [1,39 ; 21,34]_{95\%}$, $p = 0,01$ (cf. flèches (a) sur la Figure 61).

Le décès à 3 ans était aussi significativement plus fréquent parmi les chiens d'autre race (RACE_4CL = 3) que parmi les chiens de race Golden : $OR_{\text{Autre race } versus \text{ Golden}} = 5,29\ [1,27 ; 27,71]_{95\%}$, $p = 0,02$ (cf. flèches (b) sur la Figure 61).

Le décès à 3 ans était en revanche moins fréquent parmi les chiens de race croisée Golden/Labrador (RACE_4CL = 2) que parmi les chiens de race Golden : $OR_{\text{Race croisée } versus \text{ Golden}} = 0,37\ [0,04 ; 3,89]_{95\%}$, $p = 0,41$ (cf. flèches (c) sur la Figure 61), sans que cette différence de fréquence ne soit significative.

Notez que l'association globale entre la présence d'un décès à 3 ans et la race était significative ($p < 0,01$; Figure 61.d).

3. Interprétation des résultats d'une régression logistique multivariée

Supposons que l'on souhaite étudier l'association entre la présence d'un décès à 3 ans et le sexe des chiens, ajustée sur l'âge, la race des chiens (en prenant la race Golden comme classe de référence ; RACE_4CL = 0), et la cholestérolémie sous forme de variables indicatrices en prenant comme classe de référence les chiens avec une normocholestérolémie (CHOLES_3CL = 1). Je vais donc faire tourner le modèle de régression logistique ci-dessous :

$$Logit\left(\bar{P}_{/FEMELLE,AGE,(RACE\_4CL)_{"0"},(CHOLES\_3CL)_{"1"}}\right) = \alpha + \beta.FEMELLE + \gamma.AGE + \tau_0.CHOLES\_3CL(0) + \tau_2.CHOLES\_3CL(2) + \delta_1.RACE\_4CL(1) + \delta_2.RACE\_4CL(2) + \delta_3.RACE\_4CL(3)$$

Je rappelle (encore et encore) que *tous* les coefficients du modèle (β, γ, τ_0, τ_2, δ_1, δ_2, et δ_3), et pas seulement l'interprétation du coefficient γ, ne sont interprétables que si l'hypothèse de la linéarité de l'association entre l'âge et la présence d'un décès à 3 ans est linéaire. Cela dit, nous avons vu que, de façon brute (c'est-à-dire, non ajustée), l'association entre l'âge et la présence d'un décès à 3 ans pouvait être considérée comme linéaire (cf. Figure 55). Même s'il est toujours possible qu'un biais de confusion puisse modifier la forme de l'association entre une variable quantitative et le CdJ, on pourrait considérer que si l'association brute peut être considérée comme linéaire entre le CdJ et la variable quantitative, alors elle le restera après ajustement sur d'autres variables dans un modèle de régression multivarié. Donc, les estimations des coefficients β, γ, δ_1, δ_2, δ_3, τ_0, et τ_2 du modèle ci-dessus seront interprétables.

Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus.

```
PROC LOGISTIC DATA = Donnees_pour_guide;
CLASS DECES_3_ANS (DESC);
CLASS CHOLES_3CL (REF = '1') / PARAM = GLM;
CLASS RACE_4CL (REF = '0') / PARAM = GLM;
MODEL DECES_3_ANS = FEMELLE AGE CHOLES_3CL RACE_4CL;
RUN;
```

Les résultats (extrait) des lignes de programme ci-dessus sont présentés sur la Figure 62.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
FEMELLE	1	0.0985	0.7537
AGE	1	8.9005	0.0029
CHOLES_3CL	2	0.5978	0.7417
RACE_4CL	3	10.7559	0.0131

(d)

Number of Observations Read	99
Number of Observations Used	99

(a)

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
FEMELLE	0.849	0.304	2.367
AGE	1.418	1.127	1.783
CHOLES_3CL 0 vs 1	1.602	0.433	5.934
CHOLES_3CL 2 vs 1	1.391	0.381	5.082
RACE_4CL 1 vs 0	5.412	1.060	27.641
RACE_4CL 2 vs 0	0.195	0.016	2.435
RACE_4CL 3 vs 0	4.588	0.876	24.029

(b)

(c)

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.1035	1.3313	14.6950	0.0001
FEMELLE		1	-0.1642	0.5233	0.0985	0.7537
AGE		1	0.3489	0.1170	8.9005	0.0029
CHOLES_3CL	0	1	0.4714	0.6680	0.4981	0.4803
CHOLES_3CL	2	1	0.3302	0.6610	0.2496	0.6174
CHOLES_3CL	1	0	0	.	.	.
RACE_4CL	1	1	1.6886	0.8320	4.1192	0.0424
RACE_4CL	2	1	-1.6325	1.2869	1.6091	0.2046
RACE_4CL	3	1	1.5235	0.8448	3.2525	0.0713
RACE_4CL	0	0	0	.	.	.

