Analyse statistique d'une expérience d'étude de l'éco-conduite : Vers la conception d'un eco-index

Annexe : Etude de la mise en oeuvre des consignes d'éco-conduite

Cindie ANDRIEU et Guillaume SAINT PIERRE
Laboratoire sur les Interactions Véhicules-Infrastructure-Conducteurs

28 septembre 2010

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Etude de la mise en oeuvre des consignes d'éco-conduite par 3 méthodes (régression logistique, équations d'estimation généralisées (GEE) et modèles mixtes) : modélisation à partir des trajets complets.

L'objectif de ce chapitre est d'étudier la mise en pratique des quatre principales consignes d'éco-conduite données aux conducteurs et de déterminer, parmi ces quatre consignes, lesquelles ont été appliquées correctement. Nous étudierons ici diverses méthodes permettant de modéliser la probabilité d'être en éco-conduite en fonction des consignes mises en oeuvre. Dans un premier temps, nous modéliserons notre problématique par régression logistique, méthode que nous avons déjà étudiée dans les chapitres précédents. Cependant, les trajets effectués par un même conducteur étant corrélés, l'hypothèse d'indépendance de la régression logistique est violée. Nous étudierons donc deux méthodes permettant d'analyser les données corrélées : les équations d'estimation généralisées (GEE) et les modèles mixtes.

1.1 Régression logistique

Comme nous l'avons vu au chapitre IX, la méthode usuelle pour étudier la relation entre une variable réponse binaire et plusieurs variables explicatives est la régression logistique. Notre variable réponse est ici la variable Trajet. On introduit les notations suivantes :
Soit T_i le nombre d'observations réalisées sur le sujet i ; on note $Y_i = (Y_{i1}, \ldots, Y_{iT_i})$ le vecteur

des observations, ces observations étant supposées indépendantes. Autrement dit, on note :

$$Y_{ij} = \begin{cases} 1 & \text{Trajet économique} \\ 0 & \text{Trajet normale} \end{cases} \quad i = 1, \ldots, I; j = 1, \ldots, T_i$$

où I représente le nombre de conducteurs (i.e. 20) et T_i le nombre d'observations par conducteur (i.e. le nombre de trajets, soit 2).
La variable aléatoire Y_{ij} suit une loi de Bernoulli de paramètre p_{ij} où $p_{ij} = P(Y_{ij} = 1)$.
Le modèle logistique ordinaire s'écrit :

$$logit(p_{ij}) = log(\frac{p_{ij}}{1 - p_{ij}}) = \beta' X_{ij} \tag{1.1}$$

$$Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij}) \quad (fonction\ de\ variance) \tag{1.2}$$

où $\mu_{ij} = p_{ij}$, X_{ij} est le vecteur des variables explicatives et β est le vecteur des paramètres.
Dans ce cas, le vecteur β des paramètres est estimé par la méthode du maximum de vraisemblance. L'estimateur de β est solution des équations de vraisemblances (équations du score) obtenues en dérivant la fonction de log-vraisemblance (notée ℓ) par rapport à β :

$$U(\beta) = \frac{\partial \ell}{\partial \beta} = \sum_i D_i' \Sigma_i^{-1} (Y_i - \mu_i) = 0 \tag{1.3}$$

où $D_i = \frac{\partial \mu_i}{\partial \beta'}$ (matrice diagonale), Σ_i est la matrice de variances-covariances (matrice diagonale $\Sigma_i = diag(Var(Y_{ij}))$), Y_i est le vecteur des Y_{ij}, et μ_i est le vecteur des μ_{ij}. Les équations 1.3 sont appelées "independence estimating equations"(IEE) et sont résolues par des méthodes itératives : algorithme du scoring de Fisher, méthode itérative des moindres carrés pondérés (IWLS) ou algorithme de Newton modifié (Quasi-Newton). L'estimateur obtenu $\widehat{\beta}$ est alors consistent (i.e. $\widehat{\beta} \xrightarrow{P} \beta$) et distribué asymptotiquement selon une loi normale avec pour matrice de covariance $V(\widehat{\beta}) = (\sum_i D_i' \Sigma_i^{-1} D_i)^{-1}$.

Cependant, dans notre cas, les trajets effectués par un même conducteur sont corrélés et ne peuvent donc pas être supposés indépendants. En effet, les trajets effectués par un même conducteur ont les mêmes caractéristiques car le comportement de conduite reste similaire. Ce problème de données corrélées apparaît dans de nombreuses situations : mesures répétées dans le temps appelées données longitudinales (principalement en épidémiologie), mesures issues d'une même famille ou cluster. Dans le domaine des transports, les observations issues d'expérience de type "naturalistic driving" appartiennent clairement à cette dernière catégorie : les observations issues d'un même conducteur sont, en général, fortement corrélées. L'hypothèse d'indépendance étant violée, on ne peut pas utiliser les modèles classiques comme les modèles linéaires généralisées dont la régression logistique est un cas particulier, et qui reposent sur l'indépendance des observations.
Il existe principalement deux méthodes permettant de prendre en compte la liaison entre observations réalisées sur un même sujet :

1. **les modèles marginaux ou GEE (Generalized Estimating Equations)**
2. **les modèles mixtes**

Ces deux méthodes sont décrites dans les sections suivantes.

1.2 Equations d'estimation généralisées (GEE)

Les modèles marginaux basées sur le concept d'équations d'estimations généralisées ont été introduits par Liang et Zeger ([7]) pour analyser les données longitudinales (mesures répétées). Ils sont l'analogue des modèles de quasi-vraisemblance pour observations non indépendantes. En effet, l'approche GEE ne spécifie pas de fonction de vraisemblance (basée sur l'indépendance des observations) mais utilise une quasi-vraisemblance pour laquelle il n'existe pas de modèle probabiliste approprié. Pour les modèles de quasi-vraisemblance, il est suffisant de donner les deux premiers moments (moyenne et covariance) de la variable réponse, par opposition aux modèles linéaires généralisés dans lesquels toute la distribution de la variable réponse doit être définie. Ainsi les GEE sont plutôt considérées comme une méthode d'estimation que comme une véritable méthode de modélisation.

Nous avons vu à la section précédente que sous l'hypothèse d'indépendance des observations (i.e. $Corr(Y_{ij}, Y_{ij'}) = 0$), le modèle logistique s'écrivait sous la forme 1.1. Cependant, dans le cas de données corrélées (i.e. $Corr(Y_{ij}, Y_{ij'}) \neq 0$), on définit une matrice de variances-covariances de travail :

$$\Sigma_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} \tag{1.4}$$

où A_i est la matrice diagonale dont les éléments diagonaux sont $Var(Y_{ij})$ et $R_i(\alpha)$ est la matrice de corrélation de travail. Cette matrice de variances-covariances de travail est généralement différente de la vraie matrice de variances-covariances S_i.
Plusieurs choix sont possibles pour la structure de corrélation de travail :

- Indépendant : $Corr(Y_{ij}, Y_{ij'}) = 1$ pour tout j, j'.
  Cas où les observations réalisés sur un même sujet sont indépendantes.
- Echangeable (type "exchangeable") :

$$Corr(Y_{ij}, Y_{ij'}) = \left\{ \begin{array}{l} 1 \; si \; j = j' \\ \alpha \; si \; j \neq j' \end{array} \right\}$$

  Cas où les observations réalisés sur un même sujet sont corrélées de la même manière.
- Auto-regressive d'ordre 1 (type "AR(1)") : $Corr(Y_{ij}, Y_{ij'}) = \alpha^{|j'-j|}$.
  Utilisé pour les mesures répétées dans le temps : la corrélation diminue à mesure que l'écart entre j et j' augmente.
- Non spécifiée (type "unstructured") : $Corr(Y_{ij}, Y_{ij'}) = \alpha_{jj'}$ pour $j \neq j'$.
  Structure la plus générale mais beaucoup de paramètres à estimer. (choix possible uniquement si le nombre d'observations par sujet est constant).

