Compte rendu final
Projet ASM

T.Kdous, D.Sportouch, R.Riche, R. Fellous, T.Ho-Duc,
C.Fotso-Talla, X.Milhaud

Table des matières

1 Introduction

Pour cette étude, nous disposons d'un jeu de données concernant 1856 accidents pour lesquels la tranche horaire à laquelle ils ont eu lieu, la tranche d'âge des victimes et la vitesse du véhicule ont été relevées. A partir de ces données et des outils statistiques dont nous disposons, nous allons analyser ces données pour tenter de donner des informations utiles à la gendarmerie concernant l'accidentologie. Nous devons faire très attention à l'acuité de nos résultats car la vie d'êtres humains est en jeu. Pour cela, nous allons essayer de tirer un maximum d'informations du jeu de données, en appliquant divers techniques d'analyse statistique multidimensionnelle.

2 Régression linéaire simple

2.1 Introduction

La première technique que nous allons utiliser est la régression linéaire simple. Cette analyse va nous permettre de déterminer si de prime abord il y a une relation entre les vitesses enregistrées et les créneaux horaires. Il semble raisonnable de nous renseigner sur cette relation, ce qui nous permettrait alors de dire à la gendarmerie de multiplier leurs contrôles à un certain créneau horaire.

2.2 Etude préalable

Avant de réaliser notre régression linéaire, il s'agit tout d'abord de choisir l'origine des temps qui garantira une pertinence maximale à notre régression. Nous avons ainsi réaliser 8 régressions en changeant à chaque fois l'origine de temps (le créneau horaire correspondant à la valeur 0 est à chaque fois changé). Au vu de ces études, nous avons retenu l'étude de régression pour laquelle la p-valeur du test de pertinence de la régression était la plus faible. Notre but est d'obtenir une étude la plus pertinente possible.

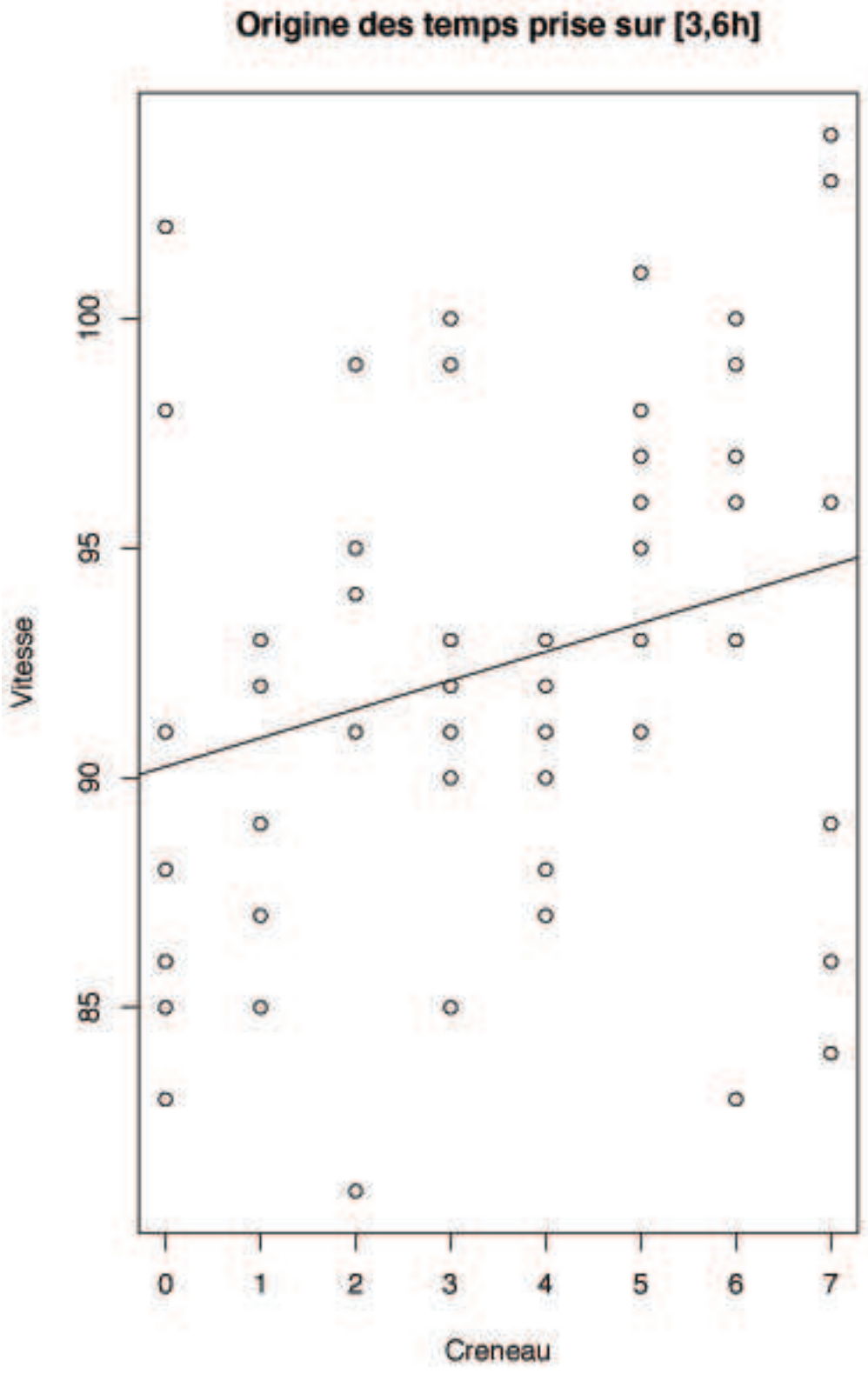
Aux vues des résultats obtenus pour les 8 régressions possibles, nous avons donc décidé de choisir le créneau horaire 3h-6h comme origine des temps pour l'étude de la régression linéaire. En effet, en prenant ce créneau horaire comme origine des temps nous obtenons la p-valeur (du test de pertinence de la regression) la plus faible : 0.05457 (cf sorties R plus bas).