Figure 62

La flèche (a) sur la Figure 62 indique que le modèle multivarié a tourné sur 99 chiens (la totalité de l'échantillon).

En reprenant les résultats précédents (cf. Figure 52.e et Figure 52.f pour la variable FEMELLE, Figure 56.b, Figure 56.c, et Figure 56.d pour la variable AGE, Figure 60.b, Figure 60.c, et Figure 60.e pour la variable CHOLES_3CL, Figure 61.a, Figure 61.b, et Figure 61.c pour la variable RACE_4CL), on peut remplir les colonnes du tableau ci-dessous sous « Crude association ». Et à partir des résultats présentés sur la Figure 62 (cf. Figure 62.b et Figure 62.c), on peut remplir les colonnes du tableau ci-dessous sous « Adjusted association ». Le tableau ci-dessous présente ainsi les OR bruts et ajustés quantifiant les association brutes et ajustées entre les variables du tableau et la présence d'un décès

dans les 3 ans, tel que l'on pourrait le présenter dans un article. Des « 1 » sans chiffre après la virgule doivent être mis dans les colonnes « OR » et « aOR » pour mettre en évidence la classe de référence choisie, et donc celle par rapport à laquelle les autres classes sont comparées dans la présence du CdJ.

	Crude association			Adjusted association		
Exposure	OR	95% CI	P-value	aOR	95% CI	P-value
Female (*versus* male)	0.85	(0.36-2.01)	0.71	0.85	(0.30-2.37)	0.75
Age (x 1-year increase)	1.33	(1.10-1.60)	< 0.01	1.42	(1.13-1.78)	< 0.01
Cholesterolemia						
Hypocholesterolemia	1.46	(0.50-4.22)	0.49	1.60	(0.43-5.93)	0.48
Normocholesterolemia	1			1		
Hypercholesterolemia	3.17	(1.10-9.12)	0.03	1.40	(0.38-5.08)	0.62
Breed						
Golden	1			1		
Labrador	5.46	(1.39-21.34)	0.01	5.41	(1.06-27.64)	0.04
Golden-Labrador	0.37	(0.04-3.89)	0.41	0.20	(0.02-2.44)	0.20
Other breed	5.93	(1.27-27.71)	0.02	4.59	(0.88-24.03)	0.07

OR, Odds Ratio; aOR, Odds Ratio adjusted for all the exposures listed in the table; CI, confidence interval.

Pour information, indépendamment du sexe, de l'âge, et de la race, la cholestérolémie n'était *globalement* pas significativement associée à la présence d'un décès à 3 ans (p = 0,74 ; cf. Figure 62.d). Mais indépendamment du sexe, de l'âge, et de la cholestérolémie, la race était quant à elle *globalement* significativement associée à la présence d'un décès à 3 ans (p = 0,01 ; cf. Figure 62.d). Il est par ailleurs normal que les degrés de signification des variables FEMELLE et AGE dans les tableaux « Type 3 Analysis of Effects » (cf. Figure 62.d) et « Analysis of Maximum Likelihood Estimates » (cf. Figure 62.c) soient identiques car comme je l'ai écrit plus haut (à la toute fin de la sous-partie « Interprétation des résultats d'une régression linéaire multivariée – En pratique avec SAS® », page 54), lorsque la variable est incluse telle quelle dans un modèle, le test statistique local de son coefficient (cf. Figure 62.c) correspond au test statistique global de ce coefficient (cf. Figure 62.d).

E. Le modèle (à risques proportionnels) de Cox

1. Introduction

Vous devez avoir acquis les connaissances de base en analyse de survie et en épidémiologie (analytique) (cf. page 8), et avoir lu et compris tout ce qui précède dans ce guide avant de poursuivre.

Je vous rappelle qu'un modèle de Cox s'utilise lorsque le CdJ est binaire et assorti d'un temps de survenue (par exemple, dans une étude de cohorte). Je vous recommande fortement de coder votre variable relative au CdJ (c'est-à-dire la variable relative au fait que l'événement étudié est, ou n'est pas, survenu au cours du temps) de la façon suivante : « 0 » pour les individus censurés, et « 1 » pour les individus ayant présenté le CdJ au cours du temps. Toutes les lignes de programme qui vont être présentées dans cette partie sur le modèle de Cox reposent sur le fait que le CdJ est codé selon la recommandation ci-dessus.

A part le fait que le modèle de Cox fournit un (ou plusieurs) HR alors que la régression logistique fournit un (ou plusieurs) OR, et à part le fait que le modèle de Cox repose sur une hypothèse qui s'appelle l'hypothèse des risques proportionnels, dont je parlerai plus loin dans ce guide, tout ce que j'ai écrit pour la régression logistique est valable pour le modèle de Cox. Ainsi, je vais passer beaucoup moins de temps sur le modèle de Cox que je n'en ai passé sur la régression logistique. Je vais par conséquent directement vous montrer comment faire tourner un modèle de Cox multivarié et interpréter les sorties SAS®.

