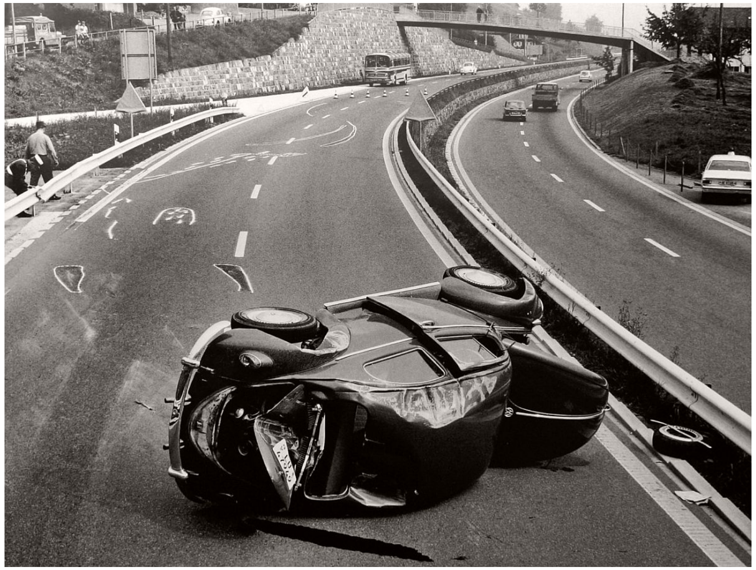
Actuariat de l'Assurance Non-Vie # 5

A. Charpentier (Université de Rennes 1)



ENSAE 2017/2018

credit: Arnold Odermatt

Surdispersion?

Références: de Jong & Heller (2008), sections 6.2 et 6.3, Hilbe (2011).

Plusieurs documents sur le site http://www.casact.org/pubs/forum/ évoquent le problème de la surdispersion. Les données sont les mêmes que pour la section #3,

```
> 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"
```

...

```
> freq = merge(contrat,nombre_RC)
> freq = merge(freq,nombre_DO)
```

Surdispersion ? *(overdispersion)*

(extrait de Hilbe (2011), chapitre 7)

1. **What is overdispersion?** *Overdispersion in Poisson models occurs when the response variance is greater than the mean.*

2. **What causes overdispersion?** *Overdispersion is caused by positive correlation between responses or by an excess variation between response probabilities or counts. Overdispersion also arises when there are violations in the distributional assumptions of the data, such as when the data are clustered and thereby violate the likelihood independence of observations assumption.*

3. **Why is overdispersion a problem?** *Overdispersion may cause standard errors of the estimates to be deflated or underestimated, i.e. a variable may appear to be a significant predictor when it is in fact not significant.*

4. **How is overdispersion recognized?** *A model may be overdispersed if the*

value of the Pearson (or χ^2) statistic divided by the degrees of freedom (dof) is greater than 1.0. The quotient of either is called the dispersion.

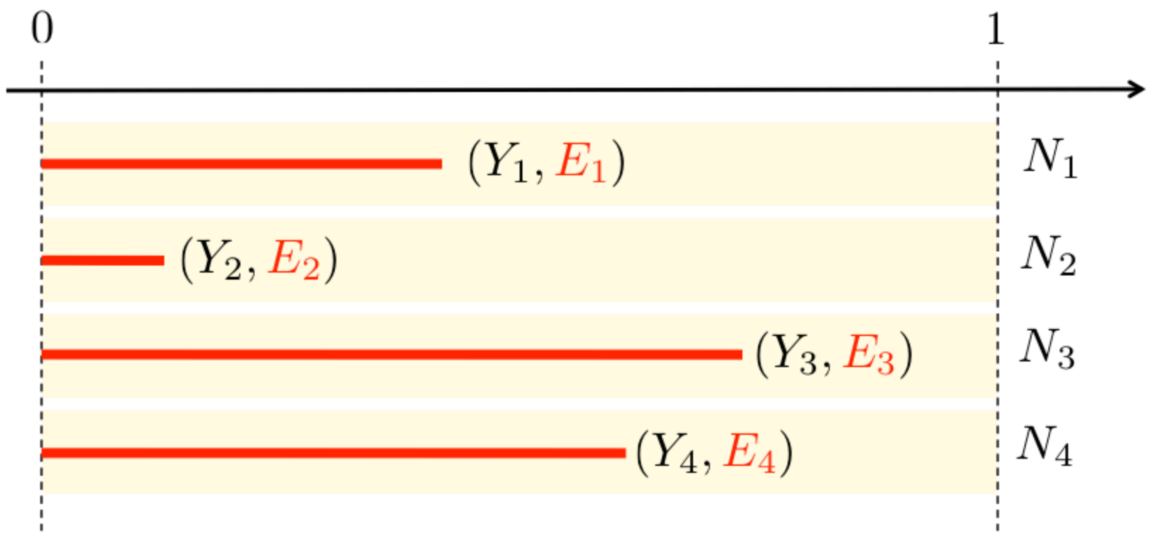
Small amounts of overdispersion are of little concern; however, if the dispersion statistic is greater than 1.25 for moderate sized models, then a correction may be warranted. Models with large numbers of observations may be overdispersed with a dispersion statistic of 1.05.

5. **What is apparent overdispersion; how may it be corrected?**
   Apparent overdispersion occurs when:

   - *the model omits important explanatory predictors;*
   - *the data include outliers;*
   - *the model fails to include a sufficient number of interaction terms;*
   - *a predictor needs to be transformed to another scale;*
   - *the assumed linear relationship between the response and the link function and predictors is mistaken, i.e. the link is misspecified.*

La surdispersion...

Rappelons que l'on observe des $(Y_i, E_i, \boldsymbol{X}_i)$, où E_i est l'exposition, Y_i est le nombre de sinistres observés sur la période $[0, E_i]$. On peut voir cela comme un problème de données censurées,



on voudrait N_i le nombre de sinistres **non**-observés sur la période $[0, 1]$

La surdispersion...

Pour estimer la moyenne de N (cf partie 3),

$$m_N = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n E_i}$$

et pour estimer la variance de N

$$S_N^2 = \frac{\sum_{i=1}^n [Y_i - m_N \cdot E_i]^2}{\sum_{i=1}^n E_i}$$

soit, numériquement

```
> Y <- baseFREQ$nbre
> E <- baseFREQ$exposition
> m=sum(vY)/sum(vE)
> v=sum((vY-m*vE)^2)/sum(vE)
> cat("average =",m," variance =",v,"phi =",v/m,"\n")
average = 0.0697  variance = 0.0739 phi = 1.0597
```

où phi vérifie $S_N^2 = \varphi \cdot m_X$.

La surdispersion...

