Contents
1. Project: Motor vehicle insurance
2. Classical Methods
3. Convex risk minimization
4. Robustness
5. Cont.: Motor vehicle insurance
6. Summary
1. PROJECT: MOTOR VEHICLE INSURANCE

- Verband öffentlicher Versicherer, Düsseldorf, Germany
- non-aggregated data from 15 insurance companies
- 3 GB per year, > 4.5 million customers, > 70 explanatory variables
- primary response: pure premium [year] ⇒ insurance tariffs
- secondary response: probability of having at least 1 claim per year

- complex dependencies, empty cells, missing values, some extreme high costs
**Statistical Objectives**

- claim amount for each customer
- \( x \in \mathbb{R}^p \) explanatory variables

- **Actual premium.** Charged to the customer
  - pure premium + safety loading + administrative costs + desired profit

- **Primary response:** Pure premium. \( E(Y|X = x) \)
  - observed pure premium = \( \frac{\text{sum of individual claim sizes}}{\text{number of days under risk } / 360} \) for each customer

- **Secondary response:** Prob. of claim. \( P(Y > 0|X = x) \)

- **Precision criterion:** MSE. \( E((Y - \hat{Y})^2|X = x) = \text{min}! \)

- **Fair criterion:** Bias. \( E(Y - \hat{Y}|X = x) \approx 0! \)
  - in whole population and in subpopulations
**Characteristic Properties**

- Most of the policy holders have no claim within a year or a certain period.
- The claim sizes are extremely skewed to the right, but there is atom in zero.
- There is a complex, high dimensional dependency structure between variables.
- There are only imprecise values available for some explanatory variables.
- Some claim sizes are only estimates.
- The data sets to be analyzed are huge.
- Extreme high claim amounts are rare events, but contribute enormously to the total sum of all claims.
Exploratory Data Analysis

Pure premium per policy holder per year

total mean \approx 360 \text{ EUR}

<table>
<thead>
<tr>
<th>Pure premium</th>
<th>% obs.</th>
<th>% of total sum</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td></td>
<td></td>
<td>364</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>94.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,2000]</td>
<td>2.2</td>
<td>6.7</td>
<td>1097</td>
<td>1110</td>
</tr>
<tr>
<td>(2000,10000]</td>
<td>2.4</td>
<td>27.1</td>
<td>4156</td>
<td>3496</td>
</tr>
<tr>
<td>(10000,50000]</td>
<td>0.4</td>
<td>19.8</td>
<td>18443</td>
<td>15059</td>
</tr>
<tr>
<td>&gt;50000</td>
<td>0.07</td>
<td>46.4</td>
<td>234621</td>
<td>96417</td>
</tr>
</tbody>
</table>

Max. individual claim size: some million EUR
smoothed prob. for claim
Average cost [EUR]:

- Green: 0–150
- Orange: 350
- Light Green: 200
- Light Gray: 250
- Yellow: 300
- Dark Red: 400
- Red: 450+
2. METHODS

- Classical approach (in Germany): 'Marginal Sum Model' → Poisson-Regression

- Generalized Linear Models (GLIM):
  ▶ $Y_i$ has distribution from exponential family
  ▶ $E(Y_i) = \mu_i = g^{-1}(x_i'\beta)$ and $\text{Var}(Y_i) = \phi \frac{V(g^{-1}(x_i'\beta))}{w_i}$
    
    $g = \text{link function}, V = \text{variance function}$

Special cases:
▶ Poisson: $E(Y_i) = \text{Var}(Y_i)$
▶ Gamma: $V(\mu_i) = \mu_i^2$
▶ Inverse Gaussian: $V(\mu_i) = \mu_i^3$
▶ Negative Binomial: $V(\mu_i) = \mu_i + k\mu_i^2$

- Tweedie’s compound Poisson Model (Smyth & Jørgensen, ’02):
Check of GLIM assumption

QQ plot for GPD

\[
\text{Gamma: } \text{Var} = kE^2 \\
\text{Poisson: } \text{Var} = kE
\]

overdispersion \( k \gg 1 \)
Denote: $Y$ pure premium, $x$ vector of explanatory variables

1: Define additional variable $C$ to classify $Y$, e.g.
   