Pour l'ensemble des exemples ci-dessous, je vais utiliser les mêmes deux variables que celles utilisées dans la partie « Analyse de survie à l'aide des courbes de Kaplan-Meier » : les variables SURVIE et DECES.

2. Interprétation des résultats d'un modèle de Cox multivarié

Supposons que l'on souhaite étudier l'association entre la survenue d'un décès et la démarche des chiens (variable binaire DEMARCHE_ANORMALE), ajustée sur l'âge (variable quantitative AGE), la cholestérolémie (variable qualitative ordinale CHOLES_3CL en trois classes : hypocholestérolémie, normocholestérolémie, et hypercholestérolémie), et la race des chiens (variable qualitative nominale RACE_4CL en quatre classes). Je vais de plus faire l'hypothèse que l'association entre l'âge est la survenue d'un décès est linéaire, mais que cette hypothèse n'est pas vérifiée pour la cholestérolémie. Je vais choisir comme classe de référence la race Golden pour la race (RACE_4CL = 0) et la classe « normocholestérolémie » pour la cholestérolémie (CHOLES_3CL = 1). Je vais donc faire tourner le modèle de Cox multivarié ci-dessous :

$$Ln\left(\overline{\lambda(t)}_{/DEMARCHE\_ANORMALE,AGE,(CHOLES\_3CL)_{"1"},(RACE\_4CL)_{"0"}}\right) = Ln\left(\lambda_0(t)\right) + \beta.DEMARCHE\_ANORMALE + \gamma.AGE + \tau_0.CHOLES\_3CL(0) + \tau_2.CHOLES\_3CL(2) + \delta_1.RACE\_4CL(1) + \delta_2.RACE\_4CL(2) + \delta_3.RACE\_4CL(3)$$

Dans la mesure où j'ai fait l'hypothèse que l'association entre la survenue d'un décès et l'âge était linéaire (il s'agit de la seule variable non binaire incluse telle quelle dans le modèle), les estimations de tous les coefficients du modèle (β, γ, τ_0, τ_2, δ_1, δ_2, et δ_3) seront interprétables.

Les lignes de programme ci-dessous permettent de faire tourner le modèle ci-dessus.

```
PROC PHREG DATA = Donnees_pour_guide;
CLASS CHOLES_3CL (REF = '1') / PARAM = GLM;
CLASS RACE_4CL (REF = '0') / PARAM = GLM;
MODEL SURVIE * DECES(0) = DEMARCHE_ANORMALE AGE CHOLES_3CL RACE_4CL / RL;
RUN;
```

Les résultats (extrait) des lignes de programme ci-dessus sont présentés sur la Figure 63 et sur la Figure 64. Comme cela été déjà le cas dans la procédure PROC LIFETEST, le chiffre « 0 » entre parenthèses dans la 4ème ligne de programme ci-dessus indique à SAS® la valeur qui a été attribuée aux individus censurés pour la variable relative au CdJ (DECES).

(b)

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
99	76	23	23.23

Number of Observations Read	99
Number of Observations Used	99

(a)

Type 3 Tests

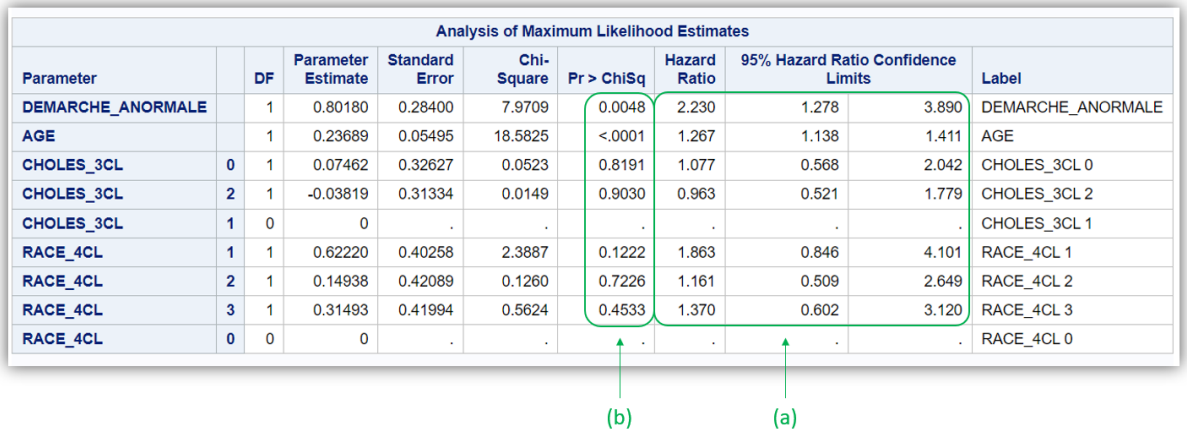
Effect	DF	Wald Chi-Square	Pr > ChiSq
DEMARCHE_ANORMALE	1	7.9709	0.0048
AGE	1	18.5825	<.0001
CHOLES_3CL	2	0.0791	0.9612
RACE_4CL	3	3.2081	0.3606

(c)

Figure 63

La flèche (a) sur la Figure 63 indique que le modèle multivarié de Cox a tourné sur 99 chiens (la totalité de l'échantillon). Les informations pointées par la flèche (b) sur la Figure 63 permettent de voir que

parmi les 99 chiens qui ont participé à l'analyse de survie, 76 ont présenté l'événement au cours du suivi (le décès) et 23 ont été censurés au cours du suivi.



Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
DEMARCHE_ANORMALE		1	0.80180	0.28400	7.9709	0.0048	2.230	1.278	3.890	DEMARCHE_ANORMALE
AGE		1	0.23689	0.05495	18.5825	<.0001	1.267	1.138	1.411	AGE
CHOLES_3CL	0	1	0.07462	0.32627	0.0523	0.8191	1.077	0.568	2.042	CHOLES_3CL 0
CHOLES_3CL	2	1	-0.03819	0.31334	0.0149	0.9030	0.963	0.521	1.779	CHOLES_3CL 2
CHOLES_3CL	1	0	0	.	.	.	.	.	.	CHOLES_3CL 1
RACE_4CL	1	1	0.62220	0.40258	2.3887	0.1222	1.863	0.846	4.101	RACE_4CL 1
RACE_4CL	2	1	0.14938	0.42089	0.1260	0.7226	1.161	0.509	2.649	RACE_4CL 2
RACE_4CL	3	1	0.31493	0.41994	0.5624	0.4533	1.370	0.602	3.120	RACE_4CL 3
RACE_4CL	0	0	0	.	.	.	.	.	.	RACE_4CL 0

(b) (a)

Figure 64

A partir des résultats pointés par les flèches (a) (les HR ajustés avec leur $IC_{95\%}$) et (b) (les degrés de signification) sur la Figure 64, on peut dresser le tableau ci-dessous.

	Adjusted association		
Exposure	aHR	95% CI	P-value
Abnormal gait (*versus* normal gait)	2.23	(1.28-3.89)	< 0.01
Age (x 1-year increase)	1.27	(1.14-1.41)	< 0.01
Cholesterolemia			
Hypocholesterolemia	1.08	(0.57-2.04)	0.82
Normocholesterolemia	1		
Hypercholesterolemia	0.96	(0.52-1.78)	0.90
Breed			
Golden	1		
Labrador	1.86	(0.85-4.10)	0.12
Golden-Labrador	1.16	(0.51-2.65)	0.72
Other breed	1.37	(0.60-3.12)	0.45

aHR, Hazard Ratio adjusted for all the exposures listed in the table; CI, confidence interval.

Pour information, indépendamment de la démarche, de l'âge, et de la race, la cholestérolémie n'était *globalement* pas significativement associée à la survenue d'un décès (p = 0,96 ; cf. Figure 63.c). De même qu'indépendamment de la démarche, de l'âge, et de la cholestérolémie, la race n'était pas non plus *globalement* significativement associée à la survenue d'un décès (p = 0,36 ; cf. Figure 63.c).

3. Vérification de l'hypothèse des risques proportionnels (HRP)

a) Introduction

Le modèle de Cox repose sur l'hypothèse des risques proportionnels (HRP). Plus spécifiquement, chaque variable E_i incluse dans un modèle de Cox doit vérifier l'HRP. Cette hypothèse pour une exposition E_i est vérifiée si et seulement si la valeur du HR quantifiant l'association entre E_i et la survenue du CDJ est constante au cours du temps, au sein de la population cible. Supposons une exposition binaire E, avec des animaux exposés à E (E+) et non exposés à E (E-). La valeur du HR associé à cette variable E ne serait pas constante au cours du temps s'il se passait, par exemple, la chose suivante, dans la population cible : peu de temps après J0, les animaux E+ présentent deux fois plus rapidement le CdJ que les animaux E- (HR = 2) mais plus longtemps après J0, la survenue du CdJ est

aussi rapide entre les animaux E+ et les animaux E- (HR = 1). Cette situation peut arriver lorsque l'exposition étudiée a un effet seulement à court terme sur la survenue du CdJ (peu de temps après J0).

L'analyse de survie réalisée dans l'article de Stablein et coll. (Stablein et al., 1981) présente une situation de non proportionnalité des risques (Cf. courbes de Kaplan-Meier présentées sur la Figure 65).