On peut aussi prendre en compte de variables explicatives

$$m_{N,\boldsymbol{x}} = \frac{\sum_{i,\boldsymbol{X}_i=\boldsymbol{x}} Y_i}{\sum_{i,\boldsymbol{X}_i=\boldsymbol{x}} E_i}$$

et pour estimer la variance de N

$$S^2_{N,\boldsymbol{x}} = \frac{\sum_{i,\boldsymbol{X}_i=\boldsymbol{x}} [Y_i - m_N \cdot E_i]^2}{\sum_{i,\boldsymbol{X}_i=\boldsymbol{x}} E_i}$$

soit, numériquement

```
> X=as.factor(baseFREQ$carburant)
> for(i in 1:length(levels(X))){
+       Ei=E[X==levels(X)[i]]
+       Yi=Y[X==levels(X)[i]]
+  meani=weighted.mean(Yi/Ei,Ei)
+  variancei=sum((Yi-meani*Ei)^2)/sum(Ei)
```

La surdispersion....

```
+ cat("Carburant, zone",levels(X)[i],"average =",meani," variance =",
    variancei,"\n")
+ }
Carburant, zone D average = 0.07881126  variance = 0.08425971
Carburant, zone E average = 0.06917643  variance = 0.07753765
```

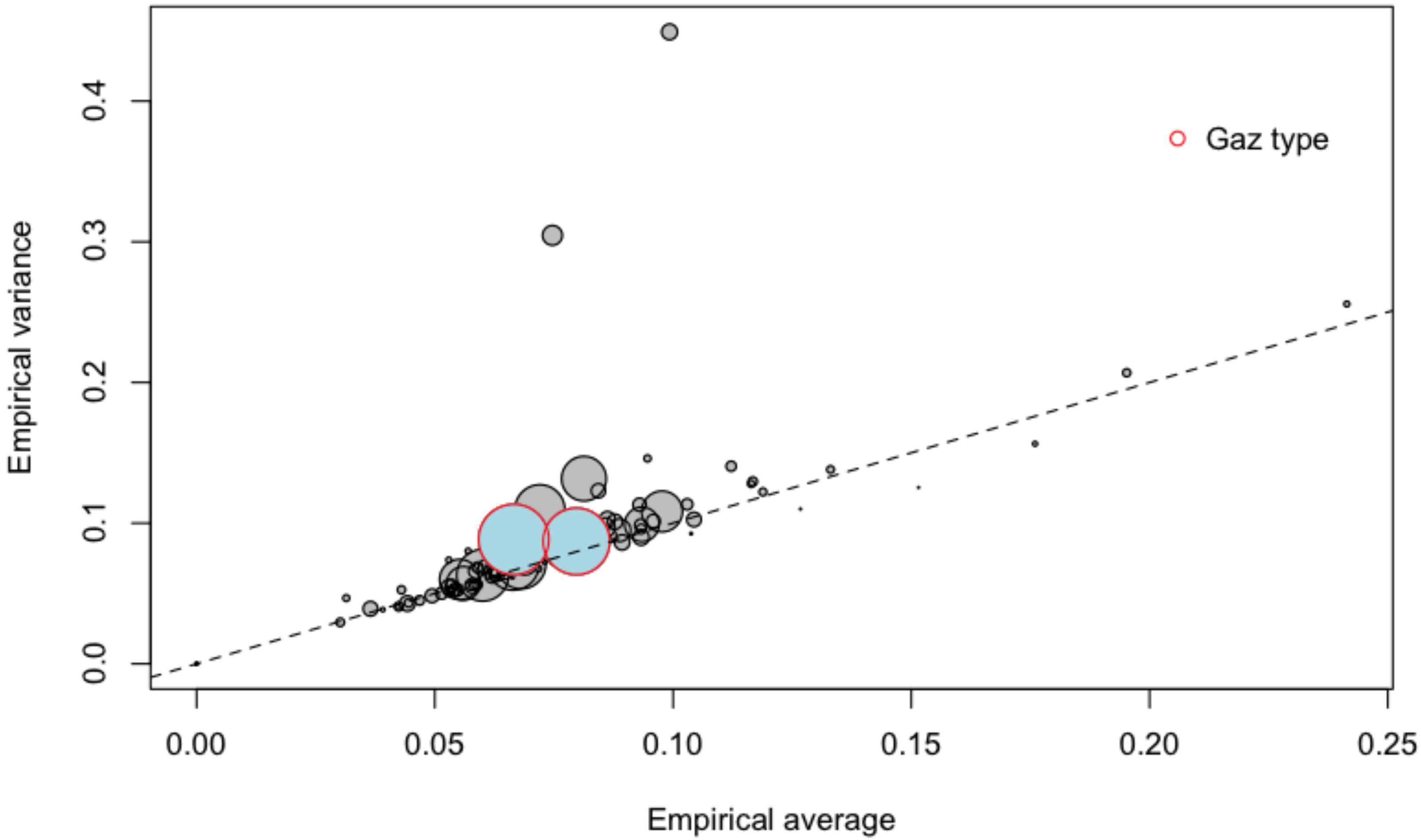
que l'on peut représenter graphiquement

```
> plot(meani,variancei,cex=sqrt(Ei),col="grey",pch=19,
+ xlab="Empirical average",ylab="Empirical variance")
> points(meani,variancei,cex=sqrt(Ei))
```

La taille du disque est proportionnelle à l'exposition total dans la classe $\{i, \boldsymbol{X}_i = \boldsymbol{x}\}$, et la première diagonale correspond au cas où $\mathbb{E}(N) = \text{Var}(N)$.

