Actuariat de l'Assurance Non-Vie # 9

A. Charpentier (Université de Rennes 1)

ENSAE 2017/2018

credit: Arnold Odermatt

Fourre-Tout sur la Tarification

- modèle collectif vs. modèle individuel
- cas de la grande dimension
- choix de variables
- choix de modèles

Modèle *individuel* ou modèle *collectif* ? La loi Tweedie

Consider a Tweedie distribution, with variance function power $p \in (1,2)$, mean μ and scale parameter ϕ, then it is a compound Poisson model,

- $N \sim \mathcal{P}(\lambda)$ with $\lambda = \dfrac{\phi\mu^{2-p}}{2-p}$
- $Y_i \sim \mathcal{G}(\alpha,\beta)$ with $\alpha = -\dfrac{p-2}{p-1}$ and $\beta = \dfrac{\phi\mu^{1-p}}{p-1}$

Consversely, consider a compound Poisson model $N \sim \mathcal{P}(\lambda)$ and $Y_i \sim \mathcal{G}(\alpha,\beta)$, then

- variance function power is $p = \dfrac{\alpha+2}{\alpha+1}$
- mean is $\mu = \dfrac{\lambda\alpha}{\beta}$
- scale parameter is $\phi = \dfrac{[\lambda\alpha]^{\frac{\alpha+2}{\alpha+1}-1}\beta^{2-\frac{\alpha+2}{\alpha+1}}}{\alpha+1}$

Modèle *individuel* ou modèle *collectif* ? La régression Tweedie

In the context of regression

$$N_i \sim \mathcal{P}(\lambda_i) \text{ with } \lambda_i = \exp[\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_\lambda]$$

$$Y_{j,i} \sim \mathcal{G}(\mu_i, \phi) \text{ with } \mu_i = \exp[\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_\mu]$$

Then $S_i = Y_{1,i} + \cdots + Y_{N,i}$ has a Tweedie distribution

- variance function power is $p = \dfrac{\phi+2}{\phi+1}$
- mean is $\lambda_i\mu_i$
- scale parameter is $\dfrac{\lambda_i^{\frac{1}{\phi+1}-1}}{\mu_i^{\frac{\phi}{\phi+1}}}\left(\dfrac{\phi}{1+\phi}\right)$

There are $1 + 2\dim(\boldsymbol{X})$ degrees of freedom.

Modèle *individuel* ou modèle *collectif* ? La régression Tweedie

Remark Note that the scale parameter should not depend on i.

A Tweedie regression is

- variance function power is $p \in (1,2)$
- mean is $\mu_i = \exp[\boldsymbol{X}_i^\top \boldsymbol{\beta}_{\text{Tweedie}}]$
- scale parameter is ϕ

There are $2 + \dim(\boldsymbol{X})$ degrees of freedom.

Double Modèle Frq́uence - Coût Individuel

Considérons les bases suivantes, en RC, pour la fréquence

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

pour les coûts individuels

```
> sinistre_RC = sinistre[(sinistre$garantie=="1RC")&(sinistre$cout>0)
    ,]
> sinistre_RC = merge(sinistre_RC,contrat)
```

et pour les co ûts agrégés par police

```
> agg_RC = aggregate(sinistre_RC$cout, by=list(sinistre_RC$nocontrat)
    , FUN='sum')
> names(agg_RC)=c('nocontrat','cout_RC')
> global_RC = merge(contrat, agg_RC, all.x=TRUE)
> global_RC$cout_RC[is.na(global_DO$cout_RC)]=0
```

Double Modèle Frq́uence - Coût Individuel

```
> library(splines)
> reg_f = glm(nb_RC~zone+bs(ageconducteur)+carburant, offset=log(
    exposition),data=freq,family=poisson)
> reg_c = glm(cout~zone+bs(ageconducteur)+carburant, data=sinistre_RC
    ,family=Gamma(link="log"))
```

Simple Modèle Coût par Police

```
> library(tweedie)
> library(statmod)
> reg_a = glm(cout_RC~zone+bs(ageconducteur)+carburant, offset=log(
    exposition),data=global_RC,family=tweedie(var.power=1.5, link.
    power=0))
```

Comparaison des primes

```
> freq2 = freq
> freq2$exposition = 1
> P_f = predict(reg_f,newdata=freq2,type="response")
> P_c = predict(reg_c,newdata=freq2,type="response")
prime1 = P_f*P_c
```