2.3 Sorties R

Origine des temps prises pour t=3h (le premier créneau horaire est 3h-6h).

```
Call:
lm(formula = Vitesse ~ Creneau, data = AccidentTable36)

Residuals:
      Min        1Q    Median        3Q       Max
-10.99320  -3.86395   0.07143   3.51020  11.76190

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  90.2381     1.3324  67.726   <2e-16 ***
Creneau       0.6259     0.3185   1.965   0.0546 .
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 5.461 on 54 degrees of freedom
  (1799 observations deleted due to missingness)
Multiple R-Squared: 0.06673,Adjusted R-squared: 0.04945
F-statistic: 3.861 on 1 and 54 DF,  p-value: 0.05457
```

Voici maintenant le nuage des points avec la droite des moindres carrés associée.


Origine des temps prise sur [3,6h]
Vitesse
85
90
95
100
0
1
2
3
4
5
6
7
Creneau

2.4 Interprétation des résultats de la régression linéaire simple

La sortie R que nous avons est obtenue grâce à la commande lm(vitesse creneau) en prenant le créneau horaire 3h-6h comme origine des temps.

Les estimateurs par la méthode des moindres carrés des coefficients de la régression linéaire simple β_0 et β_1 valent : $\hat{\beta}_0 = 90.2381$ et $\hat{\beta}_1 = 0.6259$.
D'après les résultats R, nous pouvons rejeter largement l'hypothèse $\beta_0 = 0$ étant donné que la p-valeur est extrêmement faible (<2e-16).

La p-valeur correspondant à la pertinence de la régression linéaire pour notre test est de 0.05457. Cette p-valeur est relativement faible mais pas assez pour conclure que la régression est réellement pertinente. Il faudrait que la p-valeur soit inférieure à 10^{-4} pour conclure qu'il y a une véritable dépendance affine entre la vitesse et le créneau horaire.

Nous avons par ailleurs effectué une régression linéaire simple de la vitesse en fonction de l'age. La p-valeur pour cette régression linéaire est de 47.92%, cette régression linéaire est donc très peu pertinente. Ceci montre qu'il n'y a pas de relation affine entre la vitesse des accidents et l'age des conducteurs.

Pour le moment, nous ne pouvons pas nous permettre de donner des informations claires et précises à la gendarmerie. En effet, les résultats ne nous permettent pas d'annoncer clairement qu'il existe une relation entre la vitesse et les créneaux horaires, ou bien la vitesse et les tranches d'âges. Nous allons poursuivre notre étude en réalisant à présent une régression linéaire multiple sur la vitesse.

3 Régression linéaire multiple sur la vitesse

3.1 Introduction

Il est a présent intéressant de voir si la vitesse peut dépendre des deux facteurs Créneau et Age, à la fois. Nous ne prendrons en considération que les données pour lesquelles la vitesse n'est pas donnée.

3.2 Choix de l'origine temporelle pour l'étude

Nous avons choisi le même créneau d'origine que pour la régression linéaire simple car c'est encore avec ce créneau (3-6h) que l'on a obtenu la

plus petite p-valeur et par conséquent la régression la plus pertinente.

Pour réaliser notre étude nous avons eu besoin de "codifier" nos données. Ainsi, chaque créneau d'âge ou d'horaire est remplacé par une valeur entière. Le créneau 0-3h est le numéro 0, le créneau 3-6h est le numéro 1, etc...

3.3 Application de la régression linéaire multiple sur la vitesse

Voici donc le résultat de l'application de la régression linéaire multiple sur la vitesse avec le créneau 3-6h comme origine :

```
Call:
lm(formula = Vitesse ~ Creneau + Age, data = AccidentTable)

Residuals:
    Min      1Q  Median      3Q     Max
-11.797  -3.290   0.270   3.187  11.762

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  89.4345     1.7319  51.639   <2e-16 ***
Creneau       0.6259     0.3199   1.956   0.0557 .
Age           0.2679     0.3665   0.731   0.4681
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.485 on 53 degrees of freedom
  (1799 observations deleted due to missingness)
Multiple R-Squared: 0.07604,        Adjusted R-squared: 0.04118
F-statistic: 2.181 on 2 and 53 DF,  p-value: 0.1230
```

3.3.1 Interprétations

Les valeurs éstimées de $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ sont respectivement : 89.4345, 0.6259, 0.2679. Cela confirme une faible dépendance entre la vitesse et l'âge. La p-valeur du test d'hypothèse de $\beta_2 = 0$ contre β_2 différent de 0 est grande (0.4681) ce qui consolide le raisonnement précédent.

3.3.2 Intervalles de confiance pour les paramètres β

Pour nous assurer de la véracité des valeurs des β, nous avons réalisé un script R qui calcule les 3 intervalles de confiance de seuil $\alpha = 5\%$ pour β_0, β_1, et β_2 . Les résultats sont :

- pour β_0 : $[85.04009; 91.82891]$
- pour β_1 : $[-0.09286147; 1.34466147]$
- pour β_2 : $[-0.3530357; 0.8888357]$

D'après les résultats de ces intervalles de confiance, on est sur à 95% d'avoir une valeur maximale de 1.34466147 pour β_1 (Créneau horaire) et de 0.8888357 pour β_2 (Age). $\hat{\beta}_1$ est strictement supérieur à $\hat{\beta}_2$ donc la vitesse dépend plus du créneau horaire que de l'âge. Des tests d'hypothèses permetteront d'appronfondir ces conjectures.