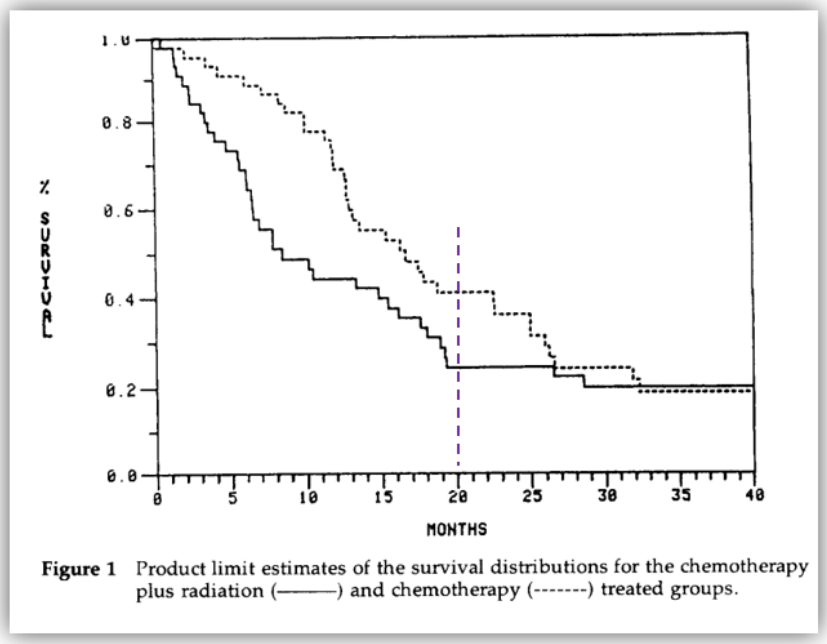


Figure 1 Product limit estimates of the survival distributions for the chemotherapy plus radiation (———) and chemotherapy (-------) treated groups.

Figure 65

Sur la Figure 65., on peut remarquer que le groupe en trait plein (« Chemotherapy + radiation ») présente l'événement plus rapidement que le groupe en trait pointillé (« Chemotherapy ») dans les premiers temps de suivi, et à partir d'environ 20 mois après J0 (droite en pointillé verticale), le phénomène s'inverse, avec le groupe en trait pointillé qui cette fois-ci présente l'événement un peu plus rapidement que le groupe en trait plein.

La situation où il n'y a pas de différence de survenue d'événement entre les individus E+ et les individus E- peu de temps après J0, et avec une différence de survenue d'événement qui commence à apparaître un certain temps après J0, est aussi une situation où l'HRP ne serait pas vérifiée pour E.

L'HRP doit être vérifiée pour chaque variable incluse dans un modèle de Cox (surtout pour celles concernées par le message clinique délivré à l'issu des analyses statistiques), tout en gardant à l'esprit que ce qui peut être observé dans l'échantillon peut ne pas être le reflet de ce qu'il se passe dans la population cible à cause de la fluctuation d'échantillonnage (par exemple ici, une HRP qui serait vérifiée dans l'échantillon mais pas dans la population cible, ou bien une HRP non vérifiée dans l'échantillon mais qui serait vérifiée dans la population cible).

Il existe de nombreuses façons de vérifier l'HRP (Hess, 1995). Je vais vous présenter une méthode graphique et une méthode statistique pour vérifier l'HRP. Je ne vais pas vous démontrer dans ce guide les raisons pour lesquelles les méthodes que je vais décrire permettent effectivement de savoir si l'HRP semble, ou pas, vérifiée.

b) *Méthode graphique de vérification de l'HRP avec SAS®*

Une méthode graphique simple est de modifier les axes des abscisses et des ordonnées d'un graphique de Kaplan-Meier, en mettant en ordonnée Ln(-Ln(S(t))) (au lieu de S(t) dans un graphique de Kaplan-Meier) et en abscisse Ln(t) (au lieu de t dans un graphique de Kaplan-Meier). Si, dans ce graphique, les courbes ont un écart relativement constant, alors l'HRP peut être acceptée.

Supposons que l'on veuille vérifier l'HRP pour la variable binaire DEMARCHE_ANORMALE. Les lignes de programme ci-dessous permettent de la vérifier.

```
PROC LIFETEST DATA = Donnees_pour_guide PLOTS = (LLS);
TIME SURVIE * DECES(0);
STRATA DEMARCHE_ANORMALE;
RUN;
```

Le résultat graphique des lignes de programme ci-dessus est présenté sur la Figure 66.