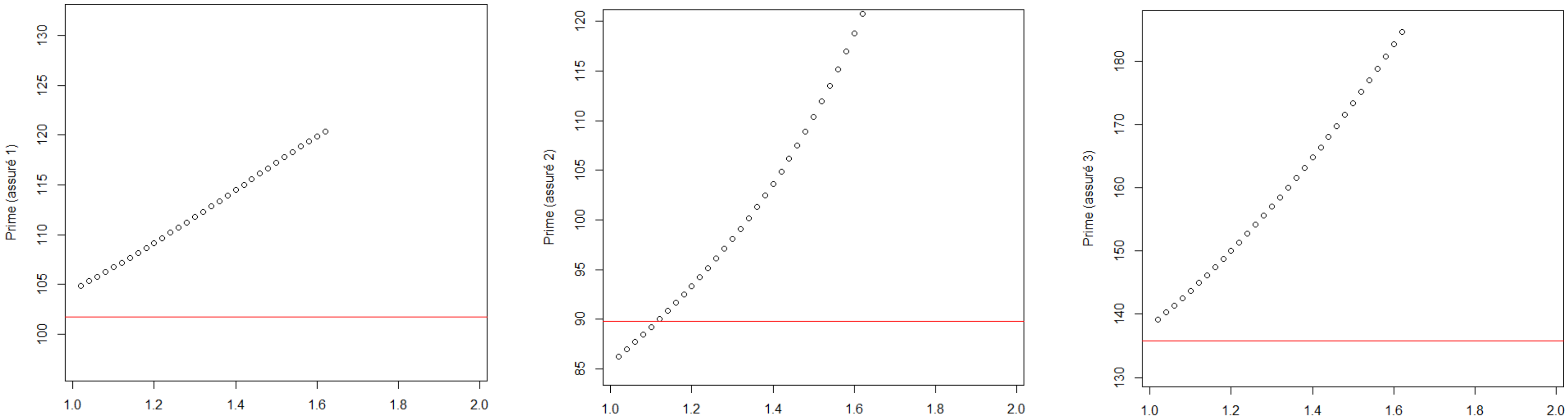
```
> k = 1.5
> reg_a = glm(cout_DO~zone+bs(ageconducteur)+carburant,  offset=log(
    exposition),data=global_DO,family=tweedie(var.power=k, link.power
    =0))
> prime2 = predict(reg_a,newdata=freq2,type="response")
```

```
> arrows(1:100,prime1[1:100],1:100,prime2[1:100],length=.1)
```

Impact du degré Tweedie sur les Primes Pures

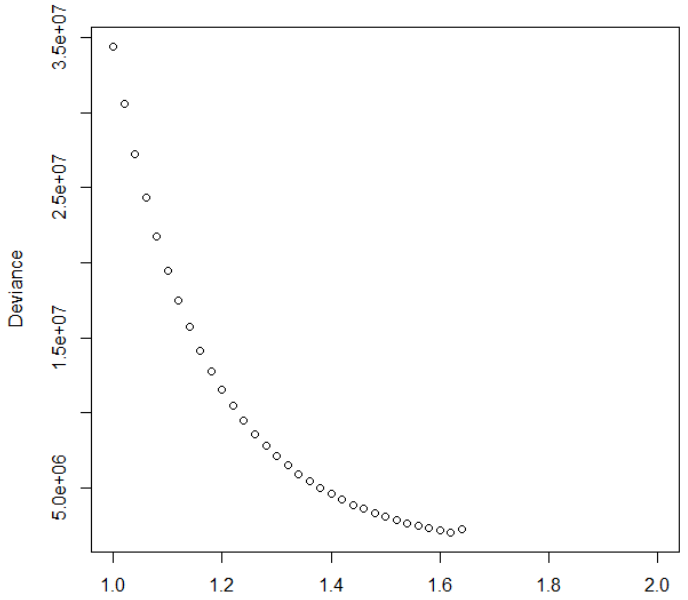
Impact du degré Tweedie sur les Primes Pures

Comparaison des primes pures, assurés no1, no2 et no 3 (DO)



‘Optimisation’ du Paramètre Tweedie

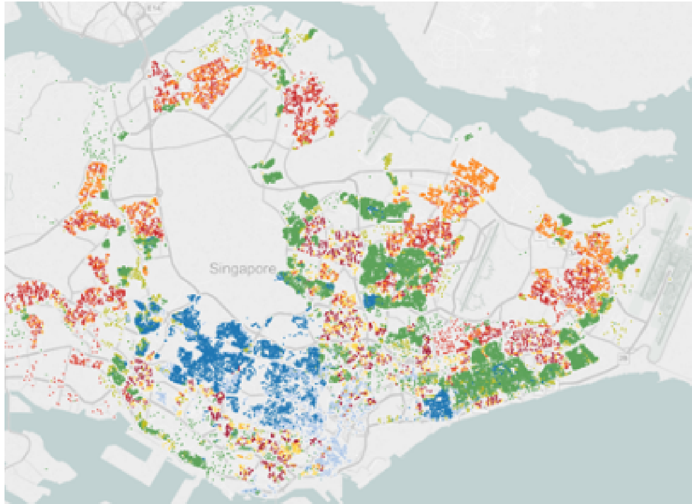
```
> dev = function(k){
+ reg = glm(cout_RC~zone+bs(ageconducteur)+
    carburant, data=global_RC, family=
    tweedie(var.power=k, link.power=0),
    offset=log(exposition))
+ reg$deviance
+ }
```

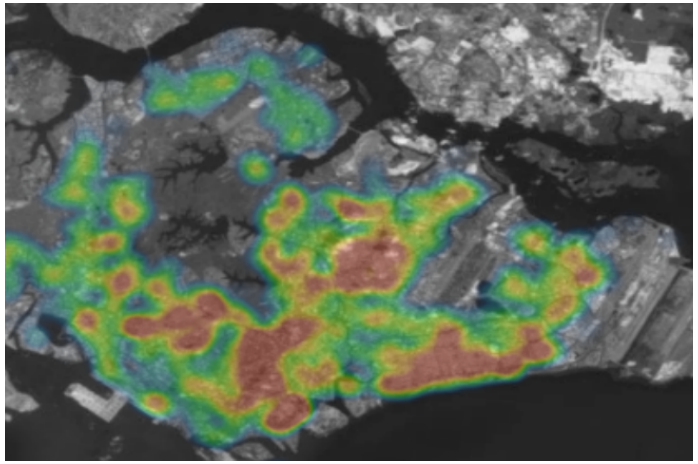


Tarification et données massives (*Big Data*)

Problèmes classiques avec des données massives

- beaucoup de variables explicatives, k grand, $\boldsymbol{X}^\top\boldsymbol{X}$ peut-être non inversible
- gros volumes de données, e.g. données télématiques
- données non quantitatives, e.g. texte, localisation, etc.