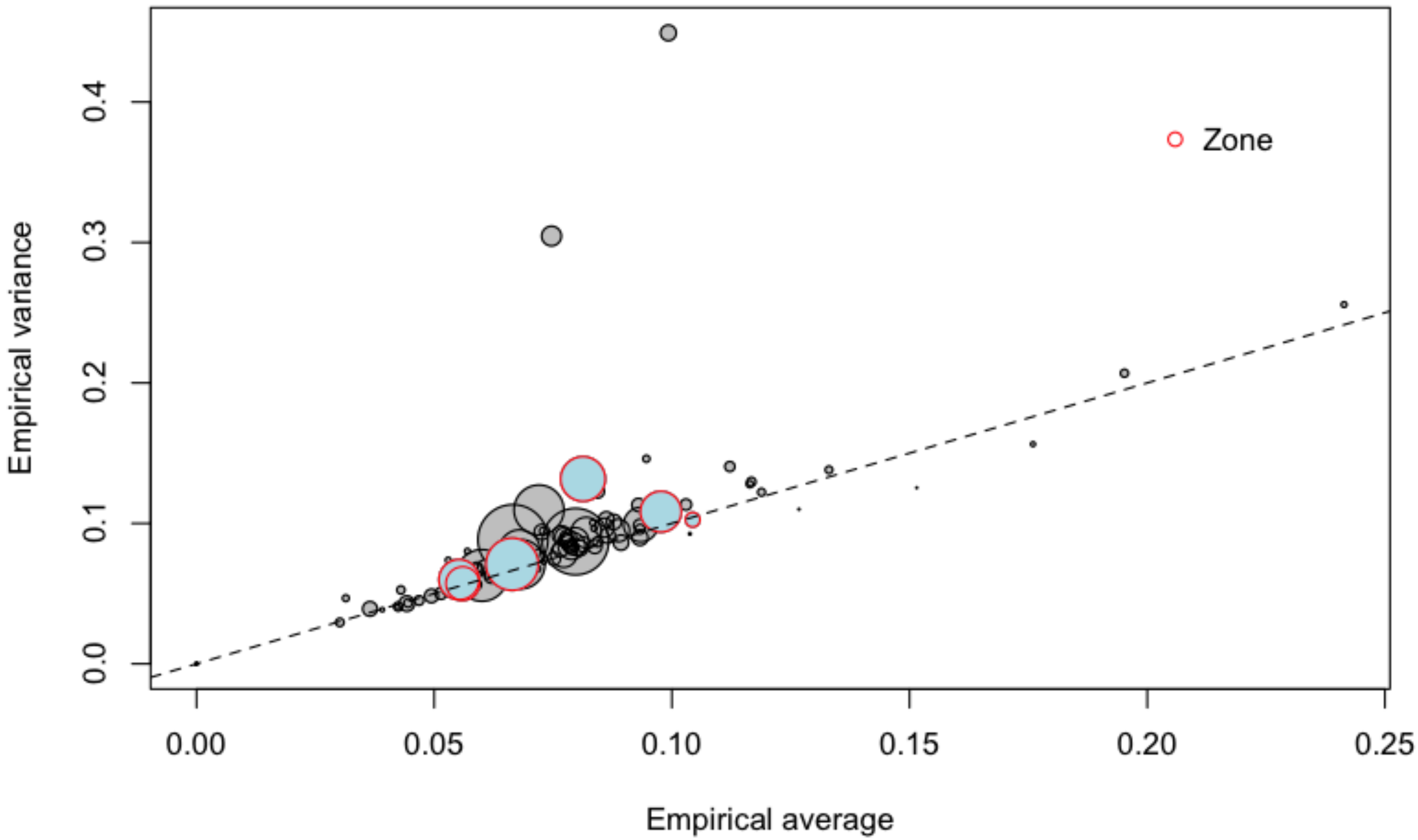
La surdispersion, et type de carburant

Estimation de $\text{Var}(N|X)$ versus $\mathbb{E}(N|X)$, si X est le carburant



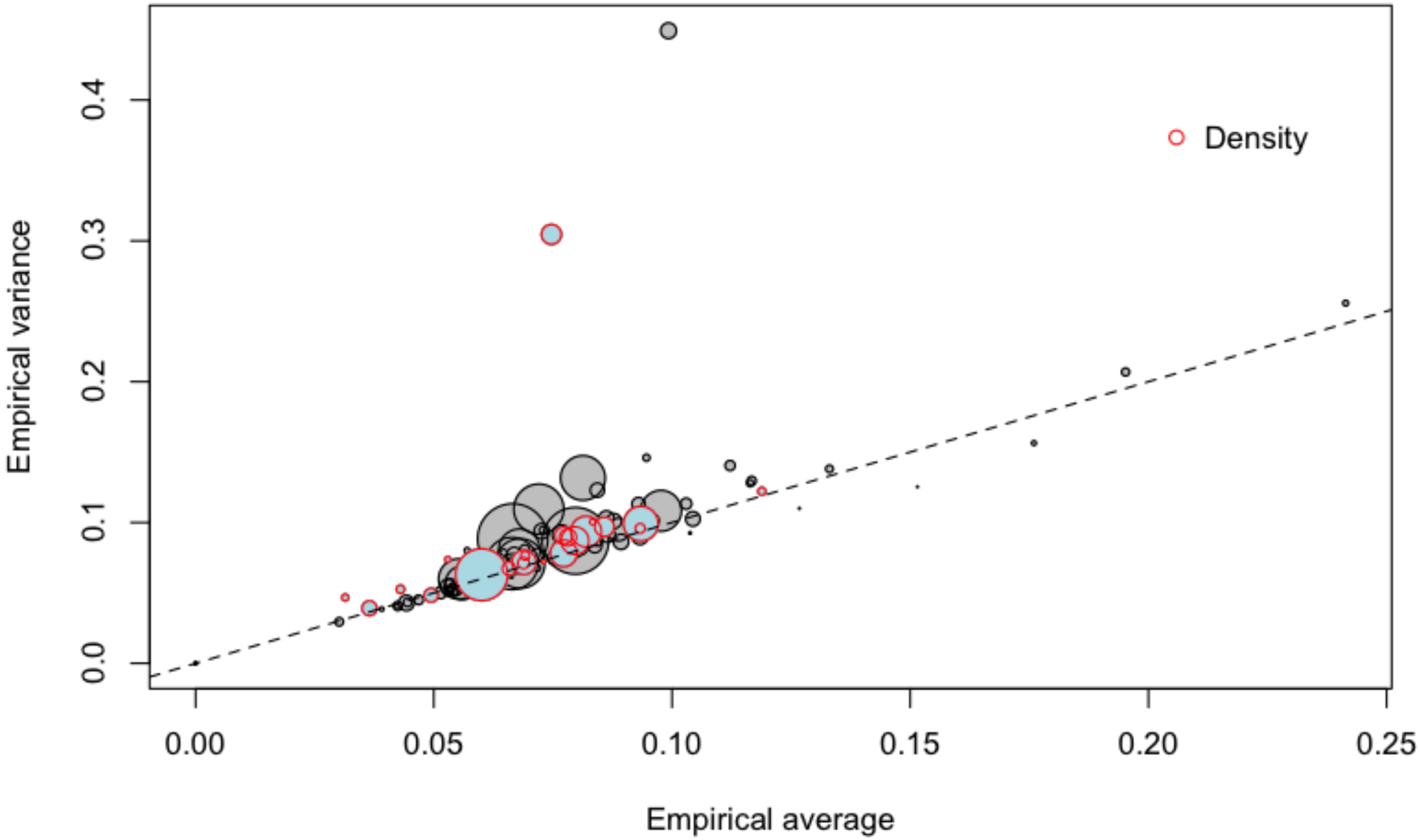
La surdispersion, et zone géographique

Estimation de $\text{Var}(N|X)$ versus $\mathbb{E}(N|X)$, si X est la zone géographique



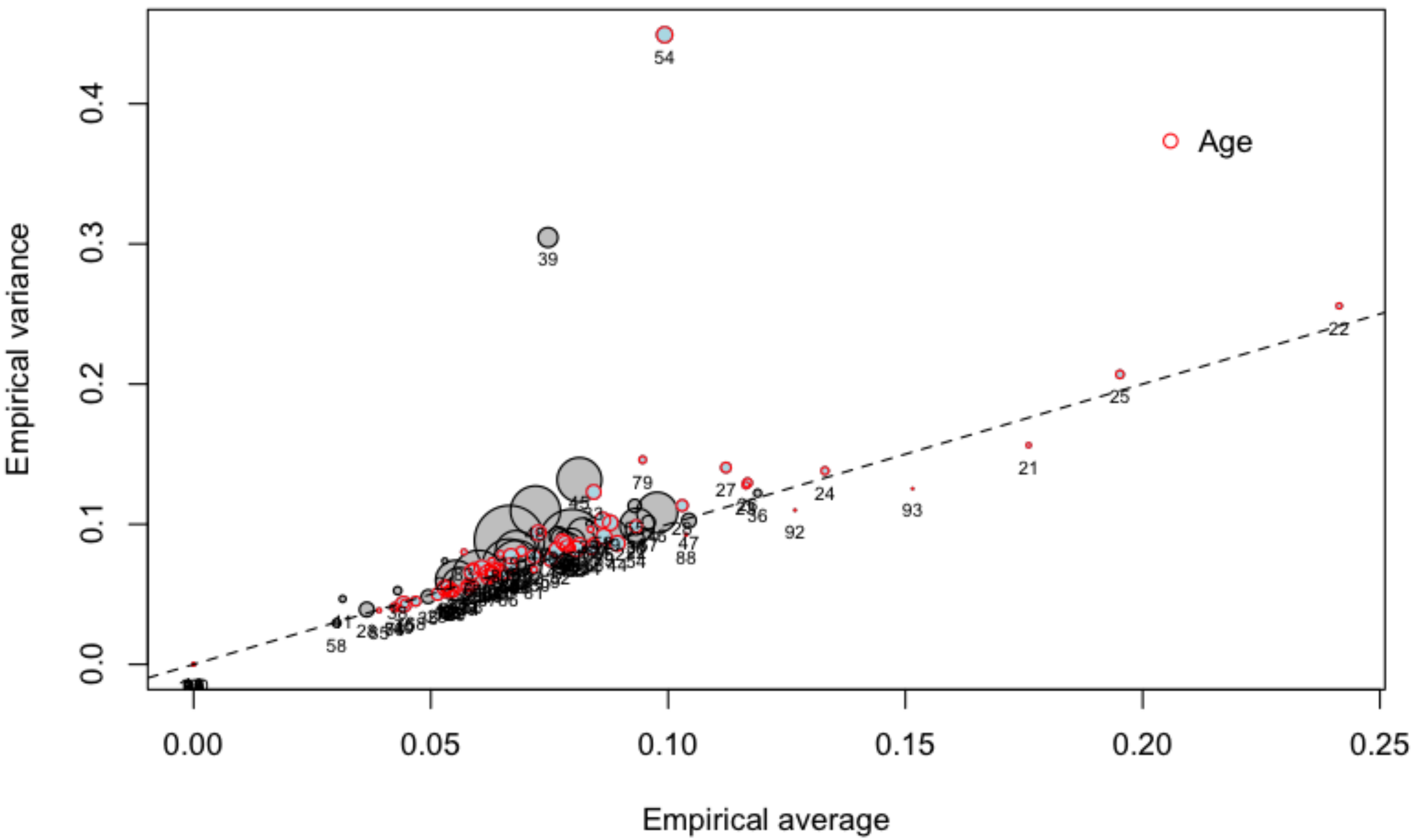
La surdispersion, et densité de population