L'estimateur de β est alors solution des équations de quasi-vraisemblance appelées équations d'estimation généralisées ou "generalized estimating equations" (GEE) :

$$\sum_i D_i' \Sigma_i^{-1} (Y_i - \mu_i) = 0 \tag{1.5}$$

où Σ_i est défini par 1.4. Lorsque la matrice de corrélation de travail est la matrice identité, on retrouve l'équation 1.3 obtenue dans le cas d'indépendance des observations. L'estimation de β se fait à l'aide d'un processus itératif : algorithme de Fisher modifié. Le terme "modifié" signifie que l'on utilise Σ_i à la place de la vraie matrice de variances-covariances S_i.
L'intérêt de la méthode GEE repose sur la robustesse des estimations obtenues : les inférences sur β sont corrects même si la matrice de variances-covariances Σ_i n'est pas correctement spécifiée. On utilise pour cela un esimateur "sandwich". Ainsi, sous des conditions de régularité faibles, $\widehat{\beta}$ est asymptotiquement distribué selon une loi multinormale :

$$\widehat{\beta} \xrightarrow{d} N(\beta, I_0^{-1} I_1 I_0^{-1}) \tag{1.6}$$

où $I_0 = \sum_i D_i' \Sigma_i^{-1} D_i$ et $I_1 = \sum_i D_i' \Sigma_i^{-1} S_i \Sigma_i^{-1} D_i$ évalué à $\beta = \widehat{\beta}$ et $\alpha = \widehat{\alpha}$, et en estimant S_i par $(Y_i - \mu_i(\widehat{\beta}))(Y_i - \mu_i(\widehat{\beta}))'$.
Si la matrice de variances-covariances est correctement spécifiée (i.e. $\Sigma_i = S_i$), alors $I_0 = I_1$ et on retrouve le résultat classique :

$$\widehat{\beta} \xrightarrow{d} N(\beta, I_0^{-1}) \tag{1.7}$$

I_0 étant la matrice d'information de Fisher.

1.3 Modèles mixtes

Une autre approche pour modéliser des données corrélées est l'utilisation des modèles linéaires généralisés mixtes (GLMM). La méthode vue à la section précédente modélise les distributions marginales en traitant la corrélation comme un paramètre de nuisance. La méthode présentée ici introduit un effet aléatoire spécifique à chaque sujet. On modélise ainsi la variable réponse conditionnellement à ces effets aléatoires. L'inférence est donc individuelle ("subject-specific approach"), contrairement aux modèles marginaux présentés à la section précédente où l'on modélisait la moyenne de la population ("population-averaged approach").
Le modèle logistique mixte s'écrit :

$$logit[P(y_{ij} = 1|u_i)] = x_{ij}'\beta + z_{ij}'u_i \tag{1.8}$$

où x_{ij} et β sont définis comme dans le modèle 1.1, u_i est le vecteur des effets aléatoires associés au sujet i et z_{ij} est la matrice associée.
Un cas particulier important est le cas où $z_{ij} = 1$. Ce modèle est appelé modèle logistique avec effet aléatoire sur la constante ("random-intercept logistic model" ou "logistic-normal model") et s'écrit :

$$logit[P(y_{ij} = 1|u_i)] = x_{ij}'\beta + u_i \tag{1.9}$$

où les u_i sont indépendants et $u_i \sim N(0, \sigma^2)$. Soit y le vecteur des observations et u le vecteur des effets aléatoires. La vraisemblance marginale ("marginal likelihood") s'écrit :

$$\ell(\beta, \sigma^2; y) = f(y; \beta, \sigma^2) = \int f(y|u; \beta) f(u; \sigma^2) du \tag{1.10}$$

où $f(y|u; \beta)$ est la densité conditionnelle de y sachant u et $f(u; \sigma^2)$ est la densité normale de u. En particulier, la vraisemblance marginale du modèle 1.9 s'écrit :

$$\ell(\beta, \sigma^2; y) = \int \prod_i \prod_j \left[\frac{exp(x'_{ij}\beta + u_i)}{1 + exp(x'_{ij}\beta + u_i)} \right]^{y_{ij}} \left[\frac{1}{1 + exp(x'_{ij}\beta + u_i)} \right]^{1-y_{ij}} f(u_i; \sigma^2) du_i \tag{1.11}$$

L'estimation des paramètres se fait par approximation puis maximisation de cette vraisemblance. Il existe pour cela différentes méthodes :

1. **Quadrature de Gauss-Hermite**
   Cette méthode consiste à approximer l'intégrale 1.10 par des méthodes numériques puis à maximiser cette approximation de la vraisemblance. L'approximation est réalisée par la formule de la quadrature de Gauss-Hermite :
   $$\int_{-\infty}^{\infty} f(u) exp(-u^2) du \approx \sum_{k=1}^{q} c_k f(s_k) \tag{1.12}$$
   où les c_k sont des poids et les s_k sont des points de quadrature. L'approximation est d'autant meilleure que le nombre de points de quadrature q est grand.
   Cette approximation de la vraisemblance est ensuite maximisée par des algorithmes classiques comme l'algorithme de Newton-Raphson.
2. **Quasi-vraisemblance pénalisée ou "Penalized Quasi-Likelihood" (PQL)**
   Cette méthode consiste à approximer le modèle linéaire généralisé mixte par un modèle linéaire mixte ("linearization method"). La vraisemblance marginale 1.10 résulte de l'intégration de la distribution jointe de y et u par rapport aux effets aléatoires u. En utilisant la représentation sous forme de famille exponentielle de chaque composant de cette loi jointe, l'intégrale 1.10 se ramène à une fonction exponentielle de u. Cette fonction peut alors être approximée par une série de Taylor du 2nd ordre ("Laplace approximation"). Cette approximation, appelé quasi-vraisemblance pénalisée, est maximisée en utilisant les méthodes de maximisation des modèles linéaires mixtes (REML : Restricted Maximum Likelihood).
3. **Autres méthodes**
   Enfin, on peut citer aussi les méthodes de Monte Carlo ainsi que l'approche Bayesienne.

1.4 Application : Analyse de la mise en oeuvre des consignes d'éco-conduite

L'objectif ici est de déterminer quelles sont les consignes d'éco-conduite qui ont été mises en oeuvre lors du trajet "Economique". Pour cela, chaque consigne d'éco-conduite a été associée

à un indicateur, et la variabilité de ces indicateurs en fonction du style de conduite (Normale ou Economique) a été étudiée à l'aide des trois méthodes décrites aux sections précédentes : régression logistique ordinaire, GEE et modèles mixtes.

1.4.1 Consignes d'éco-conduite et indicateurs associés

Lors de leur trajet effectué en conduite économique, il a été demandé aux conducteurs d'appliquer au mieux les quatres consignes suivantes issues du projet Ecodrive :

1. **Passer à la vitesse supérieure dès que possible.**
   Passer à la vitesse supérieure entre 1500 et 2000 tr/min.
2. **Maintenir une allure constante.**
   Enclencher la plus haute vitesse possible et conduire avec un régime moteur faible.
3. **Anticiper le trafic.**
   Regarder le plus loin possible et anticiper le trafic environnant.
4. **Décélérer progressivement.**
   S'il faut ralentir ou s'arrêter, décélérer progressivement en relâchant l'accélérateur à temps et en laissant la voiture en prise.

Afin d'étudier la mise en pratique de ces quatre consignes d'éco-conduite, chacune de ces consignes a été associée à un indicateur représentatif de celle-ci.
Ainsi la consigne 1 indique que les changements de vitesse doivent s'effectuer à un régime moteur faible (2000-2500 tr/min). Il est donc naturel de lui associer l'indicateur Avg_RPM_Shift qui représente la moyenne du régime moteur auquel s'effectue le changement de vitesse.
La consigne 2 est à la fois liée au rapport de boite (rapport élevé) et au régime moteur (régime moteur faible). Aucun des indicateurs déjà calculés ne représente à la fois ces deux notions (cf. Tableau des indicateurs, Annexe A). Il a donc été nécessaire de créer un nouvel indicateur résumant à la fois la distribution des rapports de boite et le régime moteur moyen associé à chaque rapport. Cet indicateur, appelé Index_Gear_RPM, est donné par la formule suivante :

$$Index\_Gear\_RPM = Time\_Neutral \times \frac{Avg\_RPM\_Neutral}{3500} + Gear\_1 \times \frac{Avg\_RPM\_Gear1}{3500} + Gear\_2 \times \frac{Avg\_RPM\_Gear2}{3500} + \cdots + Gear\_5 \times \frac{Avg\_RPM\_Gear5}{3500} \tag{1.13}$$

où

Time_Neutral = % de temps passé au point mort ;

Avg_RPM_Neutral = Moyenne du régime moteur au point mort ;

Gear_1 = % de temps passé en vitesse 1 ;

Avg_RPM_Gear1= = Moyenne du régime moteur en vitesse 1 et avec appui sur la pédale d'accélérateur ;

$\vdots$

Gear_5 = % de temps passé en vitesse 5 ;

Avg_RPM_Gear5= = Moyenne du régime moteur en vitesse 5 et avec appui sur la pédale d'accélérateur.