La fascination pour les estimateurs sans biais

En statistique mathématique, on aime les estimateurs sans biais car ils ont plusieurs propriétés intéressantes. Mais ne peut-on pas considérer des estimateurs biaisés, potentiellement meilleurs ?

Consider a sample, i.i.d., $\{y_1, \cdots, y_n\}$ with distribution $\mathcal{N}(\mu, \sigma^2)$. Define $\widehat{\theta} = \alpha\overline{Y}$. What is the optimal $\alpha^\star$ to get the best estimator of μ ?

- bias: $\text{bias}\left(\widehat{\theta}\right) = \mathbb{E}\left(\widehat{\theta}\right) - \mu = (\alpha - 1)\mu$
- variance: $\text{Var}\left(\widehat{\theta}\right) = \dfrac{\alpha^2\sigma^2}{n}$
- mse: $\text{mse}\left(\widehat{\theta}\right) = (\alpha - 1)^2\mu^2 + \dfrac{\alpha^2\sigma^2}{n}$

The optimal value is $\alpha^\star = \dfrac{\mu^2}{\mu^2 + \dfrac{\sigma^2}{n}} < 1.$

Linear Model

Consider some linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ for all $i = 1, \cdots, n$.

Assume that ε_i are i.i.d. with $\mathbb{E}(\varepsilon) = 0$ (and finite variance). Write

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\boldsymbol{y},n\times 1} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix}}_{\boldsymbol{X},n\times(k+1)} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta},(k+1)\times 1} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon},n\times 1}.$$

Assuming $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\mathbb{I})$, the maximum likelihood estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\{\|\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{\beta}\|_{\ell_2}\} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

... under the assumtption that $\boldsymbol{X}^\top\boldsymbol{X}$ is a full-rank matrix.

What if $\boldsymbol{X}_i^\top\boldsymbol{X}$ cannot be inverted? Then $\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^\top\boldsymbol{X}]^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ does not exist, but $\widehat{\boldsymbol{\beta}}_\lambda = [\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I}]^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ always exist if $\lambda > 0$.

Ridge Regression

The estimator $\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I}]^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ is the Ridge estimate obtained as solution of

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\left\{\sum_{i=1}^n [y_i - \beta_0 - \boldsymbol{x}_i^\top\boldsymbol{\beta}]^2 + \lambda \underbrace{\|\boldsymbol{\beta}\|_{\ell_2}}_{\mathbf{1}^\top\boldsymbol{\beta}^2}\right\}$$

for some tuning parameter λ. One can also write

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_2}\leq s}{\text{argmin}}\{\|\boldsymbol{Y} - \boldsymbol{X}^\top\boldsymbol{\beta}\|_{\ell_2}\}$$

Remark Note that we solve $\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\{\text{objective}(\boldsymbol{\beta})\}$ where

$$\text{objective}(\boldsymbol{\beta}) = \underbrace{\mathcal{L}(\boldsymbol{\beta})}_{\text{training loss}} + \underbrace{\mathcal{R}(\boldsymbol{\beta})}_{\text{regularization}}$$

Going further on sparcity issues

In severall applications, k can be (very) large, but a lot of features are just noise: $\beta_j = 0$ for many j's. Let s denote the number of relevent features, with $s << k$, cf Hastie, Tibshirani & Wainwright (2015),

$$s = \text{card}\{\mathcal{S}\} \text{ where } \mathcal{S} = \{j; \beta_j \neq 0\}$$

The model is now $y = \boldsymbol{X}_{\mathcal{S}}^{\mathsf{T}}\boldsymbol{\beta}_{\mathcal{S}} + \varepsilon$, where $\boldsymbol{X}_{\mathcal{S}}^{\mathsf{T}}\boldsymbol{X}_{\mathcal{S}}$ is a full rank matrix.

Going further on sparcity issues

Define $\|\boldsymbol{a}\|_{\ell_0} = \sum \mathbf{1}(|a_i| > 0)$. Ici $\dim(\boldsymbol{\beta}) = s$.

We wish we could solve

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_0} \leq s}{\text{argmin}} \{\|\boldsymbol{Y} - \boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}\|_{\ell_2}\}$$