Estimation de $\text{Var}(N|X)$ versus $\mathbb{E}(N|X)$, si X est la densité de population



La surdispersion, et âge du conducteur

Estimation de $\text{Var}(N|X)$ versus $\mathbb{E}(N|X)$, si X est l'âge du conducteur



La surdispersion

L'équidispersion est une propriété de la loi de Poisson: si $Y \sim \mathcal{P}(\lambda)$, $\mathbb{E}(Y) = \text{Var}(Y) = \lambda$.

Si on suppose maintenant que $Y|\Theta \sim \mathcal{P}(\lambda\Theta)$, où Θ suit une loi Gamma de paramètres identiques α (de telle sorte que $\mathbb{E}(\Theta) = 1$), on obtient la loi binomiale négative,

$$\mathbb{P}(Y = k) = \frac{\Gamma(k + \alpha^{-1})}{\Gamma(k+1)\Gamma(\alpha^{-1})}\left(\frac{1}{1+\lambda/\alpha}\right)^{\alpha^{-1}}\left(1 - \frac{1}{1+\lambda/\alpha}\right)^{k}, \forall k \in \mathbb{N}$$

On peut réécrire cette loi, en posant $r = \alpha^{-1}$ et $p = \dfrac{1}{1+\alpha\lambda}$

$$f(y) = \binom{y}{y+r-1} p^r [1-p]^y, \forall k \in \mathbb{N}$$

La surdispersion

ou encore

$$f(y) = \exp\left[y \log(1-p) + r \log p + \log \binom{y}{y+r-1}\right], \forall k \in \mathbb{N}$$

qui est une loi de la famille exponentielle, en posant $\theta = \log[1-p]$, $b(\theta) = -r \log(p)$ et $a(\varphi) = 1$.

Si on calcule la moyenne, on obtient

$$\mathbb{E}(Y) = b'(\theta) = \frac{\partial b}{\partial p}\frac{\partial p}{\partial \theta} = \frac{r(1-p)}{p} = \lambda,$$

et si on calcule la variance

$$\text{Var}(Y) = b''(\theta) = \frac{\partial^2 b}{\partial p^2}\left(\frac{\partial p}{\partial \theta}\right)^2 + \frac{\partial b}{\partial p}\frac{\partial^2 p}{\partial \theta^2} = \frac{r(1-p)}{p^2}$$

La surdispersion

Autrement dit

$$\text{Var}(Y) = \frac{1}{p}\mathbb{E}(Y) = [1 + \alpha \cdot \lambda] \cdot \lambda$$

Pour une régression binomiale négative de type 2 (NB2),

$$\mathbb{E}(Y) = \lambda = \mu \text{ et } \text{Var}(Y) = \lambda + \alpha\lambda^2.$$

Le lien canonique est $g(\lambda) = \theta$, i.e. $g(\mu) = \log\left(\frac{\alpha\mu}{1+\alpha\mu}\right)$

Remarque: si $\alpha = 0$, on a une loi de Poisson; si $\alpha = 1$, on a une loi géométrique.

Remarque: sous R, la régression NB2 se fait avec la fonction `glm.nb` de `library(MASS)`

Régression de Poisson

```
> regpoisson=glm(nbre~carburant+zone+ageconducteur+offset(log(
    exposition)),family=poisson("log"),data=baseFREQ)
> summary(regpoisson)
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.0176256  0.0536858 -18.955  < 2e-16 ***
carburantE    -0.2930830  0.0258247 -11.349  < 2e-16 ***
zoneB         -0.0815939  0.0516502  -1.580  0.11417
zoneC          0.1074269  0.0402553   2.669  0.00762 **
zoneD          0.2027631  0.0421105   4.815 1.47e-06 ***
zoneE          0.3041587  0.0430113   7.072 1.53e-12 ***
zoneF          0.4430020  0.0821833   5.390 7.03e-08 ***
ageconducteur -0.0092141  0.0009236  -9.976  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

Régression Binomiale Négative

```
> library(MASS)
> regnb=glm.nb(nbre~carburant+zone+ageconducteur+offset(log(
  exposition)),data=baseFREQ)
> summary(regnb)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.979677   0.063003 -15.550  < 2e-16 ***
carburantE    -0.294948   0.030338  -9.722  < 2e-16 ***
zoneB         -0.090319   0.059876  -1.508   0.1314
zoneC          0.099992   0.047132   2.122   0.0339 *
zoneD          0.208227   0.049360   4.219 2.46e-05 ***
zoneE          0.300395   0.050604   5.936 2.92e-09 ***
zoneF          0.407373   0.099042   4.113 3.90e-05 ***
ageconducteur -0.009268   0.001078  -8.596  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
(Dispersion parameter for Negative Binomial(0.4231) family taken to
    be 1)
```

Autres régression

En fait, dans la fonction `glm`, il est possible de spécifier `family=negative.binomiale`, mais il faut alors indiquer la valeur de α (qui n'est alors plus estimé).

Par exemple, `negative.binomiale(1)` permet d'avoir un régression géométrique, i.e.

$$\mathbb{E}(Y) = \lambda = \mu \text{ et } \text{Var}(Y) = \lambda + \lambda^2.$$

Le lien canonique est alors $g(\lambda) = \theta$, i.e. $g(\mu) = \log\left(\frac{\mu}{1+\mu}\right)$