La condition d'appui sur la pédale d'accélérateur permet de ne pas tenir compte du temps passé en frein moteur (cas où la consommation de carburant est nulle). En effet, l'utilisation du frein moteur est représentée par la consigne 4. De plus, il faut noter que la division par 3500 est une normalisation, cette valeur représentant le régime moteur maximal.
La consigne 3 faisant référence à l'anticipation du trafic est associée à la variable PKE (Positive Kinetic Energy). En effet, cette variable reflète bien le maintien ou non d'une vitesse constante et sa significativité a déjà été démontrée dans les chapitres précédents.
Enfin la consigne 4 fait clairement référence à l'utilisation du frein moteur et elle est donc naturellement associée à l'indicateur Time_Engine_Brake.
Cette association entre une consigne et un indicateur est résumée dans le *tableau 1.1*.

Consigne	**Nom de l'indicateur associé**	**Description de l'indicateur**
Consigne 1	Avg_RPM_Shift	Moyenne du régime moteur auquel s'effectue le changement de vitesse
Consigne 2	Index_Gear_RPM	Indice de distribution des rapports de boite et de régimes moteurs moyens associés à ces rapports
Consigne 3	PKE	Positive Kinetic Energy
Consigne 4	Time_Engine_Brake	% de temps passé en frein moteur

TABLE 1.1 - Consignes d'éco-conduite et indicateurs associés.

1.4.2 Régression logistique ordinaire

Dans un premier temps, on utilise un modèle de régression logistique ordinaire afin de modéliser la relation entre la variable binaire Trajet (qui prend la valeur 0 si la conduite est "normale", et 1 si la conduite est "économique") et les quatre indicateurs associés à chacune des consignes (cf. tableau 1.1) : Avg_RPM_Shift, Index_Gear_RPM, PKE et Time_Engine_Brake. Cette approche suppose l'indépendance des données même si l'on a vu à la section 1.1 que cette hypothèse était fausse puisque les observations issues d'un même conducteur sont corrélées. Le modèle logistique s'écrit alors :

$$logit[P(Y_{ij} = 1)] = \beta_0 + \beta_1 \times Avg\_RPM\_Shift + \beta_2 \times Index\_Gear\_RPM + \beta_3 \times PKE + \beta_4 \times Time\_Engine\_Brake \quad (1.14)$$

Les résultats sont donnés dans les tableaux 1.2 et 1.3 et ont été obtenus à l'aide de la procédure LOGISTIC de SAS, et à partir des trajets complets, soit 40 trajets (20 en conduite normale et 20 en conduite économique). Le tableau 1.2 indique que seule la variable PKE est significative

(χ^2 = 5.80, df = 1, $p < 0.05$). Il semble donc que la consigne 3 qui consiste à anticiper le trafic et maintenir une vitesse constante soit la consigne la plus appliquée lors du trajet "Economique".

Parameter	Estimate	Standard Error	Wald Khi-2	p-value
Intercept	9.2718	8.5450	1.1774	0.2779
Avg_RPM_Shift	−0.00733	0.00489	2.2469	0.1339
Index_Gear_RPM	0.0666	0.1456	0.2090	0.6476
PKE	−37.5859	15.6079	5.7991	0.0160
Time_Engine_Brake	0.2737	0.1684	2.6401	0.1042

TABLE 1.2 - Modèle logistique ordinaire (construit à partir des 40 trajets) : Estimation des paramètres.

Ce résultat est confirmé par les valeurs des odds ratios présentés au tableau 1.3 puisque seul l'odds ratio associé à la variable PKE est significatif (l'intervalle de confiance ne contient pas la valeur 1)). Cependant cette valeur est très faible (OR < 0.001) car l'ordre de grandeur de la variable PKE est petit par rapport aux trois autres régresseurs du modèle. Même si la valeur n'est pas significative, il est surprenant que l'odds ratio associé à la variable Index_Gear_RPM soit supérieur à 1. En effet, une faible valeur de la variable Index_Gear_RPM indique normalement une conduite avec un rapport élevé et un régime moteur faible, et donc une conduite économique. La consigne 2 ne semble donc pas avoir été correctement mise en oeuvre. Enfin, on peut noter que l'odds ratio associé à la variable Time_Engine_Brake est supérieur à 1 même si la valeur n'est pas significative puisque l'intervalle de confiance contient la valeur 1. Ce résultat montre que le frein moteur a été très peu utilisé lors du trajet "Economique". En effet, nous avons vu à la section 4.2.2 du chapitre VII que le frein moteur avait été essentiellement utilisé par les moniteurs d'éco-conduite et que cette pratique n'était pas familière pour le conducteur lambda.

Parameter	Odds-Ratio	95% CI low	95% CI high
Avg_RPM_Shift	0.993	0.983	1.002
Index_Gear_RPM	1.069	0.803	1.422
PKE	<0.001	<0.001	<0.001
Time_Engine_Brake	1.315	0.945	1.829

TABLE 1.3 - Modèle logistique ordinaire (construit à partir des 40 trajets) : Odds ratio et intervalles de confiance à 95%.

1.4.3 Approche GEE

Comme nous l'avons vu à la section 1.1, les trajets effectués par un même conducteur ne sont pas indépendants. Il faut donc utiliser des méthodes prenant en compte la corrélation de nos données. On utilise dans cette section l'approche GEE pour étudier la mise en pratique des consignes d'éco-conduite. Cette approche suppose que la structure de corrélation est la même pour tous les conducteurs. Nous avons vu à la section 1.2 que différentes structures de corrélation de travail étaient possible, cependant dans notre cas, il semble naturel d'utiliser la structure échangeable puisqu'il n'y a que deux observations par conducteur.
L'estimation des paramètres est donnée dans le tableau 1.4 et a été obtenue à l'aide de la procédure GENMOD de SAS, toujours à partir des trajets complets. On observe que les valeurs estimées des paramètres sont fausses (toutes les valeurs sont proches de 0) bien que l'algorithme converge. Ces résultats aberrants sont dû à la taille de notre échantillon qui est trop faible (20 conducteurs et 2 observations par conducteur). En effet, l'utilisation des GEE nécessite que le nombre d'observations par sujet soit faible et que le nombre de sujets soit important ([4]). Ainsi Ziegler, Kastner et Blettner ([8])recommandent d'utiliser les GEE seulement si le nombre de clusters est d'au moins 30, que le nombre d'observations par cluster est de l'ordre de 4, et que la corrélation intra-cluster n'est pas trop grande (cite Ziegler, Kastner et Blettner). Dans le cas de petits échantillons, il est conseillé d'utiliser plutôt la structure de corrélation de type indépendante et donc de se ramener à une régression logistique ordinaire. Il faut tout de même noter que certaines méthodes ont été développées comme le bootstrap pour palier au problème des petits échantillons ([8]).
Cependant, les résultats des tests d'hypothèse sur la nullité de chaque coefficient semblent corrects. Les Z-statistiques et les p-value associées sont données dans le tableau 1.4, alors que les statistiques du Score sont données dans le tableau 1.5. Les deux résultats sont similaires et indiquent que seule la variable PKE est significative, confirmant les résultats obtenus à la section précédente.
Les odds-ratios étant obtenus en prenant l'exponentielle des valeurs estimées des paramètres, on ne peut rien en dire puisque nos estimations sont incorrects. De même l'estimation de la corrélation intra-cluster est fausse. Notre échantillon étant trop petit, on ne peut pas utiliser l'approche GEE pour notre étude, sauf pour l'analyse de la significativité des paramètres.

Parameter	Estimate	Standard Error	Z	Pr > |Z|
Intercept	0.0003	0.0008	0.35	0.7290
Avg_RPM_Shift	−0.0000	0.0000	−0.87	0.3857
Index_Gear_RPM	0.0000	0.0000	0.12	0.9014
PKE	−0.0024	0.0008	−3.17	0.0015
Time_Engine_Brake	0.0000	0.0000	1.28	0.2018

TABLE 1.4 - Approche GEE (construit à partir des 40 trajets) : Estimation des paramètres.

