Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2Q20

GLM #23 (example)

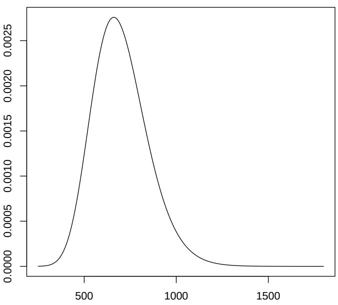
Fréquence & coût en assurance

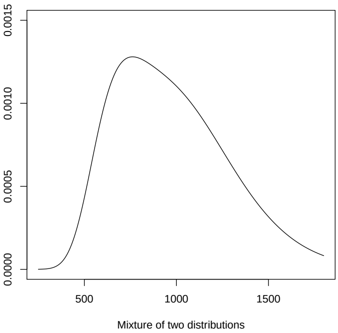
```
>  sinistre=read.table("http://freakonometrics.free.fr/sinistreACT2040.txt",
     header=TRUE,sep=";")
> contrat=read.table("http://freakonometrics.free.fr/contractACT2040.txt",header
     =TRUE,sep=";")
> contrat=contrat[,1:10]
> names(contrat)[10]="region"
> sinistre_DO=sinistre[(sinistre$garantie=="2DO")&(sinistre$cout>0),]
> sinistre_RC=sinistre[(sinistre$garantie=="1RC")&(sinistre$cout>0),]
> base_DO=merge(sinistre_DO,contrat)
> dim(base_DO)
[1] 1735   13
> base_RC=merge(sinistre_RC,contrat)
> dim(base_RC)
[1] 1924   13
```

Idée de base: si $S = \sum_{i=1}^{N} Y_i$,

$$\pi(\boldsymbol{x}) = \mathbb{E}[S|\boldsymbol{X}=\boldsymbol{x}] = \mathbb{E}[N|\boldsymbol{X}=\boldsymbol{x}] \cdot \mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]$$

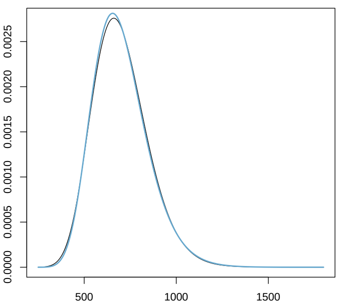
Coût : log-normale ou Gamma ?



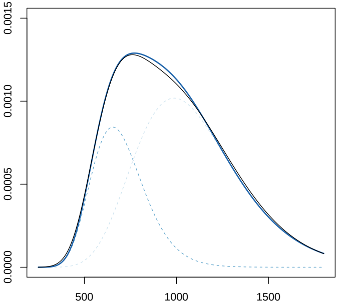


Coût : log-normale ou Gamma?

Loi Gamma ? Mélange de deux lois Gamma ?



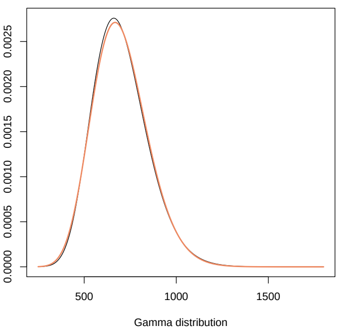
Lognormal distribution

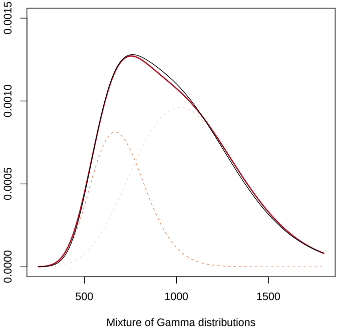


Mixture of lognormal distributions

Coût : log-normale ou Gamma?

Loi log-normale ? Mélange de deux lois log-normales ?





Coût : Gamma

Pour la régression Gamma (et un lien log i.e. $\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})=\exp[\boldsymbol{x}^\top\beta]$), on a

```
> regg=glm(cout~agevehicule+carburant+zone,data=base_RC,
+       family=Gamma(link="log"))
> summary(regg)
(Intercept)  7.72660    0.09300  83.079  < 2e-16 ***
agevehicule -0.04674    0.00855  -5.466 5.27e-08 ***
carburantE  -0.14693    0.06329  -2.321  0.02038 *
zoneB       -0.14876    0.12690  -1.172  0.24124
zoneC       -0.04275    0.09924  -0.431  0.66668
zoneD       -0.11026    0.10416  -1.058  0.28998
zoneE       -0.12129    0.10478  -1.158  0.24719
zoneF       -0.47684    0.18142  -2.628  0.00865 **

(Dispersion parameter for Gamma family taken to be 1.686782)
```