```
> reggeo=glm(nbre~carburant+zone+ageconducteur+offset(log(exposition
    )),family=negative.binomial(1),data=baseFREQ)
> summary(reggeo)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.999349   0.080774 -12.372  < 2e-16 ***
carburantE     -0.294294   0.038878  -7.570 3.80e-14 ***
zoneB          -0.085721   0.077204  -1.110 0.266871
zoneC           0.103607   0.060485   1.713 0.086730 .
zoneD           0.205556   0.063309   3.247 0.001167 **
zoneE           0.302989   0.064799   4.676 2.94e-06 ***
zoneF           0.425222   0.125571   3.386 0.000709 ***
ageconducteur  -0.009243   0.001386  -6.669 2.61e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1) family taken to be
    1.947198)
```

Les régressions pour des données de comptage

Il existe aussi une `library(counts)`

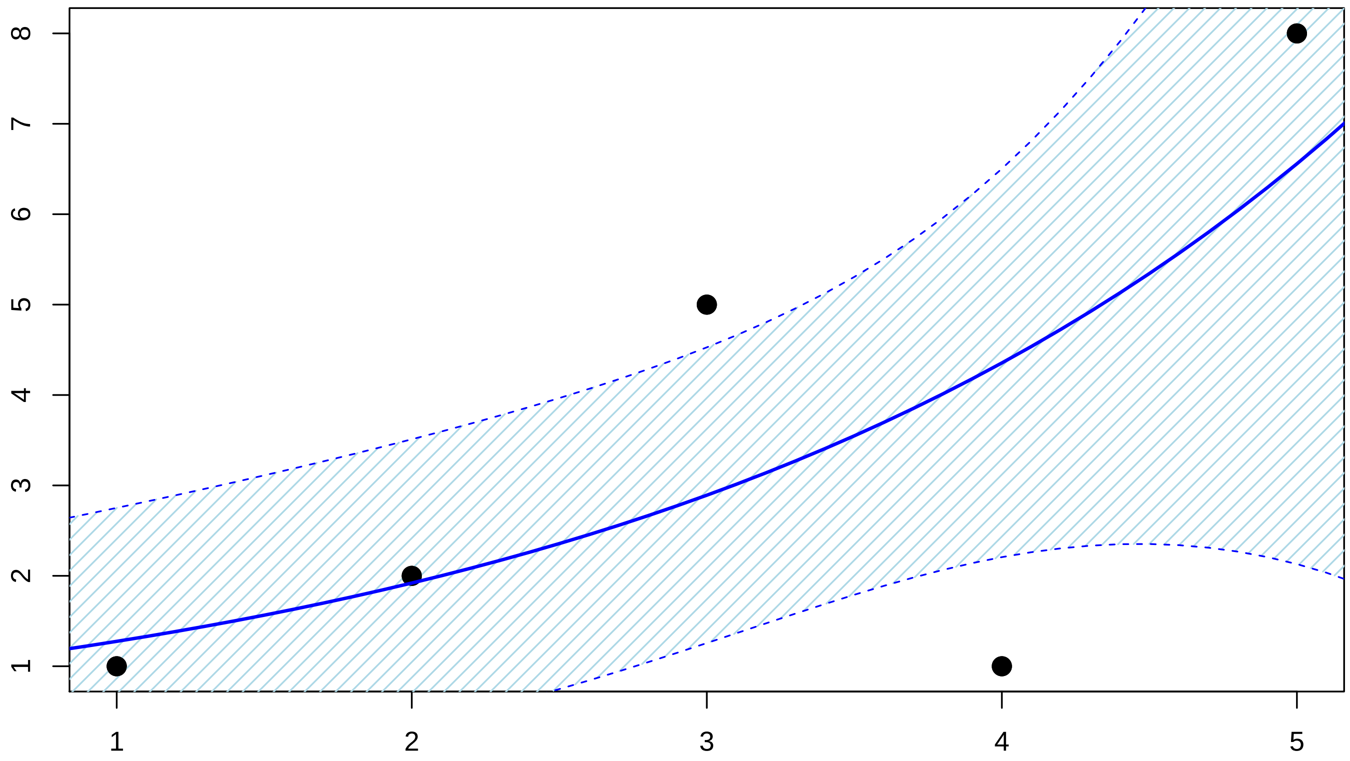
- Poisson, fonction variance $V(\mu) = \mu$, `ml.pois`
- quasiPoisson, fonction variance $V(\mu) = \varphi\mu$,
- géométrique, fonction variance $V(\mu) = \mu + \mu^2$,
- négative binomiale (NB1), fonction variance $V(\mu) = \mu + \alpha\mu$, `ml.nb1`
- négative binomiale (NB2), fonction variance $V(\mu) = \mu + \alpha\mu^2$, `ml.nb2`

Pour comparer ces modèles utilisons le petit échantillon

```
> x <- c(1,2,3,4,5)
> y <- c(1,2,5,1,8)
```

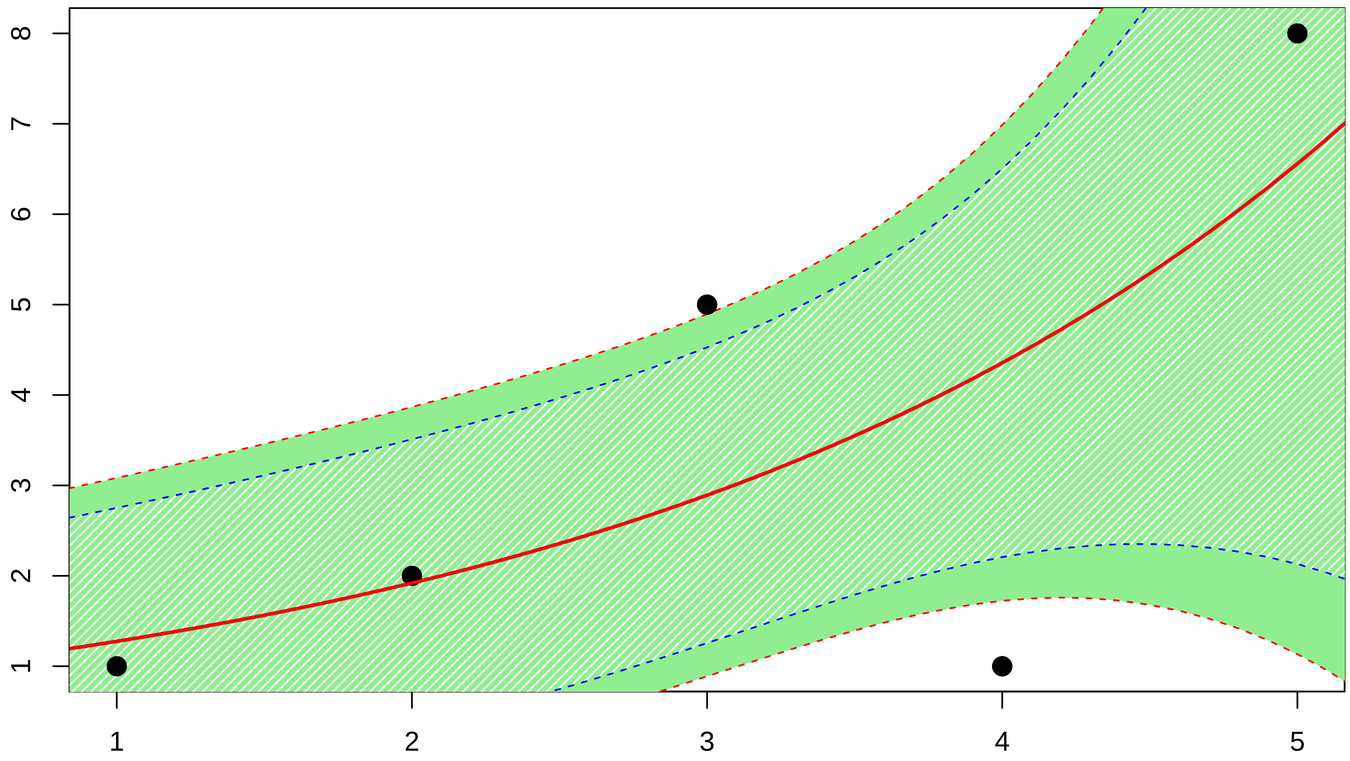
Les régressions pour des données de comptage