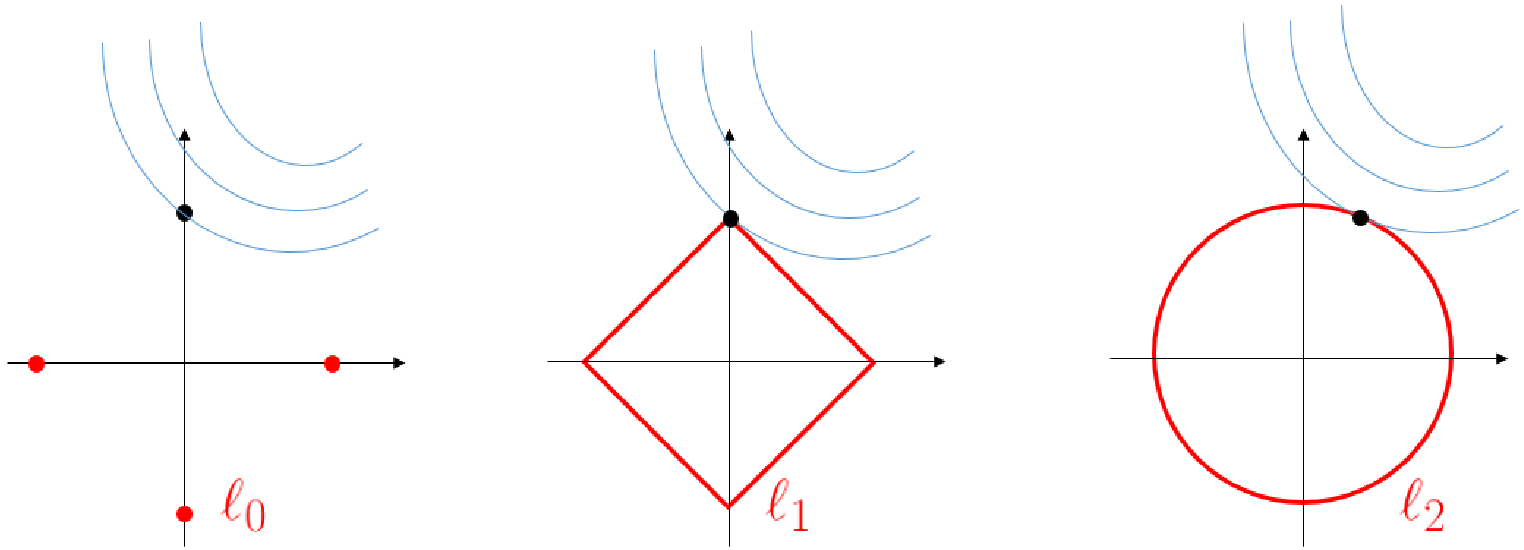
Problem: it is usually not possible to describe all possible constraints, since $\binom{s}{k}$ coefficients should be chosen here (with k (very) large).

Idea: solve the dual problem

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta};\|\boldsymbol{Y}-\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}\|_{\ell_2} \leq h}{\text{argmin}} \{\|\boldsymbol{\beta}\|_{\ell_0}\}$$

where we might convexify the ℓ_0 norm, $\|\cdot\|_{\ell_0}$.

Regularization ℓ_0, ℓ_1 et ℓ_2



Going further on sparcity issues

On $[-1,+1]^k$, the convex hull of $\|\boldsymbol{\beta}\|_{\ell_0}$ is $\|\boldsymbol{\beta}\|_{\ell_1}$

On $[-a,+a]^k$, the convex hull of $\|\boldsymbol{\beta}\|_{\ell_0}$ is $a^{-1}\|\boldsymbol{\beta}\|_{\ell_1}$

Hence,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_1}\leq\tilde{s}}{\text{argmin}}\ \{\|\boldsymbol{Y}-\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}\|_{\ell_2}\}$$

is equivalent (Kuhn-Tucker theorem) to the Lagragian optimization problem

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\{\|\boldsymbol{Y}-\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}\|_{\ell_2}+\lambda\|\boldsymbol{\beta}\|_{\ell_1}\}$$

LASSO *Least Absolute Shrinkage and Selection Operator*

$$\widehat{\boldsymbol{\beta}} \in \text{argmin}\{\|\boldsymbol{Y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\|_{\ell_2} + \lambda\|\boldsymbol{\beta}\|_{\ell_1}\}$$

is a convex problem (several algorithms⋆), but not strictly convex (no unicity of the minimum). Nevertheless, predictions $\widehat{\boldsymbol{y}} = \boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}}$ are unique

⋆ MM, minimize majorization, coordinate descent Hunter (2003).

Optimal LASSO Penalty

Use cross validation, e.g. K-fold,

$$\widehat{\boldsymbol{\beta}}_{(-k)}(\lambda) = \text{argmin}\left\{\sum_{i\notin\mathcal{I}_k}[y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}]^2 + \lambda\|\boldsymbol{\beta}\|\right\}$$

then compute the sum of the squared errors,

$$Q_k(\lambda) = \sum_{i\in\mathcal{I}_k}[y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{(-k)}(\lambda)]^2$$

and finally solve

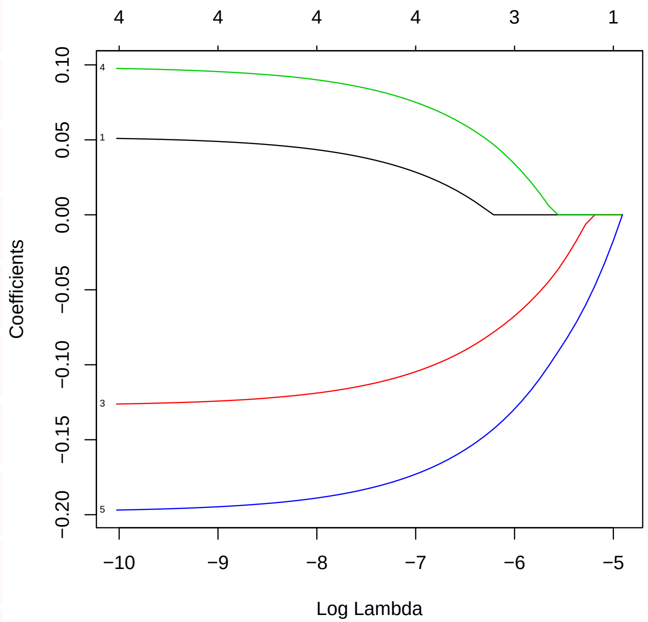
$$\lambda^\star = \text{argmin}\left\{\overline{Q}(\lambda) = \frac{1}{K}\sum_k Q_k(\lambda)\right\}$$