Parameter	Khi-2	p-value
Avg_RPM_Shift	0.77	0.3795
Index_Gear_RPM	0.02	0.9024
PKE	10.49	0.0012
Time_Engine_Brake	1.49	0.2228

TABLE 1.5 - Approche GEE (construit à partir des 40 trajets) : Test du Score.

1.4.4 Modèle mixte

Nous avons vu précédemment que l'analyse des données corrélées pouvait aussi être effectuée avec les modèles mixtes. Afin de tenir compte de la corrélation entre les trajets effectués par un même conducteur, on utilise le modèle logistique mixte suivant :

$$logit[P(y_{ij} = 1|u_{0i})] = (\beta_0 + u_{0i}) + \beta_1 \times Avg\_RPM\_Shift + \beta_2 \times Index\_Gear\_RPM + \beta_3 \times PKE + \beta_4 \times Time\_Engine\_Brake \quad (1.15)$$

où $u_{0i} \sim N(0, \sigma^2)$ représente l'effet aléatoire associé au conducteur i. Contrairement au modèle logistique 1.14, l'ajout d'un effet aléatoire sur la constante β_0 permet de prendre en compte la corrélation entre observations issues d'un même conducteur. Ainsi la valeur de u_{0i} est liée au style de conduite habituel du conducteur (conduite nerveuse ou conduite économique) : plus u_{0i} est grand, plus la probabilité que le conducteur i ait une conduite économique est grande, et inversement. Ce modèle avec effet aléatoire sur la constante est couramment appelé "random-intercept logistic model".
Il existe deux procédures SAS pour l'analyse des modèles mixtes : la procédure NLMIXED et la procédure GLIMMIX. La première utilise la méthode de quadrature de Gauss-Hermite pour l'estimation des paramètres, alors que la seconde utilise la quasi-vraisemblance pénalisée ([2],[1]). Pour notre étude, nous avons utilisé la procédure GLIMMIX qui est plus appropriée dans le cas de petits échantillons ([3]). Les résultats obtenus sont données dans les tableaux 1.6, 1.7 et 1.8. L'estimation de la variance σ^2 de la distribution des u_{0i} est de 1.741 avec un écart-type de 2.496, ce qui indique une grande variabilité entre les conducteurs. Les tableaux 1.6 et 1.7 montrent que seule la variable PKE est significative (p-value < 0.05), confirmant une nouvelle fois les résultats obtenus à la section 1.4.2. On obtient dans le tableau 1.8 des odds-ratios similaires à ceux obtenus à la section 1.4.2.

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	11.1558	9.9855	1.12	0.2778
Avg_RPM_Shift	−0.00887	0.005754	−1.54	0.1427
Index_Gear_RPM	0.07523	0.1702	0.44	0.6644
PKE	−44.0742	18.3645	−2.40	0.0289
Time_Engine_Brake	0.3292	0.2011	1.64	0.1211

TABLE 1.6 - Random-intercept logistic model (construit à partir des 40 trajets) : Estimation des paramètres.

Parameter	F Value	p-value
Avg_RPM_Shift	2.38	0.1427
Index_Gear_RPM	0.20	0.6644
PKE	5.76	0.0289
Time_Engine_Brake	2.68	0.1211

TABLE 1.7 - Random-intercept logistic model (construit à partir des 40 trajets) : Tests de nullité des effets fixes.

Parameter	Odds Ratio	95% CI low	95% CI high
Avg_RPM_Shift	0.991	0.979	1.003
Index_Gear_RPM	1.078	0.752	1.547
PKE	<0.001	<0.001	0.006
Time_Engine_Brake	1.390	0.908	2.129

TABLE 1.8 - Random-intercept logistic model (construit à partir des 40 trajets) : Odds ratio et intervalles de confiance à 95%.

1.5 Conclusion

Nous avons étudié la mise en pratique des consignes d'éco-conduite à l'aide de trois méthodes : la régression logistique ordinaire, les équations d'estimation généralisées, et les mo-

dèles mixtes. Nous avons obtenu des résultats similaires avec ces trois méthodes, à savoir que seule la variable PKE était significative. Il semble donc que la consigne 3 qui consiste à anticiper le trafic et à maintenir une vitesse constante soit la seule consigne qui ait été appliquée. Cependant cette étude ayant été faite à partir des trajets complets, l'échantillon est relativement petit (20 conducteurs et 2 observations par conducteur). Les résultats obtenus doivent donc être interprétés avec précaution. Nous avons d'ailleurs constaté les limites de la taille de notre échantillon avec la méthode GEE qui n'a pas pu être appliquée correctement. Enfin, il semble que la régression logistique ne soit pas appropriée pour notre étude car nos données sont corrélées (les trajets effectués par un même conducteur ne peuvent pas être considérés comme indépendants). Les méthodes basées sur les modèles marginaux (GEE) et les modèles mixtes semblent donc plus adaptées à notre étude.

Chapitre 2

Etude de la mise en oeuvre des consignes d'éco-conduite par 3 méthodes (régression logistique, équations d'estimation généralisées (GEE) et modèles mixtes) : modélisation à partir des sections de trajets.

L'objectif de ce chapitre est, comme au chapitre précédent, de déterminer parmi les quatre principales consignes d'éco-conduite, lesquelles sont les plus représentatives d'une conduite économique. Cependant nous avons vu au chapitre précédent que l'approche GEE ne donnait pas de résultats corrects car notre échantillon était trop petit (40 trajets). Nous avons donc décomposé nos trajets en sections correspondant aux limitations de vitesse rencontrées. Nous avons ainsi augmenté la taille de notre échantillon et pu appliquer les différentes méthodes exposées au chapitre précédent : la régression logistique, l'approche GEE et les modèles mixtes.

2.1 Décomposition des trajets en sections correspondant à une limitation de vitesse

Comme au chapitre précédent, on cherche à étudier la mise en oeuvre des quatre consignes d'éco-conduite. Pour celà, on modélise la relation entre la variable binaire Trajet (qui prend la valeur 0 si la conduite est "normale", et 1 si la conduite est "économique") et les quatre indicateurs associés à chacune des consignes (cf. tableau 1.1) : Avg_RPM_Shift, Index_Gear_RPM, PKE et Time_Engine_Brake. Cependant nous avons vu au chapitre précédent, en étudiant les trajets

complets, que notre échantillon était trop petit pour utiliser certaines méthodes. Pour palier à ce problème, nous avons décomposé nos 40 trajets par limitation de vitesse puis nous avons recalculé les quatre indicateurs (Avg_RPM_Shift, Index_Gear_RPM, PKE et Time_Engine_Brake) pour chaque section. Ainsi chaque section correspond à une limitation de vitesse et chaque changement de limitation de vitesse correspond à la fin d'une section et au début d'une autre. Cette décomposition de chaque trajet ("Normale" et "Eco") en sections est illustrée par le tableau 2.1. On obtient ainsi un échantillon total de 600 sections (15 sections pour chacun des 40 trajets). Le nombre de conducteurs est toujours le même (20 conducteurs), mais cette décomposition permet d'augmenter le nombre de trajets effectué par chaque conducteur : 30 trajets (ou sections) par conducteur au lieu de 2. Il faut noter que des valeurs manquantes ont été générées pour la variable Avg_RPM_Shift suite à cette décomposition correspondant à des sections durant lesquelles aucun changement de rapport de boite n'a été effectué. Le nombre d'observations réellement utilisé est donc de 500 au lieu de 600 (100 observations avec des valeurs manquantes n'ont pas été utilisées).

Numéro de section	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Limitation de vitesse	50	30	70	90	70	50	90	70	90	70	50	30	60	45	50
Distance (en m)	944	415	1488	538	1430	1209	1259	532	656	854	1048	464	729	1069	2075

TABLE 2.1 - Décompostion de chaque trajet.

2.2 Régression logistique

Nous modélisons dans un premier temps la relation entre la variable Trajet et les quatre indicateurs associés aux consignes d'éco-conduite à l'aide d'une régression logistique ordinaire, même si nous avons vu que l'hypothèse d'indépendance des observations n'était pas vérifiée. Le modèle est le même que celui établi au chapitre précédent (modèle 1.14). Les résultats obtenus à l'aide de la procédure LOGISTIC de SAS sont donnés aux tableaux 2.2 et 2.3. Contrairement au chapitre précédent, on obtient ici que trois paramètres sont significatifs : Index_Gear_RPM, PKE (p-value < 0.001) et dans une moindre mesure Time_Engine_Brake (p-value $= 0.0058$). Ainsi seule la consigne 1 qui consiste à passer à la vitesse supérieure à un régime moteur peu élevé n'a pas été appliquée, et ceci quel que soit la limitation de vitesse.