```
> regression <- glm(y~x,family=poisson(link="log"))
```



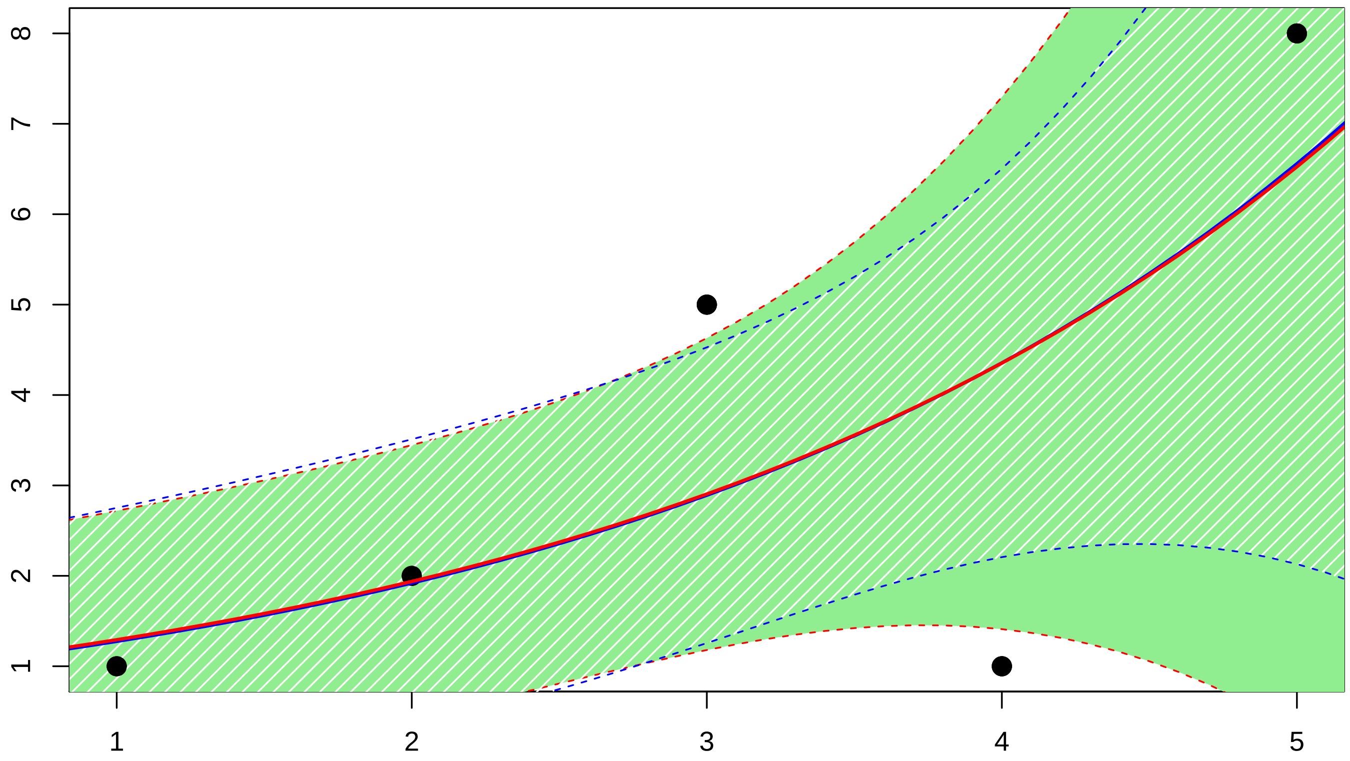
Les régressions pour des données de comptage

```
> regression <- glm(y~x,family=quasipoisson(link="log"))
```



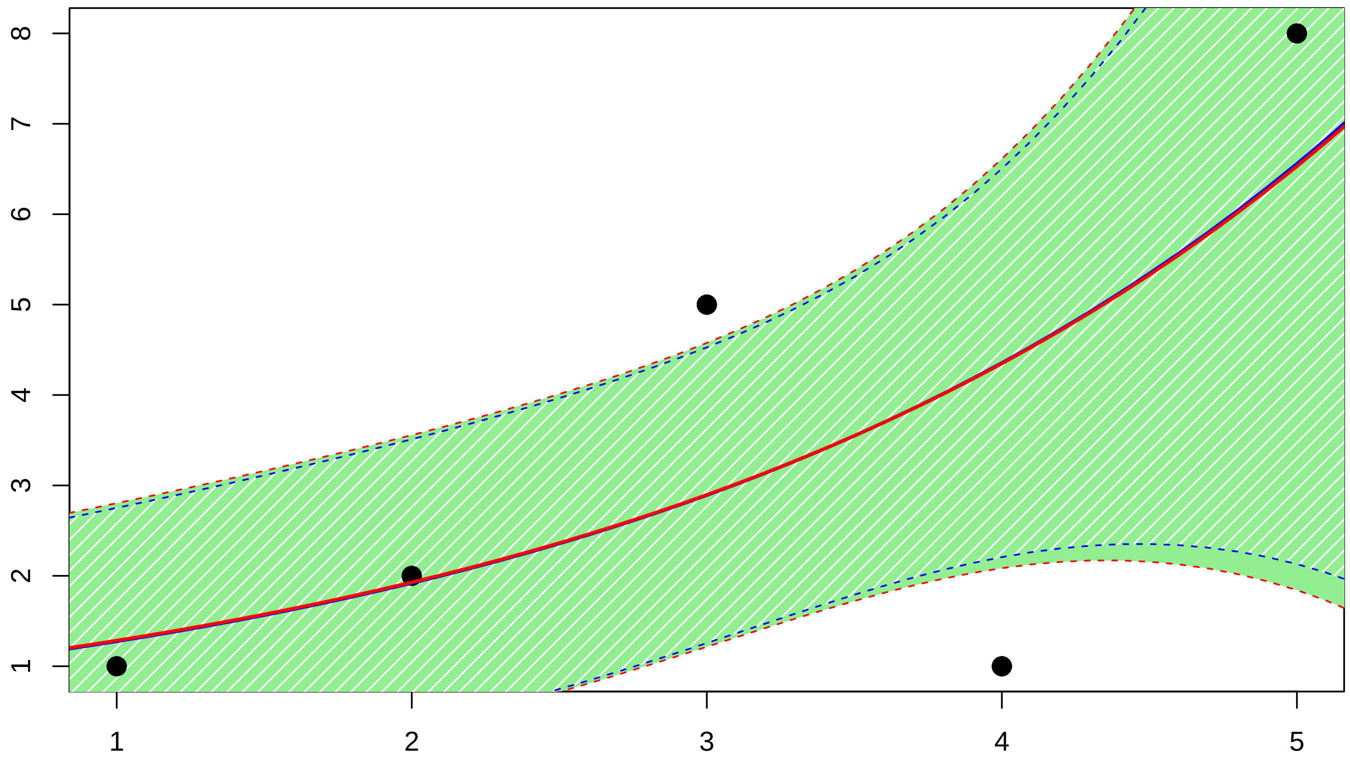
Les régressions pour des données de comptage