En outre, l'intervalle de confiance de β_2 contient 0. Ceci confirme donc le fait que la valeur de β_2 (coefficient pour l'Age) est proche de 0. La p-valeur du test d'hypothèse $\beta_2 = 0$ contre $\beta_2 \neq 0$ est de 46.81%. Ceci corrobore donc le fait que la valeur de ce coefficient est proche de 0. Les vitesses des accidents observées dépendent donc très peu de l'âge des victimes.

Le coefficient β_1 admet pour plus petite valeur -0.09286147 avec une confiance de 95%. On en déduit donc ce coefficient est positif avec une grande confiance. Ce coefficient est en rapport avec le créneau horaire, on en déduit que les vitesses auxquels les accidents ont lieu augmentent à mesure que le créneau horaire augmente et donc que la journée s'écoule.

Enfin, l'intervalle de confiance pour β_0 nous montre que les accidents ont lieu principalement à des vitesses supérieures à 85 km/h.

3.3.3 Intervalles de confiance pour le paramètre σ^2

Nous cherchons dans cette partie à donner un intervalle de confiance de seuil α pour σ^2.
Pour déduire l'intervalle de confiance de l'étude théorique, il faut faire l'hypothèse que le modèle que nous étudions est un modèle linéaire gaussien.

La valeur de σ^2 trouvee est 5.485. L'intervalle de confiance au seuil 5% pour σ^2 est : $[3.875970; 8.359278]$

σ^2 représente l'ampleur avec laquelle la vitesse varie autour de la moyenne (égale à 92km/h avec nos régresseurs).

Nous avons une certitude de 95% que cette variabilité est entre 3.88 et 8.36 nous permet d'affirmer avec certaine confiance que nos accident ce produisent entre 84km/h et 100km/h (dans le pire cas : $\sigma^2 = 8$). Cette faible variance des résidus nous donne une bonne confiance et une certaine précision dans la moyenne empirique calculée.

3.4 Tests de pertinence de la régression linéaire

Nous voulons savoir si la régression linéaire est pertinente. Nous voulons à présent réaliser le test de pertinence suivant :
$H_0 : \beta_2 = \beta_1 = 0$ contre $H_1 : \bar{H_0}$

3.4.1 Résultat R

La p-value de ce test étant déjà donnée par R lors de la regression linéaire, nous nous sommes simplement assurés qu'avec la p-value de R, nous sommes au bord de la région critique.

3.4.2 Interprétation

Nous retrouvons bien les mêmes valeurs. Notre calcul est donc bien le même que celui effectué par le logiciel R.

La p-valeur est est de $12,30\%$, ce qui est assez élevé. En conséquence, la régression linéaire multiple n'est pas très pertinente pour nos données. Les résultats qu'elle fournit sont donc à relativiser par rapport à cette p-valeur.

3.5 Comparaison relative de l'effet des 2 régresseurs

Nous avons réalisé un test d'hypothèse permettant de comparer l'influence relative des deux facteurs Age et Creneau sur la Vitesse des accidents dans le modèle de la régression linéaire multiple.

Nous voulons à présent réaliser le test de pertinence suivant :
$H_0 : \beta_2 = \beta_1$ contre $H_1 : \bar{H_0}$

En utilisant le test de Fisher (Théorème 14 page 45 du cours), on aboutit a la conclusion suivante : Pour un seuil de 5%, on peut donc rejeter l'hypothèse H_0 ($4.504 > 4.023$) et ainsi conclure que les deux facteurs n'ont pas

la même influence sur la vitesse dans la régression linéaire multiple. D'après les résultats et les interprétations précedentes, on en déduit que c'est bien le creneau qui a plus d'influence que l'age dans la régression linéaire multiple.

3.6 Bilan sur la régression linéaire multiple

La p-valeur du test de pertinence est de $12,30\%$, ce qui est supérieur à celle obtenue $(5,4\%)$ pour la régression linéaire simple de la vitesse sur le créneau. L'introduction du paramètre "âge" a donc introduit du bruit dans la modélisation. Ainsi, la vitesse ne semble pas dépendre linéairement du couple "créneau-age", et plus particulirement de l'âge. Ceci est confirmé par la forte p-valeur (46%) calculée par R pour le test $\beta_2 = 0$ contre $\beta_2 \neq 0$.

La gendarmerie ne devrait donc pas prendre en considération l'âge des conducteurs pour ses procédures de contrôle car il ne semble pas y avoir de dépendance linéaire. D'autres part, la vitesse d'occurence des accidents augmente dans la journée a partir de six heure du matin. En effet, on peut affirmer avec 95% de certitude d'avoir $\beta_1 > -0,09$. Donc la gendarmerie devra intensifier ses contrôles vers la fin de la journée.

Les résultats obtenus dans notre étude sont cohérents (estimation, intervalles de confiance, tests de pertinence), mais les valeurs obtenues ne nous permettent pas d'émettre avec certitudes des conclusions franches.

Nous allons a présent étudier l'effet de variables quantitatives controlées (les facteurs Créneau et Age) sur une variable quantitative observée (la vitesse), en appliquant l'ANOVA 1 à notre projet.

4 Analyse de variance : méthode de l'ANOVA

4.1 Les hypothèses pour l'application de l'ANOVA1

Nous allons vérifier si la vitesse dépend des 2 régresseurs suivants : facteur classe d'âge et facteur tranche horaire. Nous allons donc nous servir de notre fichier texte nettoyé ne comportant que les 56 valeurs de la vitesse. On réalise donc 56 expériences. La vitesse corespond à un vecteur de variables quantitatives. Posons Y le vecteur des vitesses.

Nous suppsons les hypothèses suivantes vérifiées :
(H1) : Les Y_i sont 2 à 2 indépendantes et normalement distribuées.
(H2) : Les facteurs ne peuvent influencer que l'espérance de la loi de Y_i.