Parameter	Estimate	Standard Error	Wald Khi-2	p-value
Intercept	10.7866	1.1235	92.1837	<.0001
Avg_RPM_Shift	−0.00004	0.000467	0.0059	0.9385
Index_Gear_RPM	−0.1482	0.0214	48.1477	<.0001
PKE	−6.1255	0.8843	47.9808	<.0001
Time_Engine_Brake	−0.0186	0.00674	7.5960	0.0058

TABLE 2.2 - Modèle logistique ordinaire (construit à partir des 500 sections) : Estimation des paramètres.

Ces résultats sont confirmés par les valeurs des odds ratios présentées au tableau 2.3 : les odds ratios associés aux variables Index_Gear_RPM, PKE et Time_Engine_Brake sont tous significatifs (les intervalles de confiance contiennent la valeur 1). Ainsi, si la variable Index_Gear_RPM augmente de 1 unité, alors la probabilité d'être en éco-conduite baisse de 13.8% (OR = 0.862). Par contre, la valeur de l'odds ratio associé au frein moteur est surprenante puisqu'on obtient un odds ratio inférieur à 1. En effet, on obtient une valeur de 0.982, ce qui signifie que si la valeur de la variable Time_Engine_Brake augmente de 1%, alors la probabilité d'être en éco-conduite baisse de 1.8% (OR = 0.982). Ce résultat montre que les conducteurs n'ont pas utilisé correctement le frein moteur. En ce qui concerne la variable PKE, une augmentation de la valeur de la variable PKE de 1 unité entraîne une baisse de la probabilité d'être en éco-conduite, la valeur de l'odds ratio associée étant particulièrement petite (OR = 0.002) à cause de l'ordre de grandeur de cette variable. En effet, une modification de 1 unité de la variable PKE représente un changement de conduite radical.

Parameter	Odds-Ratio	95% CI low	95% CI high
Avg_RPM_Shift	1.000	0.999	1.001
Index_Gear_RPM	0.862	0.827	0.899
PKE	0.002	<0.001	0.012
Time_Engine_Brake	0.982	0.969	0.995

TABLE 2.3 - Modèle logistique ordinaire (construit à partir des 500 sections) : Odds ratios et intervalles de confiance à 95%.

2.3 Approche GEE

Nous utilisons ici les GEE qui permettent de prendre en compte la corrélation entre les sections effectuées par un même conducteur. Comme au chapitre précédent, nous supposons que la

structure de corrélation est de type échangeable, c'est-à-dire que la corrélation entre les sections effectués par un même conducteur est la même, et ceci pour tous les conducteurs. La structure de corrélation de type échangeable est la plus appropriée lorsqu'il y a peu d'individus comme c'est le cas ici puisqu'il n'y a qu'un seul paramètre à estimer, contrairement à la structure de type "unstructured" qui est plus générale mais plus compliquée (beaucoup de paramètres à estimer). Les résultats obtenus à l'aide de la procédure GENMOD de SAS sont données aux tableaux 2.4 et 2.5. Contrairement au chapitre précédent, la méthode GEE fonctionne correctement. La valeur de la corrélation entre les observations issues d'un même conducteur est estimée à 0.15, ce qui confirme que l'hypothèse d'indépendance de la régression logistique est violée. Les p-value associées aux Z-statistiques et aux statistiques du Score indiquent que tous les indicateurs sont significatifs sauf Avg_RPM_Shift.

Parameter	Estimate	Standard Error	Z	Pr > |Z|
Intercept	17.0463	2.6462	6.44	<.0001
Avg_RPM_Shift	0.0003	0.0005	0.57	0.5715
Index_Gear_RPM	−0.2550	0.0504	−5.06	<.0001
PKE	−8.2553	1.3993	−5.90	<.0001
Time_Engine_Brake	−0.0305	0.0074	−4.14	<.0001

TABLE 2.4 - Approche GEE (construit à partir des 500 sections) : Estimation des paramètres.

Parameter	Khi-2	p-value
Avg_RPM_Shift	0.46	0.4979
Index_Gear_RPM	13.62	0.0002
PKE	14.57	0.0001
Time_Engine_Brake	13.12	0.0003

TABLE 2.5 - Approche GEE (construit à partir des 500 sections) : Test du Score.

La procedure GENMOD de SAS ne calcule pas les odds-ratios lorsque les variables régresseurs du modèle sont quantitatives. Nous avons donc calculé les odds ratios en prenant l'exponentielle des valeurs estimées des paramètres. Les intervalles de confiance à 95% ont été calculés par la méthode de Miettinen dont la formule est :

$$exp[log(OR) \times (1 \pm 1.96/\sqrt{\chi^2})]$$

où χ^2 est la valeur du Khi-deux de Wald.
Les résultats donnés au tableau 2.6 montrent que les odds ratios associés aux variables Index_Gear_RPM, PKE et Time_Engine_Brake sont tous significatifs. Si la valeur de la variable Index_Gear_RPM augmente de 1 unité, alors la probabilité d'être en éco-conduite baisse de 22.5% (OR = 0.7749), et si la valeur de la variable Time_Engine_Brake augmente de 1%, alors la probabilité d'être en éco-conduite baisse de 3% (OR = 0.9699). Ainsi les résultats sont semblables à ceux obtenus à la section 2.2.

Parameter	Odds Ratios	95% CI low	95% CI high
Avg_RPM_Shift	1.000	0.999	1.001
Index_Gear_RPM	0.775	0.702	0.855
PKE	<0.001	<0.001	0.004
Time_Engine_Brake	0.970	0.956	0.984

TABLE 2.6 - Approche GEE (construit à partir des 500 sections) : Odds ratios et intervalles de confiance à 95%.

2.4 Modèle mixte

Dans cette section nous modélisons la corrélation entre les observations issues d'un même conducteur à l'aide d'un modèle mixte. Le modèle utilisé est le même qu'au chapitre précédent (modèle 3.1) avec uniquement l'ajout d'un effet aléatoire sur la constante. Les résultats obtenus à l'aide de la procédure GLIMMIX de SAS sont donnés aux tableaux 2.7, 2.8 et 2.9. L'estimation de la variance σ^2 de la distribution des u_{0i} est de 1.483 avec un écart-type de 0.661, ce qui indique une grande variabilité entre les conducteurs. Comme aux deux sections précédentes, on obtient que tous les indicateurs sont significatifs sauf l'indicateur Avg_RPM_Shift (cf. tableaux 2.7 et 2.8). De même les valeurs estimées des odds ratios sont similaires à celles obtenues avec la régression logistique et les GEE (cf. tableau 2.9).

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	16.4856	1.5220	10.83	<.0001
Avg_RPM_Shift	−0.00009	0.000536	−0.17	0.8624
Index_Gear_RPM	−0.2286	0.02677	−8.54	<.0001
PKE	−8.2764	1.0539	−7.85	<.0001
Time_Engine_Brake	−0.02996	0.007526	−3.98	<.0001

TABLE 2.7 - Random-intercept logistic model (construit à partir des 500 sections) : Estimation des paramètres.

Parameter	F Value	p-value
Avg_RPM_Shift	0.03	0.8624
Index_Gear_RPM	72.87	<.0001
PKE	61.67	<.0001
Time_Engine_Brake	15.84	<.0001

TABLE 2.8 - Random-intercept logistic model (construit à partir des 500 sections) : Tests de nullité des effets fixes.

Parameter	Odds Ratio	95% CI low	95% CI high
Avg_RPM_Shift	1.000	0.999	1.001
Index_Gear_RPM	0.796	0.755	0.839
PKE	<0.001	<0.001	0.002
Time_Engine_Brake	0.970	0.956	0.985

TABLE 2.9 - Random-intercept logistic model (construit à partir des 500 sections) : Odds ratio et intervalles de confiance à 95%.