Note that this might overfit, so Hastie, Tibshiriani & Friedman (2009) suggest the largest λ such that

$$\overline{Q}(\lambda) \leq \overline{Q}(\lambda^\star) + \text{se}[\lambda^\star] \text{ with } \text{se}[\lambda]^2 = \frac{1}{K^2}\sum_{k=1}^K[Q_k(\lambda) - \overline{Q}(\lambda)]^2$$

LASSO, Fréquence RC

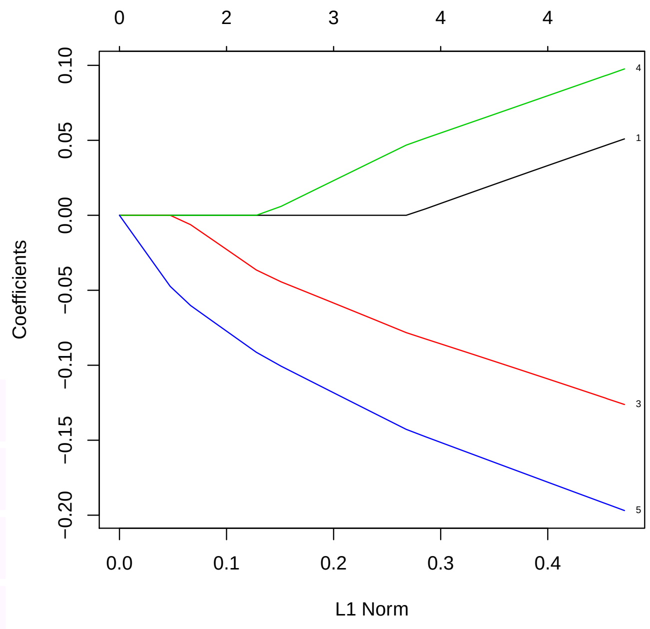
```
> freq = merge(contrat,nombre_RC)
> freq = merge(freq,nombre_DO)
> freq[,10]=as.factor(freq[,10])
> mx=cbind(freq[,c(4,5,6)],freq[,9]=="D",
    freq[,3]%in%c("A","B","C"))
> colnames(mx)=c(names(freq)[c(4,5,6)],"
    diesel","zone")
> for(i in 1:ncol(mx)) mx[,i]=(mx[,i]-mean(
    mx[,i]))/sd(mx[,i])
> names(mx)
[1]   puissance ageve hicule ageconducteur
      diesel        zone
> library(glmnet)
> fit = glmnet(x=as.matrix(mx), y=freq[,11],
     offset=log(freq[,2]), family = "poisson
    ")
> plot(fit, xvar="lambda", label=TRUE)
```

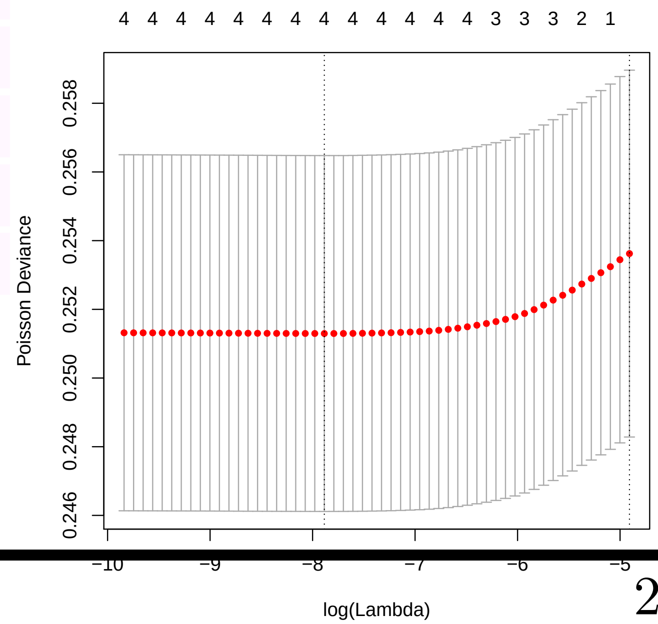


LASSO, Fréquence RC

```
> plot(fit,label=TRUE)
> cvfit = cv.glmnet(x=as.matrix(mx), y=freq
    [,11], offset=log(freq[,2]),family = "
    poisson")
> plot(cvfit)
> cvfit$lambda.min
[1] 0.0002845703
> log(cvfit$lambda.min)
[1] -8.16453
```