(H3) : Les Y_i ont toutes la même variance σ^2

La validation des trois hypothèses précédentes nous permet d'appliquer la méthode de l'ANOVA 1 à notre problème de statistiques multidimentionnelles.

4.2 Effet de la classe d'âge sur la vitesse

4.2.1 Etude théorique

D'après l'étude faite avec la régression linéaire multiple, il s'est avéré que la classe d'âge n'avait pas d'influence sur la vitesse dans le modèle de la régression linéaire. Nous allons essayer de corroborer cela, avec la méthode de l'ANOVA 1.

Les 7 intervalles correspondant aux tranches d'age vont correspondre aux modalités pour notre ANOVA.

Nous allons ici nous intéresser aux valeurs moyennes de la vitesse pour chaque tranche d'âge, ainsi qu'aux résultats du test d'influence du facteur (Test : $H_0 : m_1 = m_2 = \cdots = m_7$ contre $H_1 : \bar{H}_0$).

4.2.2 Résultats R

Voici les résultats qui ont été obtenus avec les commandes R :

```
Tables of means
Grand mean

92.42857

 age
age
  -14 14-16 16-18 18-21 21-25 25-65    65
90.38 89.75 90.62 98.25 96.38 92.50 89.12

> summary(Accident.ANOVA)
            Df  Sum Sq Mean Sq F value   Pr(>F)
age          6  600.21  100.04  4.3552 0.001340 **
Residuals   49 1125.50   22.97
---
```

```
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1
```

4.2.3 Test d'influence du facteur "classe d'âge" sur la vitesse

Les résultats de R nous donnent une p-valeur très faible égale à 0.00134, ce qui permet de dire que la vitesse a une forte dépendance en fonction de l'âge. Cette observation semble cohérente, il est facile d'imaginer que la tranche d'âge des jeunes est probablement moins consciente des risques liés à la vitesse.

Cette possibilité sera d'ailleurs validée par la suite avec nos résultats (cf figure 1 où l'on voit que la tranche 3 correspondante aux 18-21 ans est celle qui roule le plus vite !).

Grâce au théorème de Cochran, nous savons que la région critique s'écrit de la forme :
$W = \left\{ y \in \mathbb{R}; \frac{n-p}{p-q} \frac{\|\pi_{E_x \cap E_0^\perp} y\|^2}{\|\pi_{E_x^\perp} y\|^2} > f_{p-q,n-p,\alpha} \right\}$

On remarque que pour le facteur âge, on se situe bien dans la région critique puisque $4.019 < 4.3552$. Donc on rejette (H0), c'est à dire qu'on rejette le fait que l'âge n'a pas d'influence sur la vitesse. On peut affirmer que le facteur âge influe sur la vitesse.

4.2.4 Interprétations

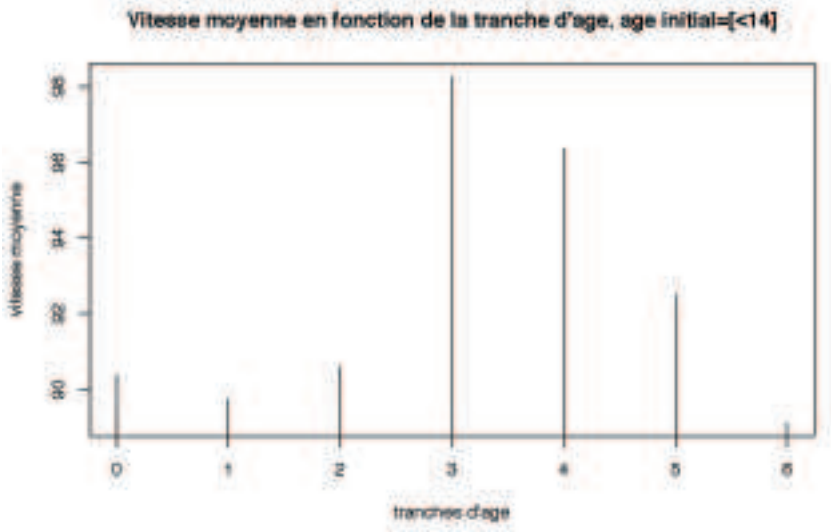
L'ANOVA 1 réalisée sur la tranche d'âge nous apporte des informations intéressantes.
En effet, les résultats obtenus ici sur l'influence des facteurs Age et Créneau sur la Vitesse corroborent les résultats obtenus lors de la régression linéraire.

Tout d'abord, la p-valeur du test d'influence de l'âge sur la vitesse est de 0.00134, ce qui est très faible. On en déduit donc que l'âge a une forte influence sur les valeurs prises par la vitesse. Lors de la régression linéaire, nous avions remarqué que l'âge introduisait du bruit dans notre modèle car l'introduction de ce facteur augmentait la p-valeur correspondante à la pertinence de la régression. Ceci est en accord avec le modèle de l'ANOVA.
En effet, vu que les valeurs de la vitesse varient beaucoup et surtout de manière non-monotone selon la modalité (intervalle) choisie pour l'âge, le modèle de la régression linéaire (Vitesse-Age) était difficile à appliquer. Le fait qu'on obtienne une p-valeur de 46% pour la régression linéaire simple de la vitesse

en fonction de l'âge corrobore cette conclusion.

L'influence des facteurs est donc relative au modèle choisi.

L'ANOVA et la régression linéaire effectuées précedemment nous permettent donc de conclure que la vitesse des accidents varie beaucoup selon la tranche d'âge.



On pourra remarquer que c'est la tranche d'âge des 18-21 ans qui roule avec la plus grande vitesse, 98.25 km/h en moyenne, alors que ce sont les plus de 65 ans qui roulent le plus lentement, 89.12 km/h en moyenne. Aux vues de ce graphique, il apparait clairement que ce sont globalement les personnes âgées de 18 à 25 ans qui roulent le plus vite parmi tous les conducteurs qui ont eu un accident, tandis que les autres personnes roulent à des vitesses autour de 90 km/h en moyenne.