2.5 Comparaison des 3 méthodes

L'objectif de cette section est de comparer les résultats obtenus à partir des trois méthodes décrites précédemment : la régression logistique, le modèle GEE et le modèle mixte. Le tableau

2.10 récapitule les valeurs estimées des paramètres obtenues avec chacune de ces 3 méthodes. Les valeurs absolues des estimations des paramètres ainsi que les écart-types associés obtenus avec la régression logistique ont les plus petites valeurs. Cependant, d'après la littérature ([6]), les écart-types issus de la régression logistique ordinaire sont biaisés lorsque les données sont corrélées. Ceci explique le fait que dans notre cas les valeurs des écart-types soient sous-estimées. En ce qui concerne le modèle GEE et le modèle mixte, les estimations des paramètres sont relativement proches. Les écart-types sont également assez proches sauf pour les variables Index_Gear_RPM et PKE, qui rappelons-le sont les plus significatives, pour lesquelles le modèle mixte donne des valeurs plus faibles. Il semble donc que les valeurs estimées des paramètres obtenues avec le modèle mixte soient plus précises.

	Logistic Model		GEE Model		Random Effect Model	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
Avg_RPM_Shift	-0.00004	0.000467	0.0003	0.0005	-0.00009	0.000536
Index_Gear_RPM	-0.1482	0.0214	-0.2550	0.0504	-0.2286	0.02677
PKE	-6.1255	0.8843	-8.2553	1.3993	-8.2764	1.0539
Time_Engine_Brake	-0.0186	0.00674	-0.0305	0.0074	-0.02996	0.007526

TABLE 2.10 - Estimations des paramètres obtenues avec les 3 méthodes : régression logistique, modèle GEE et modèle mixte.

Afin de comparer les trois méthodes, les valeurs des odds ratios ainsi que les intervalles de confiance à 95% associés sont données au tableau 2.11 et illustrées par la figure 2.1. Les odds ratios obtenus par régression logistique ont les plus grandes valeurs. Il semble que comme pour les estimations des paramètres, les valeurs des odds ratios obtenus par régression logistique, qui rappelons-le correspondent à l'exponentielle des valeurs estimées des paramètres, soient biaisées. Dans notre cas, les valeurs sont sur-estimées. Les odds ratios obtenus avec le modèle GEE et le modèle mixte ont des valeurs similaires. On peut toutefois noter que pour la variable Index_Gear_RPM, la longueur de l'intervalle de confiance à 95% de l'odds ratio obtenu avec le modèle mixte est plus petite que celle de l'intervalle de confiance à 95% de l'odds ratio obtenu avec le modèle GEE. Ainsi les odds ratios obtenus avec le modèle mixte semblent les plus fiables.

	Logistic Model			GEE Model			Random Effect Model		
Parameter	Odds Ratio	95% CI low	95% CI high	Odds Ratio	95% CI low	95% CI high	Odds Ratio	95% CI low	95% CI high
Avg_RPM_Shift	1.000	0.999	1.001	1.000	0.999	1.001	1.000	0.999	1.001
Index_Gear_RPM	0.862	0.827	0.899	0.775	0.702	0.855	0.796	0.755	0.839
PKE	0.002	<0.001	0.012	<0.001	<0.001	0.004	<0.001	<0.001	0.002
Time_Engine_Brake	0.982	0.969	0.995	0.970	0.956	0.984	0.970	0.956	0.985

TABLE 2.11 - Odds ratios obtenus avec les 3 méthodes : régression logistique, modèle GEE et modèle mixte.

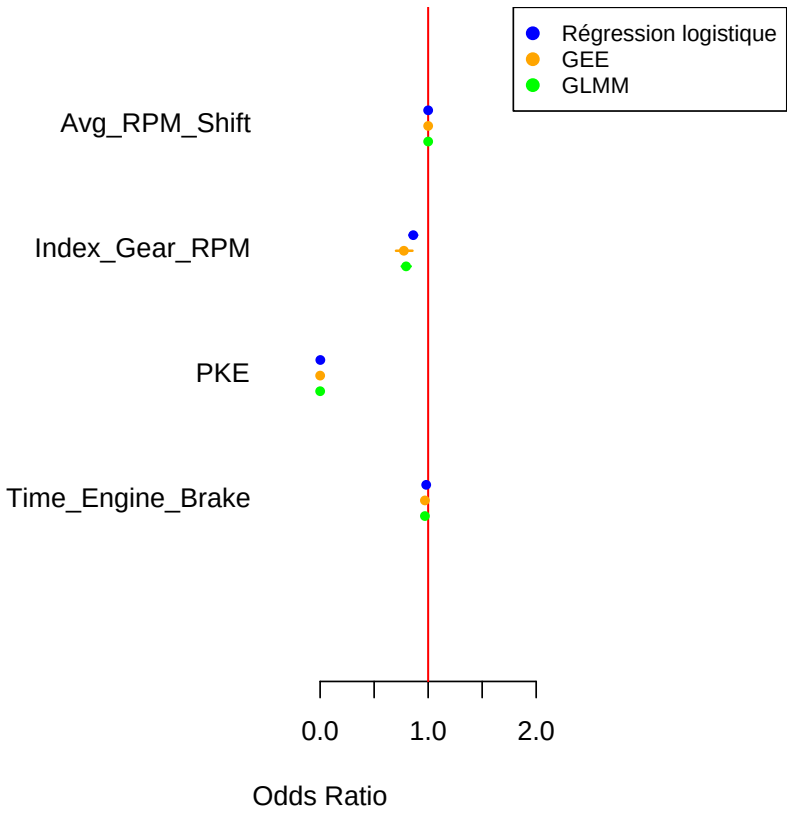

FIGURE 2.1 - Odds ratios obtenus avec les 3 méthodes : régression logistique, modèle GEE et modèle mixte.

2.6 Conclusion

La décomposition des trajets en sections nous a permis d'augmenter la taille de notre échantillon (plus d'observations par conducteur). Ainsi, contrairement au chapitre précédent, la méthode GEE a pu être appliquée correctement. Les résultats obtenus à l'aide des trois méthodes (régression logistique, modèle GEE et modèle mixte) sont similaires et indiquent que tous les paramètres du modèle sont significatifs sauf Avg_RPM_Shift. Il semble donc que seule la consigne 1, qui consiste à passer à la vitesse supérieure à un régime moteur peu élevé, n'ait pas été mise en oeuvre. Il faut toutefois noter que l'odds ratio associé à la variable Time_Engine_Brake est inférieur à 1, et ceci quel que soit la méthode utilisée, ce qui montre que le frein moteur n'a pas été utilisé correctement par les conducteurs.

Concernant les trois méthodes utilisées, il semble que le modèle mixte soit le modèle le plus approprié pour notre étude. D'une part, le modèle mixte tient compte de la corrélation entre les observations issues d'un même conducteur, contrairement à la régression logistique ordinaire. D'autre part, le fait d'introduire un effet aléatoire spécifique à chaque sujet permet de prendre en compte les différences de comportement de conduite entre les conducteurs. Ainsi, l'inférence est individuelle ("suject-specific approach"), contrairement au modèle GEE qui modélise la moyenne de la population ("population-averaged approach"). Enfin, à la section 2.5, nous avons vu que les estimations des paramètres et des odds ratios obtenues avec le modèle mixte étaient meilleures (écart-types faibles et intervalles de confiance petits).

Chapitre 3

Etude de la mise en oeuvre des consignes d'éco-conduite en fonction de la limitation de vitesse : utilisation des modèles mixtes

L'objectif de ce chapitre est d'étudier la mise en oeuvre des consignes d'éco-conduite pour chaque limitation de vitesse. Au chapitre précédent, nous avons décomposé les trajets en sections correspondant à une limitation de vitesse afin d'augmenter la taille de notre échantillon. D'une part, cette décomposition nous a permis de comparer les trois méthodes que sont la régression logistique, les GEE et les modèles mixtes. D'autre part, cette décomposition des trajets va nous permettre également de faire une étude plus fine en fonction de la consigne de limitation de vitesse. Nous avons vu au chapitre précédent que les modèles mixtes était la méthode la plus appropriée pour notre étude, c'est pourquoi nous utiliserons cette méthode ici.