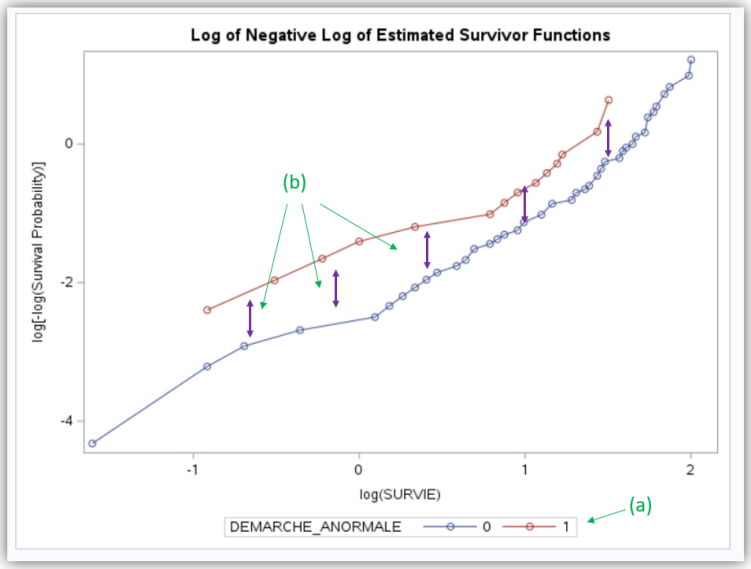


Figure 66

Vous pouvez voir sur la Figure 66 que l'axe des ordonnées est bien Ln(-Ln(S(t))) et que l'axe des abscisses est Ln(t) avec ici t représentée par la variable SURVIE (le temps écoulé depuis J0). Les courbes rouge et bleue (cf. Figure 66.a) représentent respectivement les chiens avec démarche anormale (DEMARCHE_ANORMALE = 1) et les chiens avec démarche normale (DEMARCHE_ANORMALE = 0). On peut voir que l'écart entre les deux courbes est relativement constant au cours du temps (les flèches violettes représentent l'écart entre les deux courbes, cf. Figure 66.b). Ainsi, l'HRP semble vérifiée pour la variable DEMARCHE_ANORMALE. Veuillez noter que si l'écart est tout le temps faible entre les deux courbes (⇔ faible association statistique entre la survenue du CdJ et la variable étudiée), il est attendu que les deux courbes se croisent à une ou plusieurs reprises. Une telle situation n'indique pas que l'HRP n'est pas vérifiée.

Notez que cette méthode graphique ne peut s'appliquer qu'à des variables binaires ou qualitatives. Ainsi, si l'on souhaite vérifier l'HRP pour une variable quantitative, il faut l'avoir recodée en variable qualitative, ou binaire, au préalable. Comme vous l'avez déjà vu dans la sous-partie « Situation d'une variable quantitative » (page 39), il est possible de créer une telle variable au sein de la procédure PROC LIFESTEST. Ainsi, les lignes de programmes ci-dessous vous présente la vérification de l'HRP pour la variable UREE, en prenant comme seuil pour la rendre binaire pour la vérification de l'HRP la valeur de 0,28 g/L, qui correspond à la médiane de la concentration en urée dans l'échantillon (cf. Figure 12).

```
PROC LIFETEST DATA = Donnees_pour_guide PLOTS = (LLS);
TIME SURVIE * DECES(0);
STRATA UREE(0.28);
RUN;
```

Les résultats graphiques des lignes de programme ci-dessus sont présentés sur la Figure 67.

