TP1 P3

P3

Analyse de la multicolinéarité de la matrice shéma X

Il est préférable de regarder les facteurs d'inflations de la variance avant de commencer la modélisation.

```
## # A tibble: 22 x 3
##    Variables                Tolerance    VIF
##    <chr>                        <dbl>  <dbl>
##  1 VehValue                   0.412     2.43
##  2 VehAgeoldest cars          0.622     1.61
##  3 VehAgeyoung cars           0.654     1.53
##  4 VehAgeyoungest cars        0.588     1.70
##  5 VehBodyConvertible         0.364     2.75
##  6 VehBodyCoupe               0.0586   17.1
##  7 VehBodyHardtop             0.0302   33.1
##  8 VehBodyHatchback           0.00350 285.
##  9 VehBodyMinibus             0.0634   15.8
## 10 VehBodyMotorized caravan   0.274     3.65
## # ... with 12 more rows
```

Il y a des variables avec un VIF plus grand que 10. Normalement il serait utile de retirer certaines variables. Par contre, se sont seulement des catégories d'une variable qui sont affectées. Lorsqu'il y a une variable catégorielle avec *n* catégories dans le modèle, il y a *n-1* variables qui sont créées, qui sont des indicateurs 0,1 selon la catégorie représentée par la variable. Donc il est normal que ces variable souffrent de multicolinéarité. Ce n'est pas un problème, donc on garde tous les variables pour la modélisation.

Pour chaque modèle qui sera testé il y aura un terme d'offset. Celui ci est l'exposure, c'est à dire la proportion de l'année que l'assuré est couvert. L'espérence de son nombre d'accident sera ainsi proportionel à la proportion de l'année couvert par l'assuré. La méthodologie sera la suivante:

1. Faire un glm avec la loi de poisson et un lien log. Ajuster le meilleur modèle en utilisant des techniques algorithmiques tel que le foward, backward et la combinaison des deux.
2. Faire un glm avec la loi de binomial négative et un lien log pour tenir compte de la surdispersion. Ajuster le meilleur modèle en utilisant des techniques algorithmiques tel que le foward, backward et la combinaison des deux.
3. Faire un glm avec la loi poisson et binomial négative pour les variables trouvées retenues en 1 et 2. Pour chaque combinaison, faire un test de rapport de vraissemblance pour voir si le modèle poisson est correct ou si on doit prendre le modèle avec la loi binomial négative.

Peut importe le modèle utilisé et la technique algorithmique utilisée, la sélection de variable est identique. Les variables qui sont utilisées sont *VehBody, DrivAge & VehAge*

À ce stade si, on a un glm avec la loi de poisson et un avec la loi binomial négative. La loi binomial négative tient compte de la surdispersion. Un test des rapports de vraissemblance peut être fait pour déterminé si l'amélioration du modèle par la loi binomial négative est significative. La déviance du modèle de poisson est de *14 839* et la déviance du modèle avec la loi binomial négative est de *13 312*. La statistique de rapport de vraissemblance est de *1 527*. La pvalue associé est de 0. Donc on rejette l'hypothèse H0, on peut supposer qu'il y a de la variabilité extra poissonienne dans nos données. Voici le modèle finale :

```
##
## Call:
```

```
## glm.nb(formula = ClaimNb ~ DrivAge + VehAge + VehBody, data = data_tp3,
##     weights = offset(Exposure), init.theta = 1.695107772, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7171  -0.3509  -0.2686  -0.1718   4.0378
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.44008    0.40491  -3.557 0.000376 ***
## DrivAgeolder work. people   0.19359    0.06330   3.058 0.002226 **
## DrivAgeoldest people        0.03323    0.08238   0.403 0.686673
## DrivAgeworking people       0.21641    0.06350   3.408 0.000654 ***
## DrivAgeyoung people         0.29143    0.06553   4.448 8.68e-06 ***
## DrivAgeyoungest people      0.38723    0.07829   4.946 7.56e-07 ***
## VehAgeoldest cars          -0.11943    0.05090  -2.346 0.018966 *
## VehAgeyoung cars            0.13692    0.04952   2.765 0.005690 **
## VehAgeyoungest cars         0.08850    0.05626   1.573 0.115687
## VehBodyConvertible         -1.69463    0.84821  -1.998 0.045729 *
## VehBodyCoupe               -0.77804    0.43324  -1.796 0.072516 .
## VehBodyHardtop             -0.94536    0.41666  -2.269 0.023274 *
## VehBodyHatchback           -1.20106    0.40375  -2.975 0.002932 **
## VehBodyMinibus             -1.08642    0.44503  -2.441 0.014638 *
## VehBodyMotorized caravan   -0.56492    0.53243  -1.061 0.288679
## VehBodyPanel van           -0.88518    0.42842  -2.066 0.038816 *
## VehBodyRoadster            -0.53555    0.81661  -0.656 0.511940
## VehBodySedan               -1.11378    0.40337  -2.761 0.005759 **
## VehBodyStation wagon       -1.06035    0.40374  -2.626 0.008631 **
## VehBodyTruck               -1.04230    0.41666  -2.502 0.012364 *
## VehBodyUtility             -1.26548    0.40929  -3.092 0.001989 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.6951) family taken to be 1)
##
##     Null deviance: 13403  on 67855  degrees of freedom
## Residual deviance: 13312  on 67835  degrees of freedom
## AIC: 20626
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.695
##           Std. Err.:  0.309
##
##  2 x log-likelihood:  -20582.183
```

La valeur relative du véhicule n'est pas associée au nombre de réclamations esperés. La relation entre la fréquence de sinistre et la valeur du véhicule ne fait pas beaucoup de sens. La valeur du véhicule serait plus associée à la sévérité des sinistres qu'a la fréquence.