3.1 Etude de la mise en oeuvre des consignes d'éco-conduite : zones 30km/h

On modélise ici la relation entre la variable Trajet et les quatre indicateurs associés aux consignes d'éco-conduite uniquement pour les sections dont la vitesse de consigne est 30km/h. D'après le tableau 2.1 du chapitre précédent, chaque trajet comporte deux sections limités à 30km/h. Notre jeu de données contient donc 80 observations : 20 conducteurs et 4 observations par conducteurs (2 sections limités à 30km/h pour le trajet "Normal" et 2 sections limités à 30km/h pour le trajet "Economique"). Le modèle utilisé est le même qu'aux deux chapitres précédent (modèle 3.1) :

$$logit[P(y_{ij} = 1|u_{0i})] = (\beta_0 + u_{0i}) + \beta_1 \times Avg\_RPM\_Shift + \beta_2 \times Index\_Gear\_RPM$$
$$+ \beta_3 \times PKE + \beta_4 \times Time\_Engine\_Brake \tag{3.1}$$

où $u_{0i} \sim N(0, \sigma^2)$ représente l'effet aléatoire associé au conducteur i.
Les résultats obtenus à l'aide de la procédure GLIMMIX de SAS sont données aux tableaux 3.1, 3.2 et 3.3. Cependant, on obtient une erreur de convergence dans l'estimation de la variance σ^2 de la distribution des u_{0i}. Cette erreur indique que la variation entre les conducteurs est très faible sur les zones limitées à 30km/h et que les observations peuvent donc être supposées comme indépendantes. En effet, avec la procedure LOGISTIC de SAS (régression logistique ordinaire), on obtient les mêmes estimations des paramètres et les mêmes écart-types que ceux donnés au tableau 3.1. Les tableaux 3.1 et 3.2 montrent que seule la variable PKE est significative (p-value < 0.05). On peut donc en conclure que la consigne 3 qui consiste à anticiper le trafic et à maintenir une vitesse constante est la plus mise en oeuvre sur les zones limitées à 30km/h.

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	9.4157	3.4880	2.70	0.0142
Avg_RPM_Shift	−0.00032	0.001704	−0.19	0.8505
Index_Gear_RPM	−0.1192	0.06806	−1.75	0.0873
PKE	−8.6242	3.0567	−2.82	0.0073
Time_Engine_Brake	0.02920	0.04820	0.61	0.5480

TABLE 3.1 - Zones limitées à 30km/h : Estimation des paramètres.

Parameter	F Value	p-value
Avg_RPM_Shift	0.04	0.8505
Index_Gear_RPM	3.07	0.0873
PKE	7.96	0.0073
Time_Engine_Brake	0.37	0.5480

TABLE 3.2 - Zones limitées à 30km/h : Tests de nullité des effets fixes.

Les estimations des odds ratios (tableau 3.3) confirment ces résultats puisque seul l'odds ratio associé à la variable PKE est significatif. On peut toutefois noter, même si les valeurs ne sont pas significatives, que si la variable Index_Gear_RPM augmente, alors la probabilité d'être en éco-conduite baisse (OR < 1), et si la variable Time_Engine_Brake augmente, alors la probabilité d'être en éco-conduite augmente (OR > 1). Les consignes 2 et 4 semblent donc avoir été peu appliquées sur les zones limitées à 30km/h.

Parameter	Odds Ratio	95% CI low	95% CI high
Avg_RPM_Shift	1.000	0.996	1.003
Index_Gear_RPM	0.888	0.774	1.018
PKE	<0.001	<0.001	0.086
Time_Engine_Brake	1.030	0.934	1.135

TABLE 3.3 - Zones limitées à 30km/h : Odds ratio et intervalles de confiance à 95%.

3.2 Etude de la mise en oeuvre des consignes d'éco-conduite : zones 50km/h

On s'intéresse ici aux sections dont la vitesse de consigne est de 50km/h. D'après le tableau 2.1 du chapitre précédent, chaque trajet comporte quatre sections limités à 50km/h. Notre jeu de données contient donc 160 observations : 20 conducteurs et 8 observations par conducteurs (4 sections limités à 50km/h pour le trajet "Normal" et 4 sections limités à 50km/h pour le trajet "Economique"). Le modèle mixte utilisé est le même que que celui de la section précédente et les résultats ont été obtenus à l'aide de la procédure GLIMMIX de SAS. Contrairement à la section précédente, il n'y a pas d'erreur de convergence ; on obtient pour l'effet aléatoire une variance σ^2 de 1.7383 avec un écart-type de 1.2782, ce qui indique une grande variabilité entre les conducteurs sur les zones limitées à 50km/h. Les tableaux 3.4 et 3.5 montrent que tous les indicateurs sont significatifs (p-value < 0.05) sauf la variable Avg_RPM_Shift. Il semble donc que, sur les zones limitées à 50km/h, toutes les consignes aient été appliquées sauf la consigne 1 qui consiste à passer à la vitesse supérieure dès que possible. En effet, en ville, il n'est pas naturel pour la plupart des conducteurs d'utiliser un rapport de boîte élevé. Même si nous avons vu à la section 4.1.3 du chapitre VII que la plupart des conducteurs utilisaient davantage la 5ème vitesse lors de leur trajet économique, les passages de rapport de boîte se font encore à un régime moteur relativement élevé.

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	20.7515	3.3206	6.25	<.0001
Avg_RPM_Shift	−0.00149	0.001433	−1.04	0.3004
Index_Gear_RPM	−0.2164	0.04733	−4.57	<.0001
PKE	−12.9016	2.5439	−5.07	<.0001
Time_Engine_Brake	−0.05718	0.01583	−3.61	0.0004

TABLE 3.4 - Zones limitées à 50km/h : Estimation des paramètres.

Parameter	F Value	p-value
Avg_RPM_Shift	1.08	0.3004
Index_Gear_RPM	20.91	<.0001
PKE	25.72	<.0001
Time_Engine_Brake	13.05	0.0004

TABLE 3.5 - Zones limitées à 50km/h : Tests de nullité des effets fixes.

Les odds ratios présentés au tableau 3.6 confirme ces résultats, puisque les odds ratios associés aux indicateurs Index_Gear_RPM, PKE et Time_Engine_Brake sont tous significatifs (les intervalles de confiance ne contiennent pas la valeur 1). Si la variable Index_Gear_RPM augmente de 1 unité, la probabilité d'être en éco-conduite baisse de 19.5% (OR = 0.805). Par contre, l'odds ratio associé au temps passé en frein moteur est inférieur à 1, ce qui signifie que plus on utilise le frein moteur, plus la probabilité d'être en éco-conduite est faible. Ce résultat montre que les conducteurs ont eu des difficultés à utiliser le frein moteur en ville, et que la consigne 4 n'a pas été correctement mis en oeuvre.

Parameter	Odds Ratio	95% CI low	95% CI high
Avg_RPM_Shift	0.999	0.996	1.001
Index_Gear_RPM	0.805	0.733	0.884
PKE	<0.001	<0.001	<0.001
Time_Engine_Brake	0.944	0.915	0.974

TABLE 3.6 - Zones limitées à 50km/h : Odds ratio et intervalles de confiance à 95%.

3.3 Etude de la mise en oeuvre des consignes d'éco-conduite : zones 70km/h

On s'intéresse ici aux sections dont la vitesse de consigne est de 70km/h. D'après le tableau 2.1 du chapitre précédent, chaque trajet comporte quatre sections limités à 70km/h. Notre jeu de données contient donc 160 observations : 20 conducteurs et 8 observations par conducteurs (4 sections limités à 70km/h pour le trajet "Normal" et 4 sections limités à 70km/h pour le trajet "Economique"). Le modèle mixte utilisé est toujours le même et les résultats ont été obtenus à l'aide de la procédure GLIMMIX de SAS. La variance σ^2 de la distribution des u_{0i} est de 0.8553 avec un écart-type de 0.8618, ce qui indique une grande variabilité entre les conducteurs sur les zones limitées à 70km/h, bien que cette variabilité soit plus faible que sur les zones limitées à 50km/h. Les tableaux 3.7 et 3.8 indiquent que seules les variables Index_Gear_RPM et PKE sont significatives. Il semble donc que les consignes 1 et 4 n'aient pas été appliquées sur les zones limitées à 70km/h.

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	15.3210	3.4920	4.39	0.0003
Avg_RPM_Shift	−0.00058	0.000931	−0.62	0.5364
Index_Gear_RPM	−0.2129	0.05137	−4.14	<.0001
PKE	−11.0842	2.8644	−3.87	0.0002
Time_Engine_Brake	0.02032	0.02407	0.84	0.4002

TABLE 3.7 - Zones limitées à 70km/h : Estimation des paramètres.