4.3 Effet de la tranche horaire sur la vitesse

4.3.1 Etude théorique

Ici les 8 intervalles correspondant aux tranches horaires correspondent aux modalités utilisées pour le créneau pour l'ANOVA à un facteur.
Nous allons ici nous intéresser aux valeurs moyennes de la vitesse pour chaque tranche d'âge, ainsi qu'aux résultats du test d'influence du facteur (Test : $H_0 : m_1 = m_2 = \cdots = m_8$ contre $H_1 : \bar{H}_0$).

4.3.2 Résultats R

Voici les résultats qui ont été obtenus avec les commandes R :

```
Tables of means
Grand mean

92.42857

 Creneau
Creneau
  0-3h 12-15h 15-18h 18-21h 21-24h   3-6h   6-9h  9-12h
 92.57  92.86  90.29  95.86  95.43  90.43  89.29  92.71

> summary(Accident.ANOVA)
            Df  Sum Sq Mean Sq F value Pr(>F)
Creneau      7  276.57   39.51  1.3087 0.2667
Residuals   48 1449.14   30.19
```

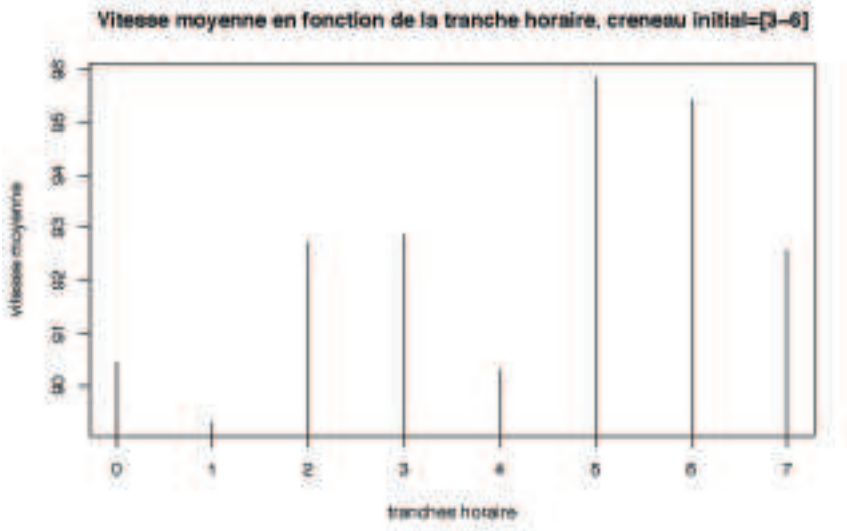
4.3.3 Test d'influence du facteur "tranche horaire" sur la vitesse

Les résultats de R nous donnent une p-valeur plutôt élevée, égale à 0.2667, ce qui laisse croire que le créneau horaire n'a pas d'influence directe sur la vitesse, autrement dit, il n'y a pas de créneau horaire où des vitesses nettement plus élevées ont été enregistrées lors des accidents.
Les résultats que l'on obtient à l'aide des sorties R nous permettent de dire que l'on ne peut pas rejetter (H0), c'est à dire que l'on ne peut pas conclure que le créneau horaire a une influence sur la vitesse.
Ceci est cohérent avec le fait que la p-valeur est nettement supérieure à $\alpha = 0.05$, puisqu'on trouve une p-valeur égale à 27%.

4.3.4 Interprétations

L'ANOVA 1 réalisée sur la tranche horaire nous apporte des informations intéressantes.
La p-valeur obtenue pour le test d'influence du Creneau sur la Vitesse des accidents dans l'ANOVA 1 est de 26%. Les valeurs de la Vitesse varient peu selon la tranche horaire. Ceci est en accord avec la régression linéaire où on a observé que la régression linéaire simple (Vitesse Creneau) était pertinente (p-valeur=5.4%).

L'ANOVA et la régression linéaire effectuées précedemment nous permettent donc de conclure que la Vitesse des accidents varie peu selon la tranche horaire.



L'ANOVA 1 nous informe que les vitesses moyennes les plus élevées ont été enregistrées pour la tranche horaire 18-21 heures, tandis que les vitesses moyennes les moins élevées ont été enregistrées pour la tranche horaire 6-9 heures.

Globalement, l'histogramme permet d'affirmer que les vitesses les plus élevées ont été enregistrées entre 18 heures et minuit.

Il est intéressant à présent d'effectuer une ANOVA 2 (étude possible car il y a deux facteurs qui agissent sur la vitesse). Cependant, cette étude de l'ANOVA 2 n'est pas faisable.

En effet, dans le cadre de notre projet, le facteur Créneau peut prendre 8 valeurs possible, et a donc 8 modalités. Le facteur Age en a 7.
Le nombre total de combinaisons des niveaux des facteurs est alors $7*8 = 56$ modalités. Le nombre d'expériences étant aussi égal à 56 (une valeur de la vitesse obsevée par combinaison), il est alors inutile de calculer la moyenne m_{ij} représentant l'effet sur la Vitesse de la combinaison du niveau i du facteur Age et du niveau j du facteur Creneau.
L'ANOVA2 est donc inapplicable pour le moment. Pour palier ce probleme, nous allons utiliser la méthode de l'AFC, afin de voir quels sont les regroupements possibles que l'on peut faire parmi des différentes modalités des facteurs.

5 ACP

Nous voulons tirer un maximum d'informations des données. Cela signifie que l'on veut pouvoir comparer les données entre elles, et de déterminer si des rapprochements sont faisables. Pour cela, il nous faut appliquer l'ACP. Cependant, dans le cadre de notre projet, l'ACP sur les variables ne fonctionne pas car les variables Age et Créneau sont des variables qualitatives. De plus, en ce qui concerne l'ACP sur les individus, elle est inutile car les accidents constituent ici les individus, et les accidents sont indifférentiables. Donc, dans le cadre de notre projet, nous ne pouvons pas appliquer l'ACP.