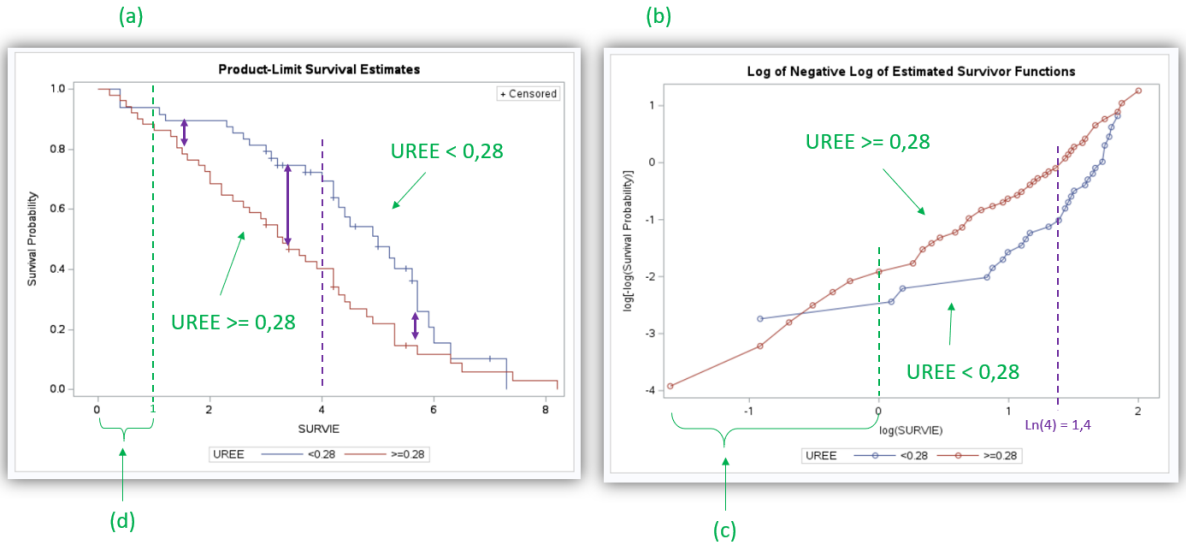


Figure 67

On peut remarquer sur la Figure 67.a que les deux courbes de Kaplan-Meier indiquent que dans l'échantillon, les chiens avec une concentration en urée inférieure à 0,28 g/L (courbe bleue, cf. légende de la Figure 67.a) décédaient moins rapidement que ceux dont la concentration en urée était supérieure ou égale à 0,28 g/L (courbe rouge) dans les premiers temps de suivi seulement (jusqu'à 4 ans de suivi environ), mais qu'ensuite, les chiens avec une concentration en urée inférieure à 0,28 g/L décédaient plus rapidement que ceux dont la concentration en urée était supérieure ou égale à 0,28 g/L (la courbe bleue commençant à rejoindre la courbe rouge à partir de 4 ans après J0). C'est *le* phénomène de non proportionnalité des risques. La Figure 67.b met plus facilement en évidence le fait que l'HRP n'est pas vérifiée pour la variable UREE prise de façon binaire : l'écart entre les deux courbes est loin d'être constant au cours du temps ! Vous pouvez remarquer que l'axe des abscisses sur la Figure 67.b est moins facile à interpréter que celui sur la Figure 67.a, car il s'agit de Ln(t). D'ailleurs, Ln(4 ans) = 1,4 (cf. ligne verticale en pointillé d'abscisse 1,4), et on remarque que l'inversion de l'évolution de l'écart entre les deux courbes sur la Figure 67.b est effectivement à partir de 1,4 (soit 4 ans) sur l'axe des abscisses. Cela dit, bien que cet axe des abscisses soit moins facile à interpréter, la vérification de l'HRP est plus simple : il « suffit » d'évaluer la constance de l'écart entre les deux courbes. Dans la mesure où l'axe des abscisses est exprimée sur l'échelle logarithmique du temps, attention à ne pas prendre en compte avec trop d'importance la partie de gauche de la Figure 67.b pour savoir si l'HRP semble, ou pas, vérifiée : la partie de la courbe repérée par la flèche (c) sur la Figure 67.b jusqu'à l'abscisse 0 représente quasiment la moitié de l'axe des abscisses de la Figure 67.b, mais seulement 1/8ème du suivi total dans l'étude (e^0 = 1 an, cf. flèche (d) sur la Figure 67.a).

c) *Méthode statistique de vérification de l'HRP avec SAS®*

Une méthode statistique simple de vérification de l'HRP, mais qui a malgré fait ses preuves (Ng'andu, 1997), consiste à introduire un terme d'interaction entre la variable dont on cherche à vérifier l'HRP et le temps. Cette méthode est simple pour la vérification de l'HRP pour une variable binaire. Elle est aussi utilisable et reste simple pour une variable qualitative ordinale ou quantitative lorsque l'hypothèse de la linéarité de l'association est vérifiée (et donc, lorsque cette variable peut être incluse telle quelle dans le modèle). La méthode est en revanche plus compliquée pour une variable qualitative nominale, ou qualitative ordinale dont l'association avec la survenue du CdJ ne peut pas être

considérée comme linéaire. Dans cette situation-là, je vous suggère la méthode graphique qui a été présentée ci-dessus.

Je vais d'abord vous présenter un exemple de vérification de l'HRP utilisant la méthode statistique avec une variable binaire (variable DEMARCHE_ANORMALE). Les lignes de programme ci-dessous permettent de le faire. La vérification pour cette variable a déjà été réalisée graphiquement ci-dessus (cf. Figure 66), et nous avions accepté l'HRP pour cette variable.

```
PROC PHREG DATA = Donnees_pour_guide;
MODEL SURVIE * DECES(0) = DEMARCHE_ANORMALE INTER;
INTER = DEMARCHE_ANORMALE * LOG(SURVIE);
RUN;
```

Sur la 3ème ligne de programme ci-dessus, c'est à vous de choisir le nom de la variable d'interaction qui a été créée au sein de la procédure PROC PHREG. J'ai choisi de la nommer « INTER ».

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 68.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
DEMARCHE_ANORMALE	1	0.59942	0.38653	2.4049	0.1210	1.821	DEMARCHE_ANORMALE
INTER	1	0.12225	0.34899	0.1227	0.7261	1.130	

(a) (b)

Figure 68

C'est la valeur du degré de signification du terme d'interaction avec le temps qui va donner des éléments pour penser que l'HRP semble, ou pas, vérifiée. Dans notre exemple avec la vérification de l'HRP pour la variable DEMARCHE_ANORMALE, le degré de signification vaut 0,73 (flèche (a) sur la Figure 68). Si ce degré de signification avait été inférieur à 0,05, on aurait eu des raisons de penser que l'HRP n'aurait pas été vérifiée. Veuillez noter que lorsqu'un terme d'interaction est introduit dans le modèle, le HR pour la variable dont on vérifie l'HRP (ici, de valeur 1,82, cf. Figure 68.b) n'est plus interprétable facilement (cette interprétation dépasse l'objectif de ce guide).