   $C = 0$ if $Y = 0$ \quad 'no cost' 94.9 \% 
   
   $= 1$ if $Y \in (0, 2000]$ \quad 'low' 2.2 \% 
   
   $= 2$ if $Y \in (2000, 10000]$ \quad 'medium' 2.4 \% 
   
   $= 3$ if $Y \in (10000, 50000]$ \quad 'high' 0.4 \% 
   
   $= 4$ if $Y > 50,000$ \quad 'extreme' 0.07 \%

2: No claims: $E(Y|X = x, C = 0) \equiv 0$ for all $x$.

\[
E(Y|X = x) = P(C > 0|X = x) \cdot E(Y|C > 0, X = x) \\
= P(C > 0|X = x) \cdot \sum_{c=1}^{k+1} P(C = c|C > 0, X = x) \cdot E(Y|C = c, X = x)
\]
\begin{align*}
E(Y|X = x) &= P(C > 0|X = x) \cdot \\
&\sum_{c=1}^{k+1} P(C = c|C > 0, X = x) \cdot E(Y|C = c, X = x)
\end{align*}

Companies have interest in \( P(C = c) \) or \( E(Y|X = x, C = c) \)

Circumvents the problem: most \( y_i = 0 \), but \( P(Y = 0) = 0 \)

Reduction of computation time possible: regression only for 5% of obs.!

Reduction of interactions: different \( x \) for different \( C = c \)

Variable selection: different vectors \( x \) for different \( C \) groups are possible

Different estimation techniques can be used for estimating \( P(C = c|X = x) \) and \( E(Y|X = x, C = c) \), e.g.

\( \rightarrow \) Multinomial logistic regression + Gamma regression

\( \rightarrow \) Robust logistic regr. + semi-parametr. regr.

\( \rightarrow \) Classification trees + semiparametric regression

\( \rightarrow \) Kernel logistic regression + \( \varepsilon \)-support vector regression

\( \rightarrow \) Combination of above pairs + extreme value theory (e.g. GPD)
3. EMPIRICAL RISK MINIMIZATION

Vapnik '98

Data set \((x_i, y_i) \in \mathbb{R}^p \times \{-1, +1\}\). Assume \((X_i, Y_i) \) iid \(\mathbb{P}\), \(\mathbb{P}\) unknown, predictor \(\hat{f}(x)\), classifier \(\text{sign}(\hat{f}(x) + \hat{b})\), loss function \(L(y, f(x) + b)\)

goal: \[
\arg \min_{f, b} \mathbb{E}_\mathbb{P} L(Y, f(X) + b)
\]

reg. emp. \((\hat{f}_n, \lambda, \hat{b}_n, \lambda) = \arg \min_{f \in H, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i) + b) + \lambda \|f\|_H^2\)

risk: where \(L\) convex, \(\lambda > 0\) reg. constant, \(H\) reproducing kernel Hilbert space (RKHS), defined by reproducing kernel \(k\)

\([k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}, k(x, \cdot) \in H, f(x) = \langle f, k(x, \cdot) \rangle]\)

reg. theor. \((f_\mathbb{P}, \lambda, b_\mathbb{P}, \lambda) = \arg \min_{f \in H, b \in \mathbb{R}} \mathbb{E}_\mathbb{P} L(Y, f(X) + b) + \lambda \|f\|_H^2\)

risk: Vapnik '98, Zhang '01, Chr & Steinwart '05: \(L\)-risk consistency
Advantages:

Kernel trick restricts class of functions to broad subset (RKHS) allows very flexible fits
RBF kernel: \( k(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2) \), \( \gamma > 0 \)

Convex loss function avoids computational intractable NP-hard problems

Regularization improves generalization property, avoids overfitting

Software:
many computational tricks

Overview: http://www.kernel-machines.org
Kernel Logistic Regression (KLR)

\[ L(x, y, f) = \ln(1 + \exp(-y[f(x) + b])) \]

smoothly decreasing risk scoring is possible: 

\[ P(Y = 1|X = x) = \frac{1}{1+e^{-(f(x)+b)}} \]

\[
\varepsilon-\text{Support Vector Regression (SVR)} \quad \text{(Vapnik '98)}
\]

\[ L_\varepsilon(x, y, f) = \max\left\{ 0, |y - f(x)| - \varepsilon \right\} \]
4. Robustness