Cependant, notre fichier est constitué de 1856 données. Or, lors de chaque étapes précédentes, nous avons été forcé d'utiliser uniquement les données pour lesquelles une valeur de la vitesse a été donnée. Cela nous fait perdre une grande partie de l'information puisque nous n'exploitons que 56 valeurs parmi les 1856 que nous possédons. L'application d'une autre méthode d'analyse, l'AFC, va nous permettre non seulement de pouvoir exploiter les 1856 données, mais aussi, en permettant le regroupement de certaines modalités, d'appliquer l'ANOVA 2 à notre projet.

6 AFC

6.1 Etude théorique

6.1.1 Etude de l'indépendance entre les facteurs Age et Creneau

Pour étudier l'indépendance entre ces deux facteurs nous allons utiliser le test du χ^2. Grâce à ce test nous allons pouvoir déterminer si deux facteurs sont indépendants ou non. L'hypothèse H_0 est que ces deux facteurs sont indépendants alors que l'hypothèse H_1 est qu'ils sont liés.
La forme de la région critique de ce test est $W = \{\delta_n^2 > z_{(l-1)(c-1),\alpha}\}$ avec $l = 7$ (tranche d'âge) et $c = 8$ (créneau horaire) ici.δ_n^2 désigne ici la statistique du test du χ_2.

6.1.2 ACP des deux nuages de profils

ACP sur les profils-lignes On réalise ici une ACP sur les individus qui va permettre de dégager les proximités entre les différentes tranches d'âges au niveau des accidents (modalités de Age).

ACP sur les profils-colonnes On réalise ici une ACP sur les individus qui va permettre de dégager les proximités entre les différents créneaux horaires au niveau des accidents (modalités de Créneau).

6.2 script R

6.2.1 test du chi-2 :

Voici le résultat du test du χ_2 :

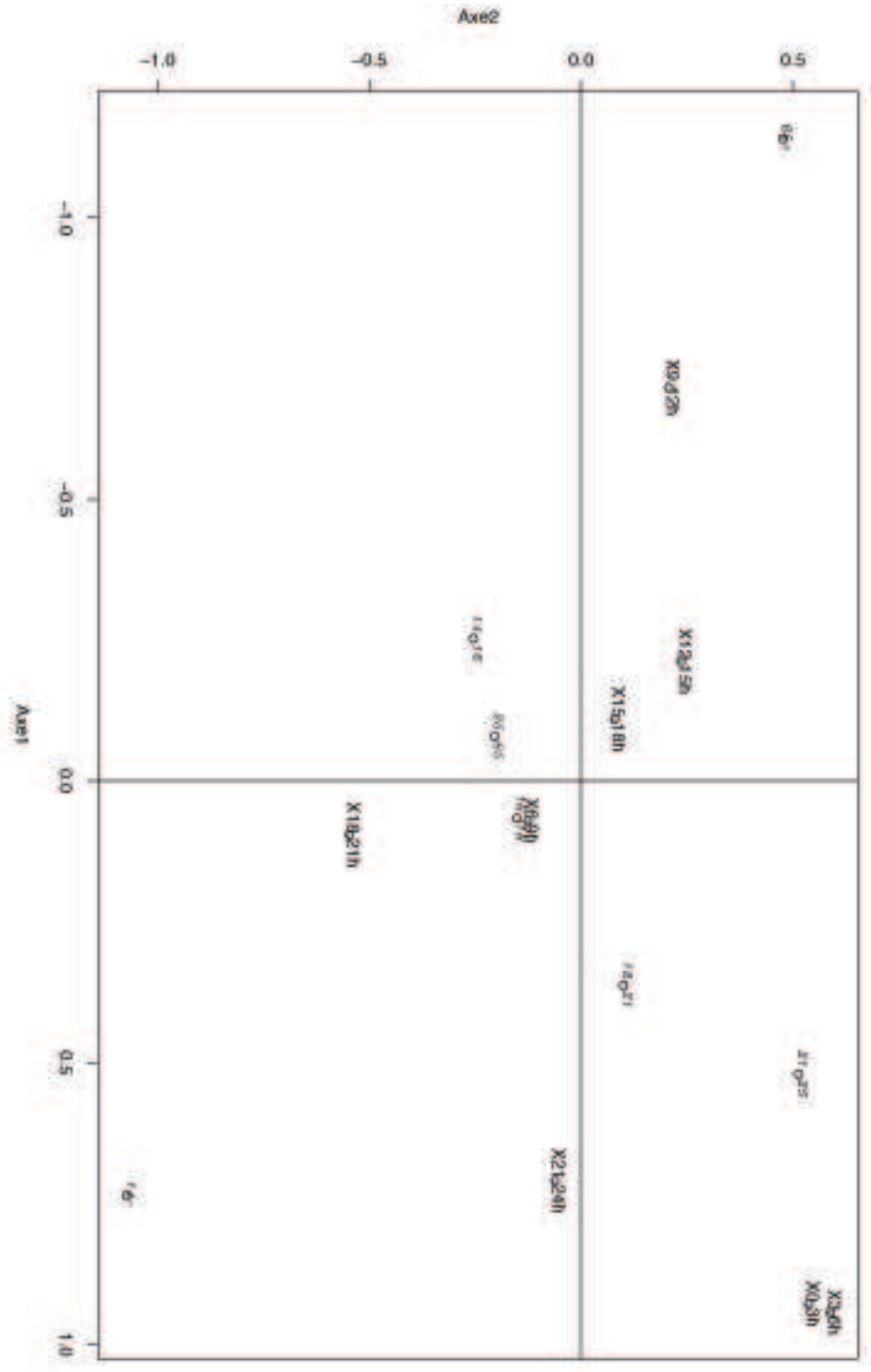
```
Pearson's Chi-squared test

data:  TabAccident
X-squared = 102.1545, df = 42, p-value = 6.361e-07
```

6.2.2 Qualité de représentation des individus :

Voici la qualité de représentation de chacun des individus (Age ou Creneau) dans le plan formé par les Axes 1 et 2. La formule permettant de calculer les représentations des individus ne dépend pas de la métrique choisie et peut donc être utilisée ici.
Nous récupérons les matrices tronquées correspondantes aux axes 1 et 2, donc les données pertinentes.