Je vais vous présenter ensuite un exemple de vérification avec la variable quantitative AGE, dont on suppose que l'hypothèse de la linéarité de l'association a été vérifiée. Les lignes de programme ci-dessous permettent de faire cette vérification.

```
PROC PHREG DATA = Donnees_pour_guide;
MODEL SURVIE * DECES(0) = AGE INTER;
INTER = AGE * LOG(SURVIE);
RUN;
```

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 69.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
AGE	1	0.24209	0.07141	11.4932	0.0007	1.274	AGE
INTER	1	-0.03657	0.06121	0.3568	0.5503	0.964	

(a)

Figure 69

La valeur du degré de signification du terme d'interaction est supérieure à 0,05 (p = 0,55 ; cf. Figure 69.a). Ainsi, les données n'apportent pas de preuves fortes que l'HRP n'est pas vérifiée pour la variable quantitative AGE. Maintenant que l'âge semble vérifier l'HRP et puisque la linéarité de l'association

avec l'âge a été vérifiée, là, vraiment, nous pouvons interpréter le résultat pour la variable AGE sur la Figure 64.

Souvenez-vous que la variable UREE, sous forme binaire (< ou ≥ 0,28 g/L), ne semblait pas vérifier l'HRP (cf. Figure 67). Nous allons voir si cette HRP semble ne pas être non plus vérifiée en utilisant la méthode statistique. Les lignes de programme ci-dessous permettent de vérifier l'HRP de façon statistique pour la variable UREE sous forme binaire (< ou ≥ 0,28 g/L).

```
PROC PHREG DATA = Donnees_pour_guide;
MODEL SURVIE * DECES(0) = UREE_BIN INTER;
IF     UREE >= 0.28 THEN UREE_BIN = 1;
IF . < UREE <  0.28 THEN UREE_BIN = 0;
INTER = UREE * LOG(SURVIE);
RUN;
```

Dans les lignes de programme ci-dessus, vous pouvez vous rendre compte qu'il est possible de créer une variable (ici, les variables UREE_BIN et INTER) au sein de la procédure PROC PHREG – alors que ce n'est pas possible pour de nombreuses autres procédures, comme la PROC GLM ou la PROC LOGISTIC. Vous remarquerez aussi que lorsque l'on crée une nouvelle variable à partir des valeurs d'une variable quantitative, il ne faut surtout pas oublier l'impact de données manquantes ! En effet, si j'avais écrit « IF UREE < 0.28 THEN UREE_BIN = 0; » sur la 4ème ligne de programme, les chiens avec une concentration en urée manquante auraient été considérés comme des chiens avec une concentration en urée inférieure à 0,28 g/L (ce qui aurait été une belle erreur).

Le résultat (extrait) des lignes de programme ci-dessus est présenté sur la Figure 70.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
UREE_BIN	1	0.89551	0.29576	9.1675	0.0025	2.449
INTER	1	-2.16048	1.14668	3.5499	0.0595	0.115

(a)

Figure 70

La valeur du degré de signification du terme d'interaction est égale à 0,06 (cf. Figure 70.a). Dans la mesure où il est non significativement différent de 1 (puisque le degré de signification est supérieur à 0,05), certains pourraient dire que l'HRP peut être acceptée pour la variable UREE sous forme binaire. Mais d'autres pourraient dire que la Figure 67.b ne permet quand même pas de l'accepter… (Et là, vous pouvez relire la sous-partie « Vérification d'hypothèses sur lesquelles repose un modèle de régression », page 40… !)

IX. Liens Internet vers l'aide de SAS® pour les procédures utilisées dans ce guide

Cliquez sur les liens ci-dessous pour accéder à l'aide en ligne de SAS® pour chacune des procédures qui ont été traitées dans ce guide.

PROC IMPORT

PROC CONTENTS

PROC FREQ

PROC UNIVARIATE

PROC MEANS

PROC SGPLOT

PROC TTEST

PROC NPAR1WAY

PROC ANOVA

PROC CORR

PROC LIFETEST

PROC GLM

PROC LOGISTIC

PROC PHREG

X. Références

Altman, D.G. and Royston, P., 2006. The cost of dichotomising continuous variables. Bmj. 332, 1080.

Brenner, H. and Blettner, M., 1997. Controlling for continuous confounders in epidemiologic research. Epidemiology. 8, 429-34.

Cox, D.R., 1972. Regression models and life tables (with discussion). J R Statist Soc B. 34, 187-220.

Darnis, E., Boysen, S., Merveille, A.C., Desquilbet, L., Chalhoub, S. and Gommeren, K., 2018. Establishment of reference values of the caudal vena cava by fast-ultrasonography through different views in healthy dogs. J Vet Intern Med. 32, 1308-1318.

Desquilbet, L., 2020. Enhancing Clinical Decision-Making: Challenges of making decisions on the basis of significant statistical associations. J Am Vet Med Assoc. 256, 187-193.

Desquilbet, L. and Mariotti, F., 2010. Dose-response analyses using restricted cubic spline functions in public health research. Stat Med. 29, 1037-57.

Hess, K.R., 1995. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Stat Med. 14, 1707-23.

Ng'andu, N.H., 1997. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. Stat Med. 16, 611-26.

Royston, P., Altman, D.G. and Sauerbrei, W., 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 25, 127-41.

Stablein, D.M., Carter, W.H., Jr. and Novak, J.W., 1981. Analysis of survival data with nonproportional hazard functions. Control Clin Trials. 2, 149-59.