• model assumptions valid?
• some extreme claim sizes are estimates
• imprecise data: driving distance during a year

Goal: \( T(\mathbb{P}) \), but \( T((1 - \varepsilon)\mathbb{P} + \varepsilon\tilde{\mathbb{P}}) \approx T(\mathbb{P}) \)?

\[
T(\mathbb{P}) = (f_\mathbb{P}, \lambda, b_\mathbb{P}, \lambda) = \arg\min_{f \in H, b \in \mathbb{R}} \mathbb{E}_\mathbb{P} L(Y, f(X) + b) + \lambda \| f \|^2_H
\]
Robustness Approaches

Influence function (F. Hampel): Gâteaux-derivative towards Dirac-distr. \( \Delta z \)

\[
IF(z; T, \mathcal{P}) = \lim_{\varepsilon \downarrow 0} \frac{T((1 - \varepsilon)\mathcal{P} + \varepsilon\Delta z) - T(\mathcal{P})}{\varepsilon}
\]

Sensitivity curve (J.W. Tukey): measures impact of one point \( z \)

\[
SC_n(z; T_n) = n(T_n(z_1, \ldots, z_{n-1}, z) - T_{n-1}(z_1, \ldots, z_{n-1}))
\]

\[
SC_n(z; T_n) = \frac{T((1 - \varepsilon_n)\mathcal{P}_{n-1} + \varepsilon_n\Delta z) - T(\mathcal{P}_{n-1})}{\varepsilon_n}, \quad \varepsilon_n = \frac{1}{n}
\]

Maxbias (P.J. Huber): measures worst case behavior in neighborhood

\[
\text{maxbias}(\varepsilon; T, \mathcal{P}) = \sup_{Q \in N_{\varepsilon}(\mathcal{P})} \|T(Q) - T(\mathcal{P})\|
\]

\[
N_{\varepsilon}(\mathcal{P}) = \{Q = (1 - \varepsilon)\mathcal{P} + \varepsilon\tilde{\mathcal{P}}; \tilde{\mathcal{P}} \text{ is distr. on } X \times Y, 0 \leq \varepsilon < \frac{1}{2}\}
\]
Assume for classification or regression problem:

- loss function $L$ is convex, continuous, $L'$ bounded
- $H$ RKHS of a bounded, continuous kernel $k$.

Then these kernel methods have good robustness properties:

- influence function
- sensitivity curve
- bias
- maxbias

are bounded, sometimes even uniformly bounded,

and they are able to “learn” the unknown distribution $\mathbb{P}$.

Details and Proofs: Chr & Steinwart (’04, ’05, ’06)
5. Cont.: motor vehicle insurance

- \( Y \): pure premium [year]
- \( X \): 8 explanatory variables:
  - age of main user, gender of main user,
  - car kept in garage?, driving distance, geographical information,
  - population density, no. of years without claims, strength of engine

- Estimation with (KLR, \( \varepsilon \)-SVR) with RBF kernel using
  \[
  E(Y|X = x) = P(C > 0|X = x) \cdot \sum_{c=1}^{k+1} P(C = c|C > 0, X = x) \cdot E(Y|C = c, X = x)
  \]

- 50% training (n=2.2E6), 25% validation (n=1.1E6), 25% test (n=1.1E6)
Estimated pure premium $E(Y|X=x)$
Estimated pure premium $E(Y|X=x)$

Age of main user

EUR

- Males
- Females
6. Summary

Empirical Risk Minimization:
+ statistics + mathematics + computer sciences
+ companies (insurance, banking, IT, telecommunication)
  ⇒ tariffs, risk scoring, data mining, CRM, CHURN
+ software companies (SAS/Enterprise Miner)

+ good predictions, very flexible

+ applicable also in very high dimensions

+ good robustness properties


[homepages.vub.ac.be/~achristm](http://homepages.vub.ac.be/~achristm)