6.2.3 Graphe biplot représentant les modalités dans le premier plan propre :

Voici maintenant le graphe représentant la projection de toutes nos modalités dans le premier plan propre de notre ACP sur les individus (Age ou Creneau horaire). Le graphe est ici un graphe biplot :


Axe2
−1.0
−0.5
0.0
0.5
Axe1
−1.0
−0.5
0.0
0.5
1.0

6.3 Interprétations

6.3.1 test du chi-2 :

Tout d'abord la p-valeur de notre test du χ_2 est égale 6.36110^{-7}, ce qui est extrêmement faible. On peut donc conclure que nos deux variables Age et Créneau horaire sont très fortement dépendantes. L'AFC est donc pertinente dans notre cas. De plus cette p-valeur extrêmement faible qui traduit une forte dépendance entre nos deux variables montre que le choix d'effectuer un graphe biplot est pertinent.
D'après le tableau d'indépendance,ceci signifie donc que les accidents pour les plus de 65 ans sont sur-représentés dans l'intervalle de temps 9h-12h (17%). De plus on remarque que les accidents concernant les personnes entre 21 et 25 ans sont sur-représentés dans la tranche horaire 3h-6h.

6.3.2 Qualité de représentation des individus :

On a déterminé la représentation de chacune de nos modalités dans le premier plan propre de projection des individus. Voici maintenant la liste des modalités bien représentées dans notre plan propre (elles ont une représentation de qualité supérieure à 50%) :

Pour nos interprétations dans le graphe, on ne s'intéressera qu'aux individus bien représentés. C'est-à-dire les individus pour lesquels la qualité de représentation sur le plan formé par les axes 1 et 2 est supérieure à 50%. Il s'agit des individus :
Pour l'Age : 14-, 18-21, 21-25, 25-65, 65+.
Pour le Créneau horaire : X0.3h, X3.6h , X9.12h, X12.15h, X15.18h, X18.21h, X21.24h.

6.3.3 Interprétation du graphe :

Dans ce graphe, on remarque tout d'abord dans le cadran en haut à gauche que les personnes agées de plus de 65 ans ont l'habitude de rouler le matin (créneau horaire : 9h-12h) et ont donc des accidents principalement dans ce créneau horaire.
De plus, dans le cadran en haut à droite, on voit que les modalités 18-21 et 21-25 pour l'âge des victimes sont très proches et peuvent donc ici être agrégées en une seule modalité 18-25 pour la variable Age. Les modalités 0h-3h et 0h-6h sont aussi très proches et peuvent donc être agrégées en une seule modalité 0h-6h. Avec la présence du créneau horaire 21h-24h non loin de ce même cadran, on peut en déduire que les jeunes (entre 18 et 25 ans)

ont l'habitude de rouler la nuit (entre 21h et 6h) et ils ont donc tendance à avoir des accidents dans ce créneau horaire et plus particulèrement entre minuit et six heures du matin.
L'axe 1 oppose donc à la fois les personnes jeunes (18-25 ans) aux personnes agées (+ de 65 ans) et les personnes roulant le matin (de 9h-12h principalement) à celles roulant la nuit (21h à 6h).

En ce qui concerne l'axe 2 il est difficile de dégager des conclusions car très peu de modalités sont bien représentées dessus. La modalité 18h-21h est bien représenté sur cette axe dans la partie inférieure du graphe, de même que la modalité 12h-15h dans la partie supérieure du graphe. On peut donc penser que l'Axe 2 du graphe oppose les personnes roulant en début de journée à celles roulant en fin de journée.
Enfin la modalité concernant les moins de 14 ans est excentrée de toutes nos modalités et n'a donc aucun lien avec nos autres modalités sur l'âge des victimes ou l'heure des accidents. Ceci est du au fait que les moins de 14 ans ne rentre pas dans le cadre de cette étude car ils n'ont pas en mesure de conduire.

7 ANOVA 2

7.1 Etude théorique

Lors du TP8, nous avions tenté d'appliquer l'ANOVA2 à nos données de l'accidentologie mais sans succès. En effet, dans le cadre de notre projet, le facteur Créneau peut prendre 8 valeurs possible, et a donc 8 modalités. Le facteur Age en a 7.
Le nombre total de combinaisons des niveaux des facteurs est alors $7*8 = 56$ modalités. Le nombre d'expériences étant aussi égal à 56 (une valeur de la vitesse obsevée par combinaison), il est alors inutile de calculer la moyenne m_{ij} représentant l'effet sur la Vitesse de la combinaison du niveau i du facteur Age et du niveau j du facteur Creneau.
L'ANOVA2 était donc inapplicable sur les données de notre projet car nous n'avons qu'une seule valeur (observation) pour chaque couple (Creneau,Age).

Grâce à l'AFC réalisée sur les variables "Créneau horaire" et "Tranche d'âge" dans la partie 1, nous avons pu mettre en évidence les proximités entre les différentes modalités du créneau horaire (0-3h, 3-6h, 6-9h, 9-12h, 12-15h, 15-18h, 18-21h, 21-24h) et de la tranche d'age (strictement moins de 14 ans, 14 à 16ans, 16 à 18ans, 18 à 21 ans, 21 à 25 ans, 26-65 ans, 65 ans ou plus).