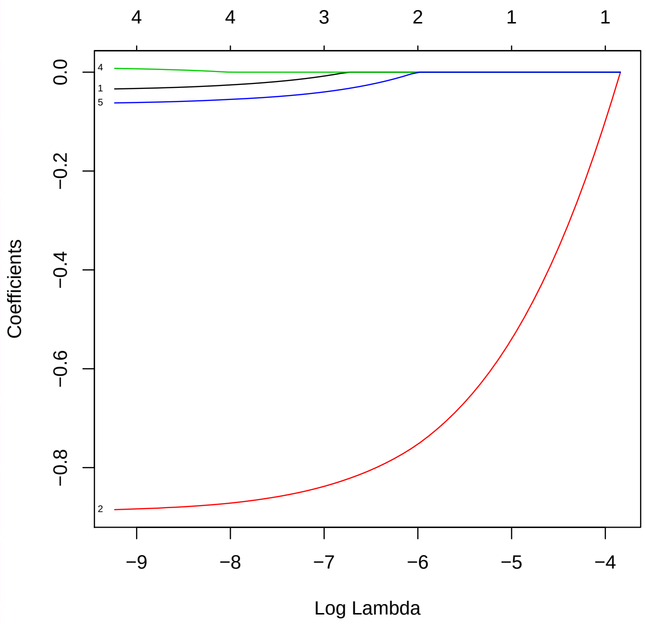
- Cross validation curve + error bars





```
> freq = merge(contrat,nombre_RC)
> freq = merge(freq,nombre_DO)
> freq[,10]=as.factor(freq[,10])
> mx=cbind(freq[,c(4,5,6)],freq[,9]=="D",
   freq[,3]%in%c("A","B","C"))
> colnames(mx)=c(names(freq)[c(4,5,6)],"
   diesel","zone")
> for(i in 1:ncol(mx)) mx[,i]=(mx[,i]-mean(
   mx[,i]))/sd(mx[,i])
> names(mx)
[1]  puissance agevehicule ageconducteur
     diesel       zone
> library(glmnet)
> fit = glmnet(x=as.matrix(mx), y=freq[,12],
    offset=log(freq[,2]), family = "poisson
   ")
> plot(fit, xvar="lambda", label=TRUE)
```

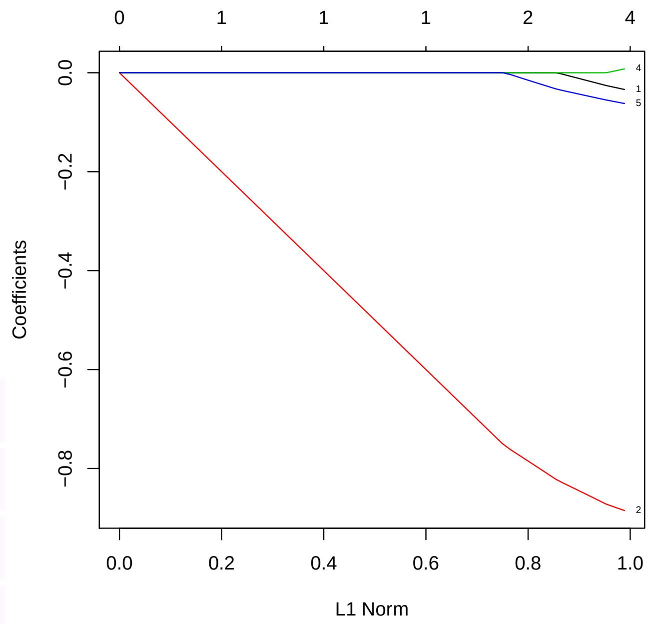
LASSO, Fréquence DO

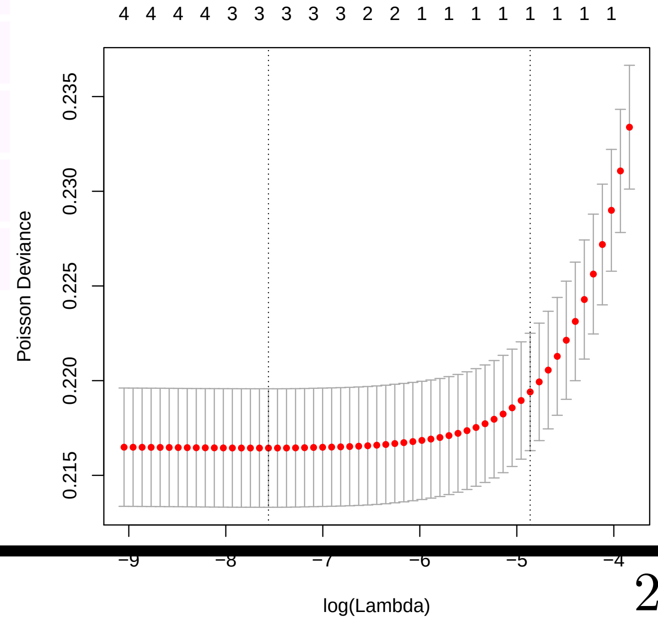


LASSO, Fréquence DO

```
> plot(fit,label=TRUE)
> cvfit = cv.glmnet(x=as.matrix(mx), y=freq
    [,12], offset=log(freq[,2]),family = "
    poisson")
> plot(cvfit)
> cvfit$lambda.min
[1] 0.0004744917
> log(cvfit$lambda.min)
[1] -7.653266
```

- Cross validation curve + error bars





Model Selection and Gini/Lorentz (on incomes)

Consider an ordered sample $\{y_1, \cdots, y_n\}$, then Lorenz curve is

$$\{F_i, L_i\} \text{ with } F_i = \frac{i}{n} \text{ and } L_i = \frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{n} y_j}$$