Parameter	F Value	p-value
Avg_RPM_Shift	0.38	0.5364
Index_Gear_RPM	17.18	<.0001
PKE	14.97	0.0002
Time_Engine_Brake	0.71	0.4002

TABLE 3.8 - Zones limitées à 70km/h : Tests de nullité des effets fixes.

Les valeurs des odds ratios données au tableau 3.9 confirment ces résultats puisque seuls les odds ratios des variables Index_Gear_RPM et PKE sont significatifs. On peut toutefois noter que l'odds ratio associé à la variable Time_Engine_Brake est supérieur à 1, même s'il n'est pas significatif, ce qui montre que les conducteurs ont peu utilisé le frein moteur sur les zones limitées à 70km/h.

Parameter	Odds Ratio	95% CI low	95% CI high
Avg_RPM_Shift	0.999	0.998	1.001
Index_Gear_RPM	0.808	0.730	0.895
PKE	<0.001	<0.001	0.004
Time_Engine_Brake	1.021	0.973	1.070

TABLE 3.9 - Zones limitées à 70km/h : Odds ratio et intervalles de confiance à 95%.

3.4 Etude de la mise en oeuvre des consignes d'éco-conduite : zones 90km/h

On s'intéresse ici aux sections dont la vitesse de consigne est de 90km/h. D'après le tableau 2.1 du chapitre précédent, chaque trajet comporte trois sections limités à 90km/h. Notre jeu de données contient donc 120 observations : 20 conducteurs et 6 observations par conducteurs (3 sections limités à 90km/h pour le trajet "Normal" et 3 sections limités à 90km/h pour le trajet "Economique"). Le modèle mixte utilisé est toujours le même et les résultats ont été obtenus à l'aide de la procédure GLIMMIX de SAS. Comme pour les zones limitées à 30km/h, on obtient une erreur de convergence dans l'estimation de la variance σ^2 de la distribution des u_{0i}. Cette erreur indique que la variation entre les conducteurs est très faible sur les zones limitées à 90km/h et que les observations peuvent donc être supposées comme indépendantes. En effet, avec la procedure LOGISTIC de SAS (régression logistique ordinaire), on obtient les mêmes

estimations des paramètres et les mêmes écart-types que ceux donnés au tableau 3.10. Les résultats présentés aux tableaux 3.10 et 3.11 montrent qu'aucun des indicateurs n'est significatif. Il semble donc que les quatre consignes d'éco-conduite n'aient pas été réellement mises en oeuvre sur les zones limitées à 90km/h. Ce résultat peut tout de même s'expliquer par le fait que notre trajet d'étude favorise plutôt une conduite économique sur les zones limitées à 90km/h, même lors du trajet "Normal" : trafic fluide, grandes descentes, ... Il n'y a donc pas de différence de style de conduite significative entre le trajet "Normal" et le trajet "Economique".

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	9.5705	2.8771	3.33	0.0035
Avg_RPM_Shift	−0.00228	0.001807	−1.26	0.2141
Index_Gear_RPM	−0.07398	0.06918	−1.07	0.2908
PKE	−2.5514	3.1701	−0.80	0.4254
Time_Engine_Brake	0.02482	0.04267	0.58	0.5637

TABLE 3.10 - Zones limitées à 90km/h : Estimation des paramètres.

Parameter	F Value	p-value
Avg_RPM_Shift	1.59	0.2141
Index_Gear_RPM	1.14	0.2908
PKE	0.65	0.4254
Time_Engine_Brake	0.34	0.5637

TABLE 3.11 - Zones limitées à 90km/h : Tests de nullité des effets fixes.

Les valeurs des odds ratios sont présentées au tableau 3.12 et aucun des odds ratios n'est significatif. On peut toutefois noter que tous les odds ratios sont inférieurs à 1 sauf l'odds ratio associé à la variable Time_Engine_Brake. Ainsi comme pour les sections limitées à 70km/h, le frein moteur a été relativement peu utilisé sur les zones limitées à 90km/h.

Parameter	Odds Ratio	95% CI low	95% CI high
Avg_RPM_Shift	0.998	0.994	1.001
Index_Gear_RPM	0.929	0.808	1.068
PKE	0.078	<0.001	46.608
Time_Engine_Brake	1.025	0.941	1.117

TABLE 3.12 - Zones limitées à 90km/h : Odds ratio et intervalles de confiance à 95%.

3.5 Conclusion

Cette étude nous a permis d'étudier plus en détail la mise en oeuvre des consignes d'éco-conduite en fonction de la limitation de vitesse. Le tableau 3.13 récapitule les résultats obtenus dans ce chapitre en indiquant les paramètres significatifs et les valeurs des odds ratios associés. Nous avons obtenu que la consigne 3, associée à la variable PKE, et qui consiste à anticiper le trafic et à maintenir une vitesse constante avait été la plus appliquée, toutes limitations de vitesse confondues. La consigne 2, associée à la variable Index_Gear_RPM, et qui consiste à conduire avec un rapport de boite élevé et un régime moteur faible a été particulièrement mise en pratique sur les zones limitées à 50km/h et 70km/h. Par contre, le frein moteur (consigne 4) a été peu utilisé sur les zones limitées à 30km/h et 70km/h, et mal utilisé sur les zones limitées à 50km/h (odds ratio inférieur à 1). Cette pratique semble donc encore peu familière des conducteurs, comme nous l'avions déjà souligné au chapitre VII. La consigne 1 qui consiste à changer de rapport à un régime moteur faible n'a pas ou peu été appliqué, et ceci quelque soit la limitation de vitesse. Enfin, il faut noter qu'aucune des consignes ne semblent avoir été mises en oeuvre sur les zones limitées à 90km/h. Cependant, notre trajet d'étude favorise plutôt une conduite économique sur les zones limitées à 90km/h, même lors du trajet "Normal" (trafic fluide, grandes descentes, ...) et peut expliquer l'absence de différence significative de style de conduite entre le trajet "Normal" et le trajet "Economique".

		Zone 30	Zone 50	Zone 70	Zone 90
Avg_RPM_Shift	Significatif	Non	Non	Non	Non
	Odds Ratio	1	0.999	0.999	0.998
Index_Gear_RPM	Significatif	Non	Oui	Oui	Non
	Odds Ratio	0.888	0.805	0.808	0.929
PKE	Significatif	Oui	Oui	Oui	Non
	Odds Ratio	<0.001	<0.001	<0.001	0.078
Time_Engine_Brake	Significatif	Non	Oui	Non	Non
	Odds Ratio	1.030	0.944	1.021	1.025

TABLE 3.13 - Tableau récapitulatif de la mise en oeuvre des consignes d'éco-conduite en fonction de la limitation de vitesse.

Bibliographie

[1] Alan Agresti. *Categorical data analysis. 2nd ed.* Wiley Series in Probability and Mathematical Statistics., Chichester, 2002. 18

[2] M. Callens and C. Croux. Performance of likelihood-based estimation methods for multilevel binary regression models. *Journal of Statistical Computation and Simulation*, 75 :1003–1017, 2005. 18

[3] P.L. Flom, McMahon J.M., and Pouget E.R. Using proc nlmixed and proc glmmix to analyze dyadic data with a dichotomous dependent variable. In *Global Forum 2007 Conference. SAS Institute Inc., Cary, NC.*, 2007. 18

[4] A. Guéguen, M. Zins, and J.P. Nakache. Utilisation des modèles marginaux et des modèles mixtes dans l'analyse de données longitudinales (1992-1996) concernant mariage et consommation d'alcool des femmes de la cohorte gazel. *Revue de statistique appliquée*, tome 48, nř3 :p. 57–73, 2000. 17

[5] F. Guo and J. Hankey. Modeling 100-car safety events : A case-based approach for analyzing naturalistic driving data. Technical report, Report No. 09-UT-006, Virginia Tech Transportation Institute, 2009.

[6] F.B. Hu, J. Goldberg, D. Hedeker, B.R. Flay, and M.A. Pentz. Comparaison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, 147 :694–703, 1998. 27

[7] K.Y. Liang and S.L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73 :13–22, 1986. 11

[8] A. Ziegler, C. Kastner, and M. Blettner. The generalised estimating equations : an annotated bibliography. *Biometrical Journal*, 40 (2) :115–139, 1998. 17