Coût : Inverse-Gaussienne

Pour la régression inverse-Gaussienne, (et un lien log i.e. $\mathbb{E}(Y|\boldsymbol{X}) = \exp[\boldsymbol{X}'\beta]$),

```
> regig=glm(cout~agevehicule+carburant+zone,data=base_DO,
+      family=inverse.gaussian(link="log"),start=coefficients(regg))
> summary(regig)
Coefficients:
(Intercept)  7.731661   0.093390  82.789  < 2e-16 ***
agevehicule -0.046699   0.007016  -6.656 3.76e-11 ***
carburantE  -0.153028   0.061479  -2.489  0.01290 *
zoneB       -0.138902   0.123192  -1.128  0.25968
zoneC       -0.054040   0.098951  -0.546  0.58505
zoneD       -0.102103   0.102734  -0.994  0.32043
zoneE       -0.127266   0.103662  -1.228  0.21973
zoneF       -0.492622   0.155715  -3.164  0.00159 **

(Dispersion parameter for inverse.gaussian family taken to be 0.001024064)
```

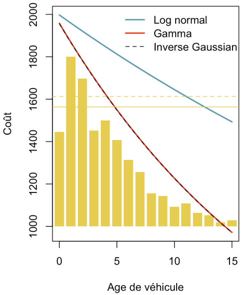
Pour la régression log-normale i.e. $\mathbb{E}(\log Y|\boldsymbol{X}) = \boldsymbol{X}'\beta$, on a

```
> regln=lm(log(cout)~agevehicule+carburant+zone,data=base_DO)
> summary(regln)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.776664   0.094371  71.809   <2e-16 ***
agevehicule -0.019397   0.008676  -2.236   0.0255 *
carburantE  -0.045508   0.064224  -0.709   0.4787
zoneB       -0.022196   0.128763  -0.172   0.8632
zoneC        0.056457   0.100695   0.561   0.5751
zoneD       -0.008894   0.105694  -0.084   0.9330
zoneE        0.017727   0.106321   0.167   0.8676
zoneF       -0.363002   0.184087  -1.972   0.0488 *

Residual standard error: 1.318 on 1727 degrees of freedom
> sigma=summary(regln)$sigma
```

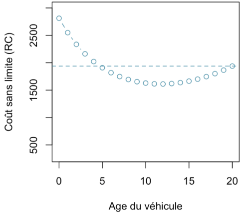
Coût : Comparison

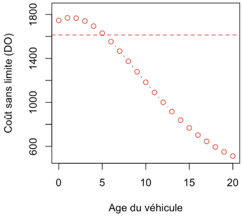
```
> nd = data.frame(agevehicule=seq(0,15,by=.25),
    carburant="E",zone="A")
> Tb = table(base_DO$agevehicule)
> ypln = exp(predict(regln,newdata=nd)+.5*sigma^2)
> ypg = predict(regg,newdata=nd,type="response")
> ypig = predict(regig,newdata=nd,type="response")
> plot(nd$agevehicule,ypln,col="blue")
> lines(nd$agevehicule,ypg,lwd=2,col=colr[2])
> lines(nd$agevehicule,ypig,lty=2)
> abline(h=mean(base_DO$cout),lty=2)
abline(h=mean(base_DO$cout[(base_DO$carburant=="E")
    &(base_DO$zone=="A")]))
```



Coût : RC vs DO

```
> library(bs)
> reg=glm(cout~bs(agevehicule),data=base_RC,family=Gamma(link="log"))
> age = 0:15
> yp=predict(reg,newdata=data.frame(agevehicule=age))
> plot(age,yp,type="b")
> reg=glm(cout~bs(agevehicule),data=base_DO,family=Gamma(link="log"))
> age = 0:15
> yp=predict(reg,newdata=data.frame(agevehicule=age))
> plot(age,yp,type="b")
```





Coût : coûts importants

On a ici quelques *gros* sinistres. L'idée est de noter que

$$\mathbb{E}(Y) = \sum_i \mathbb{E}(Y|\Theta = \theta_i) \cdot \mathbb{P}(\Theta = \theta_i)$$

Supposons que Θ prenne deux valeurs, correspondant au cas $\{Y \leq s\}$ et $\{Y > s\}$. Alors

$$\mathbb{E}(Y) = \mathbb{E}(Y|Y \leq s) \cdot \mathbb{P}(Y \leq s) + \mathbb{E}(Y|Y > s) \cdot \mathbb{P}(Y > s)$$

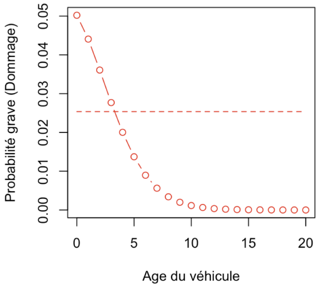
ou, en calculant l'espérance sous $\mathbb{P}_{\boldsymbol{X}}$ et plus $\mathbb{P}$,