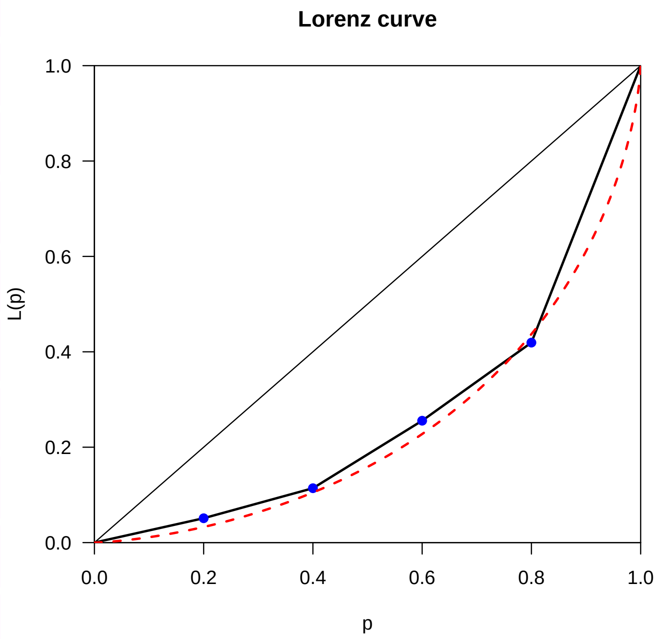
The theoretical curve, given a distribution F, is

$$u \mapsto L(u) = \frac{\int_{-\infty}^{F^{-1}(u)} t dF(t)}{\int_{-\infty}^{+\infty} t dF(t)}$$

see Gastwirth (1972, econpapers.repec.org)

Model Selection and Gini/Lorentz (on incomes)

```
> library(ineq)
> set.seed(1)
> (x<-sort(rlnorm(5,0,1)))
[1] 0.4336018 0.5344838 1.2015872 1.3902836
    4.9297132
> Lc.sim <- Lc(x)
> plot(Lc.sim)
> points((1:4)/5,(cumsum(x)/sum(x))[1:4],pch
    =19,col="blue")
> lines(Lc.lognorm, parameter=1,lty=2)
```

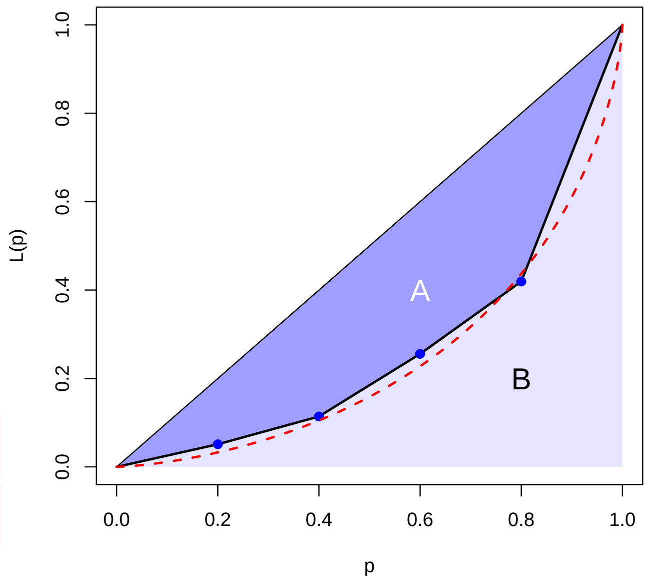


Model Selection and Gini/Lorentz (on incomes)

Gini index is the ratio of the areas $\dfrac{\mathsf{A}}{\mathsf{A}+\mathsf{B}}$. Thus,

$$
\begin{aligned}
G &= \frac{2}{n(n-1)\overline{x}}\sum_{i=1}^{n} i\cdot x_{i:n} - \frac{n+1}{n-1} \\
&= \frac{1}{\mathbb{E}(Y)}\int_0^{\infty} F(y)(1-F(y))dy
\end{aligned}
$$

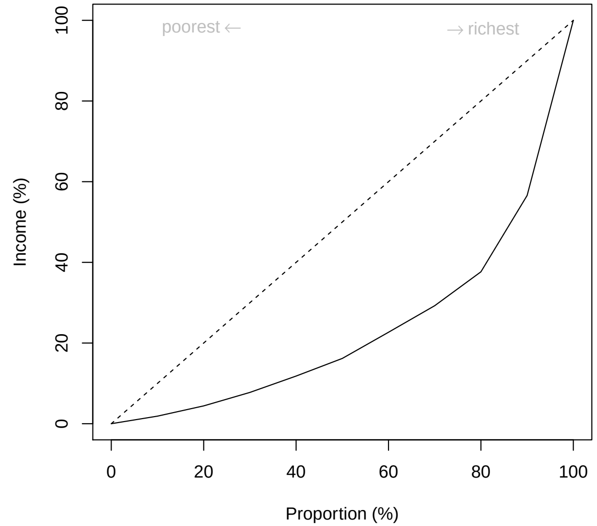
```
> Gini(x)
[1] 0.4640003
```



Comparing Models

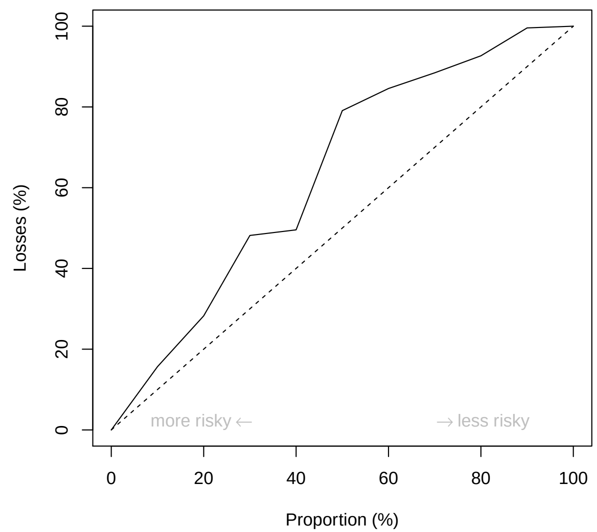
Consider an ordered sample $\{y_1, \cdots, y_n\}$ of incomes, with $y_1 \leq y_2 \leq \cdots \leq y_n$, then Lorenz curve is