```
> regression <- glm(y~x,family=negative.binomial(1))
```



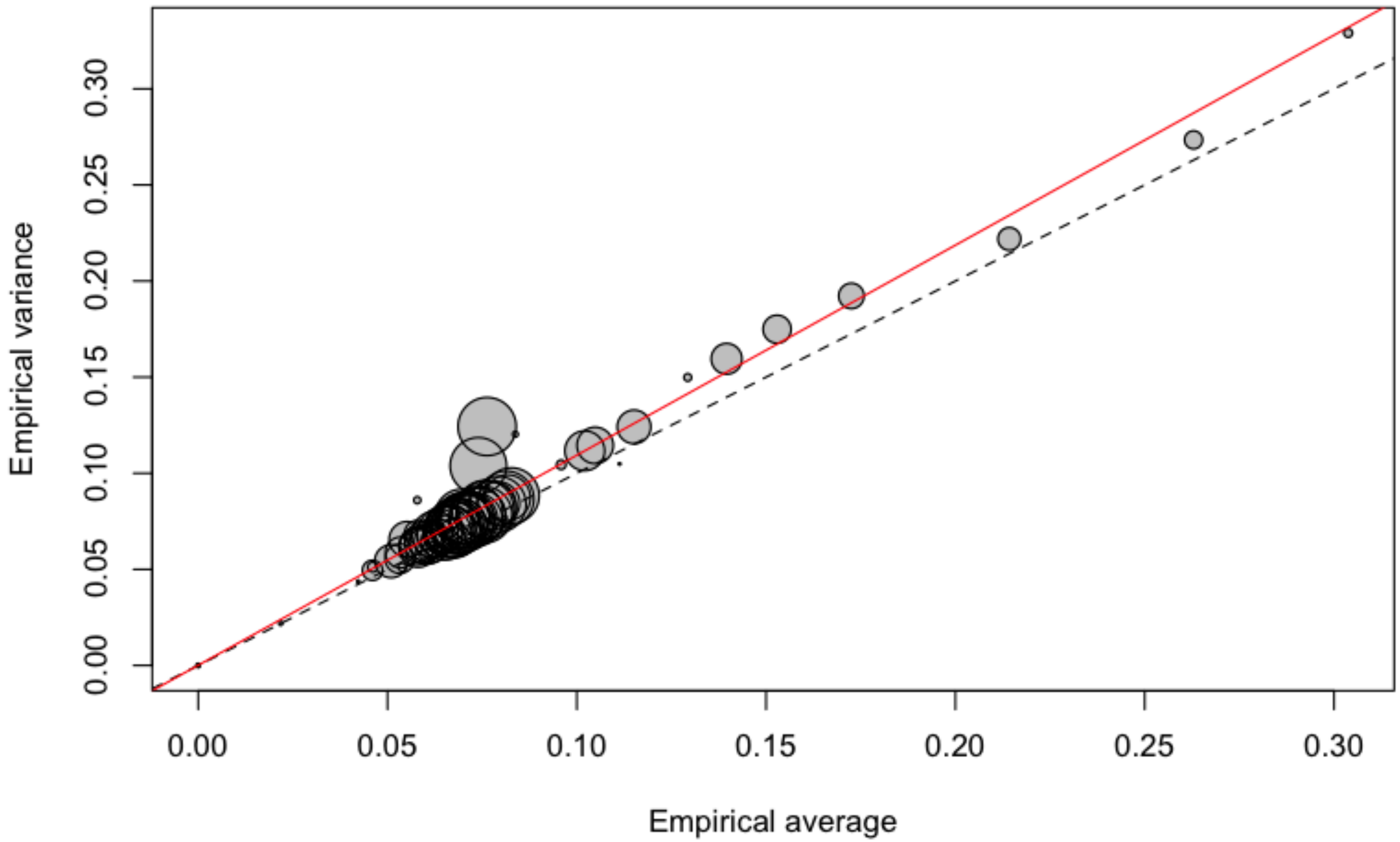
Les régressions pour des données de comptage

```
> regression <- glm.nb(y~x)
```



Test(s) de surdispersion

Sur l'exemple évoqué au début, on peut ajuster une droite de régression passant par l'origine



Test(s) de surdispersion

```
>  library(AER)
>  variance=as.vector(MV[,2])
>  moyenne=as.vector(MV[,1])
>  regression=lm(variance~0+moyenne,weight=as.vector(exposition))
>  summary(regression)

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
moyenne  1.09308    0.01171   93.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007593 on 82 degrees of freedom
Multiple R-squared: 0.9907, Adjusted R-squared: 0.9906
F-statistic:  8714 on 1 and 82 DF,  p-value: < 2.2e-16
```

Test(s) de surdispersion

On peut faire un test pour voir si la pente vaut 1, ou pas,

```
>  linearHypothesis(regression,"moyenne=1")
Hypothesis:
moyenne = 1

Model 1: restricted model
Model 2: variance ~ 0 + moyenne

  Res.Df       RSS Df Sum of Sq     F    Pr(>F)
1     83 0.0083700
2     82 0.0047272  1 0.0036428 63.19 8.807e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sinon, en utilisant la famille quasi-Poisson, on peut aussi faire un test pour tester si φ vaut 1, ou pas.

Test(s) de surdispersion

Pour cela, on note que la surdispersion correspond à une hétérogénéité résiduelle, c'est à dire un effet aléatoire. Par exemple on peut supposer que

$$(Y|\boldsymbol{X}=\boldsymbol{X},\boldsymbol{Z}=\boldsymbol{z})\sim\mathcal{P}(\exp[\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}+\boldsymbol{z}^\mathsf{T}\boldsymbol{\alpha}])$$

de telle sorte que si $u=\boldsymbol{z}^\mathsf{T}\boldsymbol{\alpha}-\mathbb{E}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{\alpha}|\boldsymbol{X}=\boldsymbol{X})$, alors

$$(Y|\boldsymbol{X}=\boldsymbol{X},\boldsymbol{Z}=\boldsymbol{z})\sim\mathcal{P}(\exp[\boldsymbol{X}^\mathsf{T}\boldsymbol{\gamma}+u])$$

On a un modèle dit à effets fixes, au sens où

$$(Y|\boldsymbol{X}=\boldsymbol{X})\sim\mathcal{P}(\exp[\boldsymbol{X}^\mathsf{T}\boldsymbol{\gamma}+U]),$$

où $U=\boldsymbol{Z}^\mathsf{T}\boldsymbol{\alpha}-\mathbb{E}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{\alpha}|\boldsymbol{X}=\boldsymbol{X})$. Par exemple, si on suppose que $U\sim\gamma(\alpha,\alpha)$, i.e. d'espérance 1 et de variance $\sigma^2=\alpha^{-1}$, alors

$$(Y|U=u)\sim\mathcal{P}(\lambda u)\text{ où }\lambda=\exp[\boldsymbol{X}^\mathsf{T}\boldsymbol{\gamma}],$$

Test(s) de surdispersion

de telle sorte que $\mathbb{E}(Y|U=u)=\text{Var}(Y|U=u)$. Mais si on regarde la loi nonconditionnelle, $\mathbb{E}(Y)=\lambda$ alors que

$$\text{Var}(Y)=\text{Var}(\mathbb{E}[Y|U])+\mathbb{E}(\text{Var}(Y|))=\lambda+\lambda^2\sigma^2.$$

On peut alors proposer un test de la forme suivante: on suppose que

$$\text{Var}(Y|\boldsymbol{X}=\boldsymbol{x})=\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})+\tau\cdot\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})^2,$$

on on cherche à tester

$$H_0:\tau=0 \text{ contre } \tau>0.$$

Parmi les statistiques de test classique, on pourra considérer

$$T=\frac{\sum_{i=1}^n[(Y_i-\widehat{\mu}_i)^2-Y_i]}{\sqrt{2\sum_{i=1}^n\widehat{\mu}_i^2}}$$

qui suit, sous H_0, une loi normale centrée réduite.

```
> regquasipoisson <- glm(nbre~bs(ageconducteur)+carburant+ offset(log
    (exposition)),
+ data=baseFREQ,family=quasipoisson)
> summary(regquasipoisson)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.8130     0.1055  -7.706 1.32e-14 ***
bs(ageconducteur)1   -0.9984     0.3388  -2.947  0.00321 **
bs(ageconducteur)2   -0.2508     0.3426  -0.732  0.46406
bs(ageconducteur)3   -1.1986     0.5307  -2.258  0.02393 *
carburantE           -0.2664     0.0381  -6.992 2.74e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 2.196859)

    Null deviance: 28271  on 49999  degrees of freedom
```

```
Residual deviance: 28024  on 49995  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6

> (summary(regquasipoisson)$dispersion)
[1] 2.196859
```