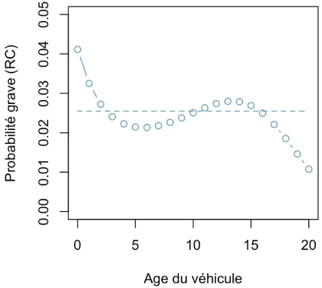
$$\mathbb{E}(Y|\boldsymbol{X}) = \underbrace{\mathbb{E}(Y|\boldsymbol{X}, Y \leq s)}_{A} \cdot \underbrace{\mathbb{P}(Y \leq s|\boldsymbol{X})}_{B} + \underbrace{\mathbb{E}(Y|Y > s, \boldsymbol{X})}_{C} \cdot \underbrace{\mathbb{P}(Y > s|\boldsymbol{X})}_{B^\star}$$

Probabilité d'un grave

Pour le terme B, il s'agit d'une régression *standard* d'une variable de Bernoulli,

```
> sRC = quantile(base_RC$cout,.975)
> sRC
  97.5%
8677.38
> sDO = quantile(base_DO$cout,.975)
> sDO
   97.5%
8203.183
> base_RC$anormal=(base_RC$cout>=sRC)
> library(splines)
> age=seq(0,20)
> regB=glm(normal~bs(agevehicule),data=base_RC,
     family=binomial)
> ypB=predict(regB,newdata=data.frame(agevehicule=
     age),type="response")
> plot(age,ypB,type="b")
```

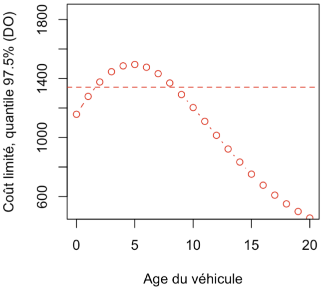


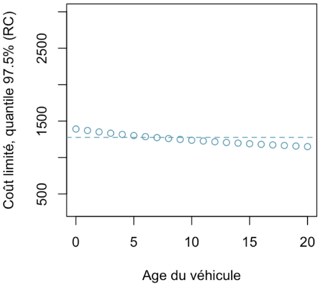


Coût d'un non-grave

Pour le terme *A*, il s'agit d'une régression *standard* Gamma,

```
> library(splines)
> indice = which(base_RC$cout<=sRC)
> mean(base_RC$cout[indice])
[1] 1269.22
> regA = glm(cout~bs(agevehicule),data=base_RC,
         subset=indice,family=Gamma(link="log"))
> ypA = predict(regA,newdata=data.frame(agevehicule=
      age),type="response")
> plot(age,ypA,type="b")
```

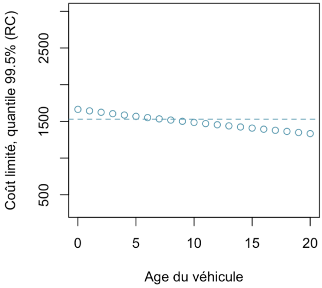


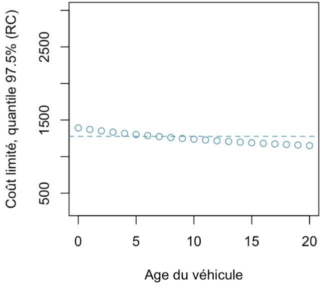


Coût d'un non-grave

On peut prendre un seuil plus élevé

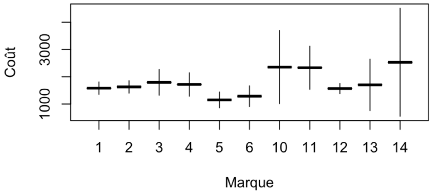
```
> sRC = quantile(base_RC$cout,.995)
> sRC
   99.5%
23978.43
> library(splines)
> indice = which(base_RC$cout<=sRC)
> mean(base_RC$cout[indice])
[1] 1529.892
> regA = glm(cout~bs(agevehicule),data=base_RC,
            subset=indice,family=Gamma(link="log"))
> ypA = predict(regA,newdata=data.frame(agevehicule=
     age),type="response")
> plot(age,ypA,type="b")
```