$$\{F_i, L_i\} \text{ with } F_i = \frac{i}{n} \text{ and } L_i = \frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{n} y_j}$$

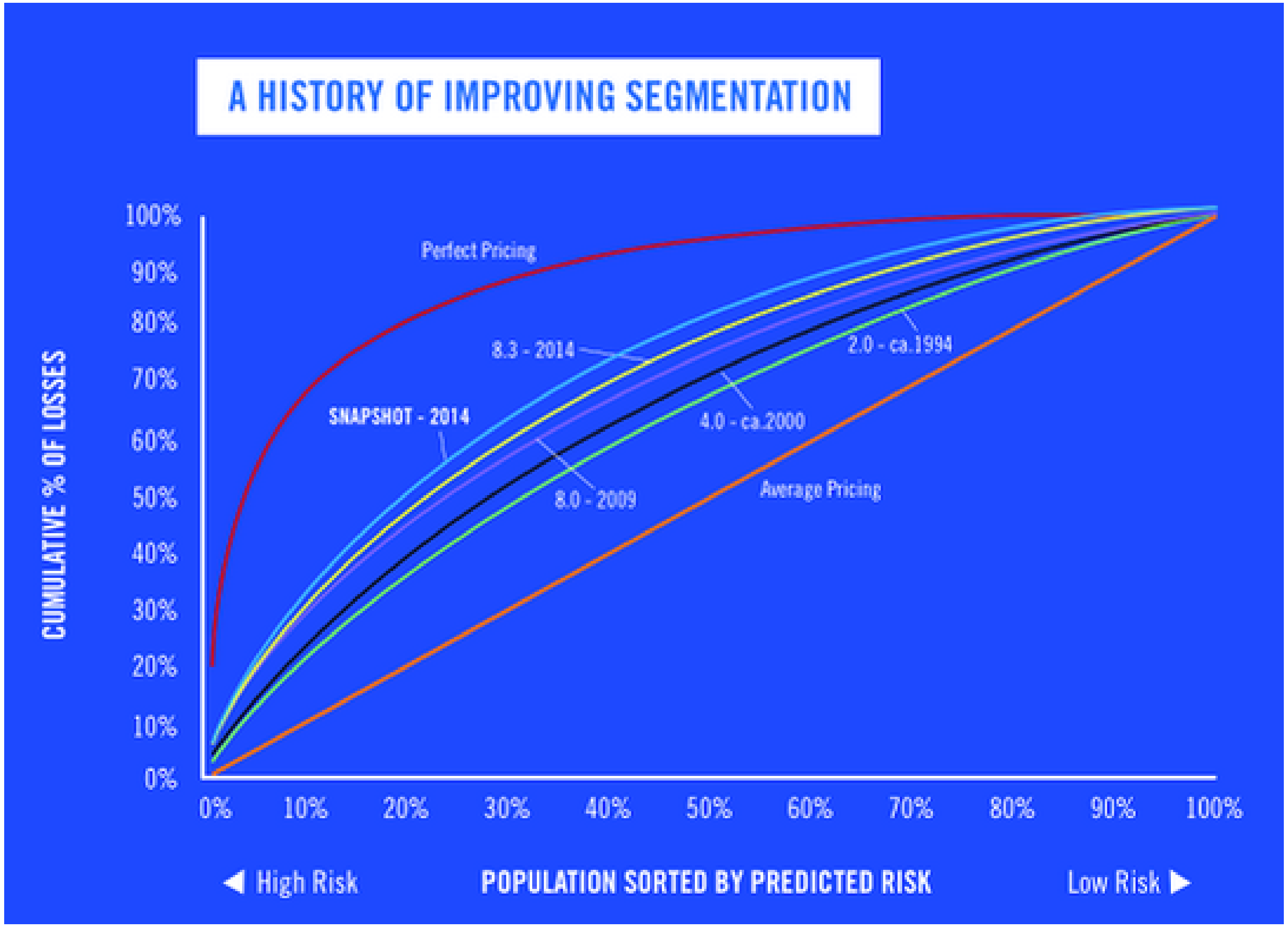


We have observed losses y_i and premiums $\widehat{\pi}(\boldsymbol{x}_i)$. Consider an ordered sample by the model, see Frees, Meyers & Cummins (2014), $\widehat{\pi}(\boldsymbol{x}_1) \geq \widehat{\pi}(\boldsymbol{x}_2) \geq \cdots \geq \widehat{\pi}(\boldsymbol{x}_n)$, then plot

$$\{F_i, L_i\} \text{ with } F_i = \frac{i}{n} \text{ and } L_i = \frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{n} y_j}$$



Choix et comparaison de modèle en tarification



See Frees *et al.* (2010) or Tevet (2013).