Le test est programmé de la manière suivante

```
> library(AER)
> regpoisson <- glm(nbre~bs(ageconducteur)+carburant+ offset(log(
    exposition)),
+ data=baseFREQ,family=poisson)
> dispersiontest(regpoisson)

  Overdispersion test

data:  regpoisson
```

```
z = 3.1929, p-value = 0.0007042
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  1.754662
```

Modèles à inflation de zéros

Un modèle dit à inflation de zéros ou zero inflated) est un mélange entre une masse en 0 et un modèle classique de comptage

Pour modéliser la probabilité de ne pas déclarer un sinistre (surpoids en 0), considérons un modèle logistique

$$\pi_i = \frac{\exp[\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}]}{1+\exp[\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}]}$$

Pour le modèle de comptage, $p_i(k)$ est la probabilité que l'individu i ait k sinistres

Alors

$$\mathbb{P}(N_i = k) = \begin{cases} \pi_i + [1-\pi_i]\cdot p_i(0) \text{ si } k = 0, \\ [1-\pi_i]\cdot p_i(k) \text{ si } k = 1, 2, \cdots \end{cases}$$

Modèles à inflation de zéros

Si p_i correspond à un modèle Poissonnien (de moyenne λ_i), alors

$$\mathbb{E}(N_i) = [1-\pi_i]\lambda_i \text{ et } \mathrm{Var}(N_i) = \pi_i\lambda_i + \pi_i\lambda_i^2[1-\pi_i].$$

La `library(gamlss)` propose la fonction `ZIP` (pour zero inflated Poisson), et `ZINBI` lorsque p_i correspond à une loi binomiale négative.

La `library(pscl)` propose également une fonction `zeroinfl` plus simple d'utilisation, proposant aussi bien un modèle de Poisson qu'un modèle binomial négatif

Remarque: Il existe aussi des modèles dits zero adapted, où l'on suppose que

$$\mathbb{P}(N_i = k) = \begin{cases} \pi_i \text{ si } k = 0, \\ [1-\pi_i] \cdot \dfrac{p_i(k)}{1-p_i(0)} \text{ si } k = 1, 2, \cdots \end{cases}$$

Dans `library(gamlss)` le modèle précédant dans le cas d'une loi de Poisson est `ZAP`, avec aussi une loi `ZANBI`.

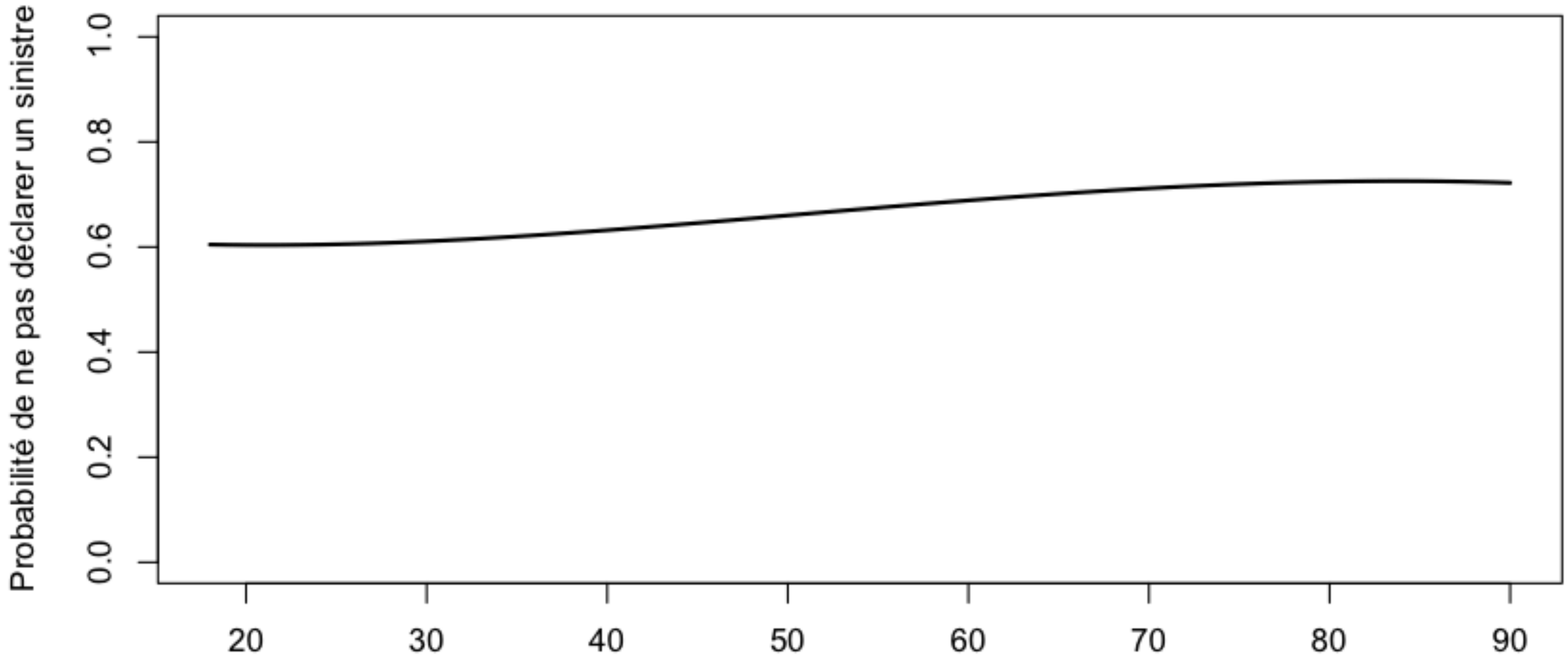
Modèles à inflation de zéros

```
> library(pscl)
> regNZI <- glm(nbre~bs(ageconducteur,5)+offset(log(exposition)),
+ data=baseFREQ,family=poisson(link="log"))
> regZI <- zeroinfl(nbre~bs(ageconducteur) |
+ bs(ageconducteur),offset=log(exposition),
+ data = baseFREQ,dist = "poisson",link="logit")
```

On peut s'intéresser plus particulièrement à l'impact de l'âge sur la probabilité de ne pas déclarer de sinistres (correspondant au paramètre de la loi binomiale)

```
> age<-data.frame(ageconducteur=18:90,exposition=1)
> pred0 <- predict(regZI,newdata=age,type="zero")
> plot(age$ageconducteur,pred0,type="l",xlab="",lwd=2,
+ ylim=c(0,1),ylab="Probabilite de ne pas declarer un sinistre")
```

Modèles à inflation de zéros



La principale explication avancée - en France - pour la non déclaration de sinistre est l'existence du système bonus-malus.

```
> regZIbm <- zeroinfl(nbre~1 |
+ bs(bonus),offset=log(exposition),
+ data = baseFREQ,dist = "poisson",link="logit")
```

```
> B <- data.frame(bonus=50:200,exposition=1)
> pred0 <- predict(regZIbm,newdata=B,type="zero")
> plot(age$ageconducteur,pred0,type="l",xlab="",lwd=2,
+ ylim=c(0,1),ylab="Probabilite de ne pas declarer un sinistre")
```