Ainsi nous avons remarqué que nous pouvons regrouper certaines modalités similaires (ayant une très grande proximité dans l'AFC) en une seule.
Nous allons appliquer l'ANOVA2 à ce nouveau jeu de données où les modalités similaires ont été regroupées en une seule. Ainsi pour un couple de deux modalités (Creneau horaire, Tranche d'age), on aura plusieurs valeurs de vitesses relevées (plusieurs accidents). On pourra donc effectuer des moyennes et ainsi appliquer l'ANOVA2.
D'après les résultats obtenus lors de l'AFC, nous allons donc agréger les modalités 18-21 et 21-25 en 18-25, ainsi que les modalités 0h-3h et 3h-6h en 0h-6h. Dans le jeu de données, chaque modalité 18-21 ou 21-25 est remplacé par 18-25, et chaque modalité 0h-3h ou 3h-6h est remplacé par 0h-6h.

7.2 Résultats R

```
             Df Sum Sq Mean Sq F value    Pr(>F)
Creneau       6 260.50   43.42  10.711 0.0001502 ***
Age           5 586.15  117.23  28.920 6.591e-07 ***
Creneau:Age  30 822.31   27.41   6.762 0.0002421 ***
Residuals    14  56.75    4.05
---
```

7.3 Interprétations

L'ANOVA2 appliquée avec les regroupements effectués grâce à l'AFC est ici pertinente car on obtient des p-valeur très faibles pour nos tests.
En effet, pour le test d'influence du Creneau sur la Vitesse des accidents, la p-valeur est de 1.510^{-4}. Cette p-valeur est très faible, on en déduit donc l'influence significative du Créneau horaire sur la Vitesse.
De même, on a des conclusions similaires pour la variable Age. En effet, la p-valeur est encore plus faible : 6.59110^{-7}, la vitesse des accidents dépend donc plus des modalités de la variable Age que des modalités de la variable Creneau. Ceci corrobore les résultats obtenus précedemment sur l'ANOVA 1 et la régression linéaire. La vitesse des accidents varie donc plus selon l'âge des victimes que selon l'horaire de l'accident
Enfin, on remarque qu'il y a une forte interaction entre l'âge des victimes et la tranche horaire correspondant aux accidents car on a la aussi une p-valeur très faible : 2.410^{-4}. Ceci est cohérent car par exemple les personnes jeunes ont tendance en général à avoir des accidents durant la nuit.

8 Bilan général de l'étude

Notre étude nous a amené à plusieurs conclusions qui vont permettre de répondre aux interrogation de la gendarmerie.

Tout d'abord, nous pouvons dire que (comme le montre la régression linéaire) la vitesse à laquelle ont lieu les accidents semble être plus élevé en fin de journée ; à l'inverse elle est plus faible en début de journée (meme inférieure à la vitesse limitée). De manière plus globale, on peut avancer le fait que la vitesse des accidents augmente durant la journée (à partir de 6h du matin jusqu'à la nuit suivante). Nous conseillons donc à la gendarmerie d'intensifier ses contrôles vers la fin de la journée, voire pendant la nuit.

De plus, cette vitesse est plus élevée lorsqu'il s'agit de jeune conducteurs. Les vitesses les plus élevées sont constatées (comme le montre l'ANOVA 1) lors d'accidents impliquant de jeunes conducteurs de 18-25 ans (97 km/h de moyenne pour cette tranche d'age). La deuxieme tranche d'age où l'on constate les vitesses les plus élevées est la tranche 25-65 ans (92,5 km/h de moyenne pour cette tranche d'age). Pour les autres tranches d'age, la vitesse n'est pas exessive puisque elle est tout juste inférieure à la limitation de 90 km/h. Nous conseillons donc à la gendarmerie d'intensifier ses controles sur la population des jeunes conducteurs, mais aussi sur la population 25-65 ans.

Par ailleurs, on a vu (grâce à l'AFC) que les contrôles effectués sur les personnes agés de 18 à 21 ans et de 21 à 25 ans doivent être les mêmes, ainsi que les contrôles effectués de minuit à 3h du matin et de 3h à 6h du matin. Il y a de plus (comme le montre l'ANOVA 2), une très forte corrélation entre l'âge des victimes et l'heure de l'accident. On a montré aussi la vitesse des accidents dépend beaucoup de l'âge des conducteurs et de l'horaire. Mais la vitesse des accidents dépend plus de l'age des conducteurs que de l'horaire comme on l'a vu grâce à l'ANOVA.

Pour conclure, nous conseillons donc à la gendarmerie d'intensifier tout d'abord ses contrôles sur les personnes âgées de 18 à 25 ans ainsi que sa campagne d'information et de prévention. Dans un deuxième temps, nous lui conseillons d'intensifier ses contrôles routiers la nuit de minuit à 6h du matin.

Cependant, nous voulons préciser la nuance avec laquelle il faut prendre en compte les conclusions précitées. En effet, les données récoltées étaient certes nombreuses, mais pour la plupart d'entre elles, la vitesse n'était pas relevée. Ceci impacte bien evidement sur la qualité des résultats obtenus,

meme si les différentes méthodes employées semblent converger. Un autre point est le nombre d'accidents relevés dans chaque créneau horaire. En effet, le nombre d'accidents relevés n'étant pas le meme pour chacun des créneaux (et ceci est a priori inévitable), cela influe notre modélisation de la vitesse des accidents. Enfin, l'étude concerne des horaires correspondant à une journée, et ce phénomène de cycle journalier peut gêner les estimations faites, par exemple lorsque l'on s'interesse au dernier créneau de la journée et au premier de la journée "suivante".