Modéliser les coûts individuels

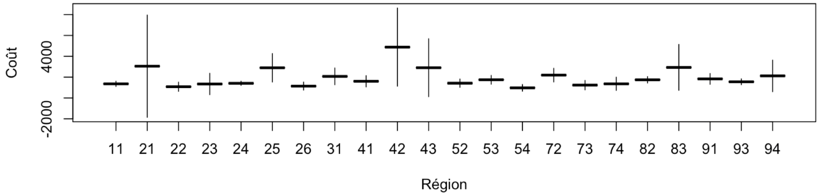
```r
k = which(names(base_DO) %in% c("garantie","no","nocontrat","exposition"))
base_DO = base_DO[,-k]
base_DO$marque = as.factor(base_DO$marque)
A= aggregate(base_DO$cout , by=list(base_DO$marque),function(x) c(mean(x)-2*sd(x)
     /sqrt(length(x)),mean(x),mean(x)+2*sd(x)/sqrt(length(x))))
plot(A$Group.1,A$x[,2],ylim=range(A$x))
segments(1:nrow(A),A$x[,1],1:nrow(A),A$x[,3])
```




```r
levels(base_DO$marque) = c("A","A","A","A","B","B","C","C","A","A","C")
levels(base_DO$zone) = c("A","A","A","A","A","F")
```

Modéliser les coûts individuels

```
base_DO$region = as.factor(base_DO$region)
A= aggregate(base_DO$cout , by=list(base_DO$region),function(x) c(mean(x)-2*sd(x)
     /sqrt(length(x)),mean(x),mean(x)+2*sd(x)/sqrt(length(x))))
plot(A$Group.1,A$x[,2],ylim=range(A$x))
segments(1:nrow(A),A$x[,1],1:nrow(A),A$x[,3])
```




```
levels(base_DO$region) = c("B","E","A","B",  "B","E","A","D", "B","F","E","B","C
     ","A","D","A",  "B","C","E","C", "B","D")
```

Modéliser les coûts individuels

```
> regg=glm(cout~.,data=base_DO , family=Gamma(link="log"))
> summary(regg)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.159940   0.229208  31.238  < 2e-16 ***
zoneF         -0.331000   0.156797  -2.111 0.034915 *
puissance      0.035415   0.014598   2.426 0.015369 *
agevehicule   -0.033725   0.007937  -4.249 2.26e-05 ***
ageconducteur -0.006137   0.002309  -2.658 0.007933 **
bonus          0.006451   0.002141   3.013 0.002626 **
marqueB       -0.258137   0.103477  -2.495 0.012702 *
marqueC        0.309273   0.141523   2.185 0.029000 *
carburantE    -0.096018   0.059037  -1.626 0.104046
regionE        0.565601   0.165092   3.426 0.000627 ***
regionA       -0.287633   0.111578  -2.578 0.010024 *
regionD        0.296621   0.101872   2.912 0.003641 **
regionF        0.860987   0.603632   1.426 0.153951
regionC        0.126424   0.071222   1.775 0.076064 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 1.442703)

    Null deviance: 2464.2  on 1734  degrees of freedom
Residual deviance: 2259.6  on 1721  degrees of freedom
AIC: 28937

Number of Fisher Scoring iterations: 7
```

Modéliser les coûts individuels

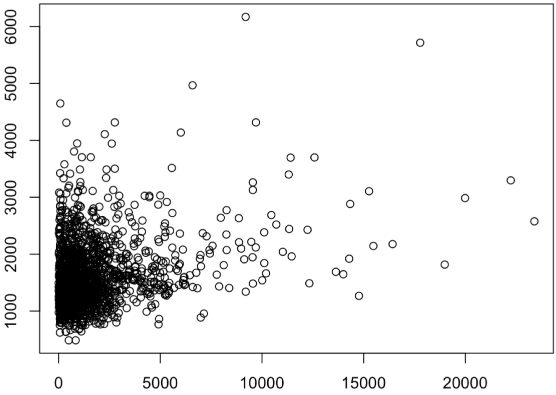
```
> regln=lm(log(cout)~.,data=base_DO)
> summary(regln)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.396580   0.248328  25.759  < 2e-16 ***
zoneF          -0.369535   0.169876  -2.175  0.02974 *
puissance       0.009819   0.015816   0.621  0.53478
agevehicule    -0.015695   0.008599  -1.825  0.06813 .
ageconducteur  -0.002956   0.002502  -1.182  0.23748
bonus           0.006678   0.002320   2.878  0.00405 **
marqueB        -0.226641   0.112109  -2.022  0.04337 *
marqueC         0.334085   0.153328   2.179  0.02948 *
carburantE     -0.022320   0.063961  -0.349  0.72716
regionE         0.408499   0.178863   2.284  0.02250 *
regionA        -0.232556   0.120885  -1.924  0.05455 .
regionD         0.190320   0.110369   1.724  0.08482 .
regionF         1.314463   0.653985   2.010  0.04459 *
regionC         0.157108   0.077163   2.036  0.04190 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.301 on 1721 degrees of freedom
Multiple R-squared:  0.03465, Adjusted R-squared:  0.02736
F-statistic: 4.751 on 13 and 1721 DF,  p-value: 3.657e-08
```

Modéliser les coûts individuels

```
> plot(base_DO$cout,predict(regg,type="response"))
```



pour aller plus loin ? ACT6100