Generalized Linear Models

Mario V. Wüthrich
RiskLab, ETH Zurich

"Deep Learning with Actuarial Applications in R"
Swiss Association of Actuaries SAA/SAV, Zurich
October 14/15, 2021

Programme SAV Block Course

- Refresher: Generalized Linear Models (THU 9:00-10:30)
- Feed-Forward Neural Networks (THU 13:00-15:00)
- Discrimination-Free Insurance Pricing (THU 17:15-17:45)

- LocalGLMnet (FRI 9:00-10:30)
- Convolutional Neural Networks (FRI 13:00-14:30)
- Wrap Up (FRI 16:00-16:30)

Contents: Generalized Linear Models

- Starting with data
- Exponential dispersion family (EDF)
- Generalized linear models (GLMs)
- Maximum likelihood estimation (MLE)
- Canonical link and the balance property
- Covariate pre-processing / feature engineering
- Parameter selection

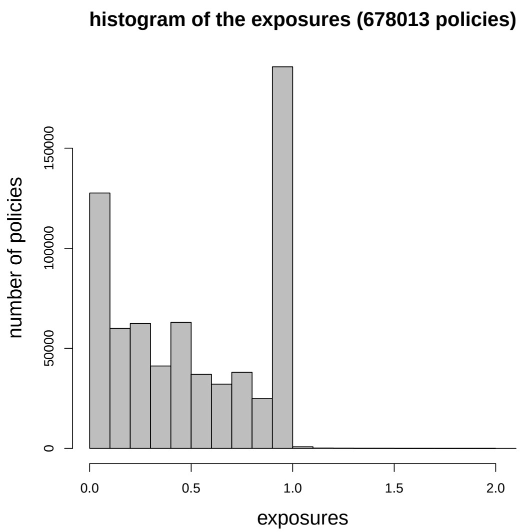
- **Starting with Data**

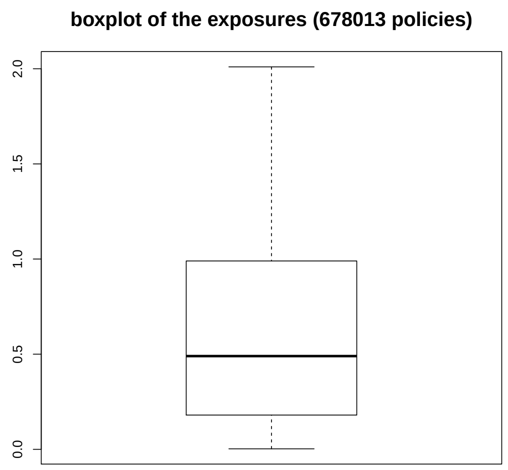
Car Insurance Claims Frequency Data

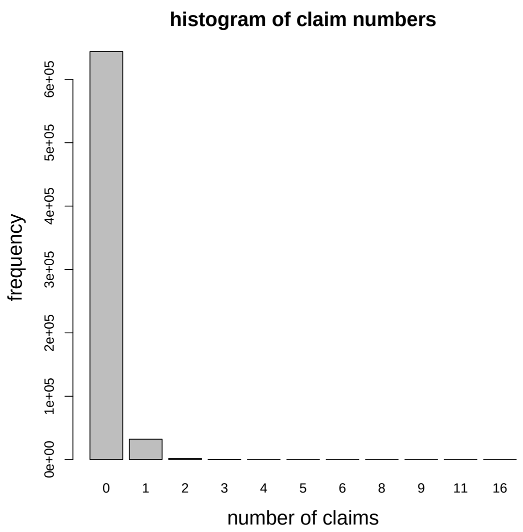
```
'data.frame':   678013 obs. of  12 variables:
 $ IDpol      : num  1 3 5 10 11 13 15 17 18 21 ...
 $ ClaimNb    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Exposure   : num  0.1 0.77 0.75 0.09 0.84 0.52 0.45 0.27 0.71 0.15 ...
 $ Area       : Factor w/ 6 levels "A","B","C","D",..: 4 4 2 2 2 5 5 3 3 2 ...
 $ VehPower   : int  5 5 6 7 7 6 6 7 7 7 ...
 $ VehAge     : int  0 0 2 0 0 2 2 0 0 0 ...
 $ DrivAge    : int  55 55 52 46 46 38 38 33 33 41 ...
 $ BonusMalus: int  50 50 50 50 50 50 50 68 68 50 ...
 $ VehBrand   : Factor w/ 11 levels "B1","B10","B11",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ VehGas     : Factor w/ 2 levels "Diesel","Regular": 2 2 1 1 1 2 2 1 1 1 ...
 $ Density    : int  1217 1217 54 76 76 3003 3003 137 137 60 ...
 $ Region     : Factor w/ 22 levels "R11","R21","R22",..: 18 18 3 15 15 8 8 20 20 12
```

- 3 categorical covariates, 1 binary covariate and 5 continuous covariates
- Goal: Find systematic effects to explain/predict claim counts `ClaimNb`.

Exposures and Claims

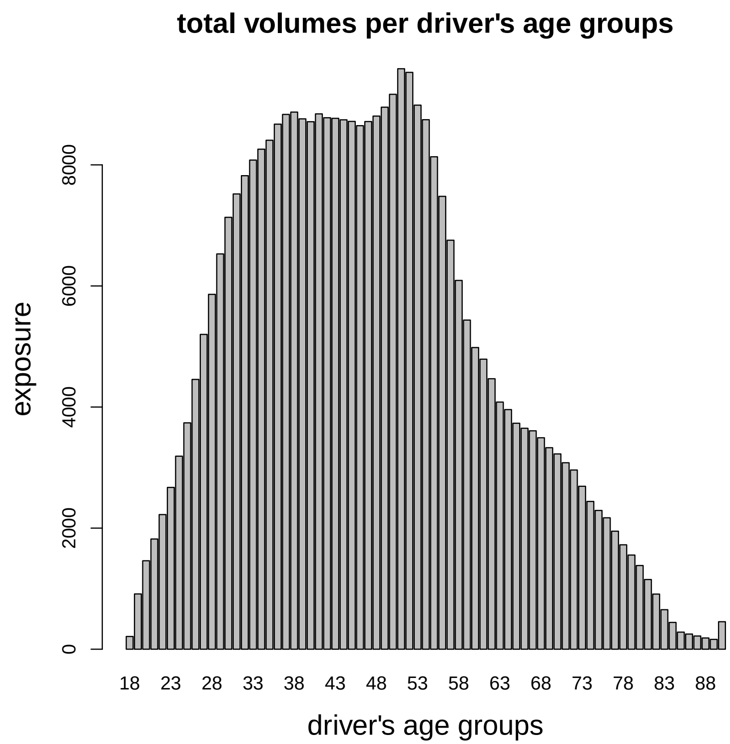


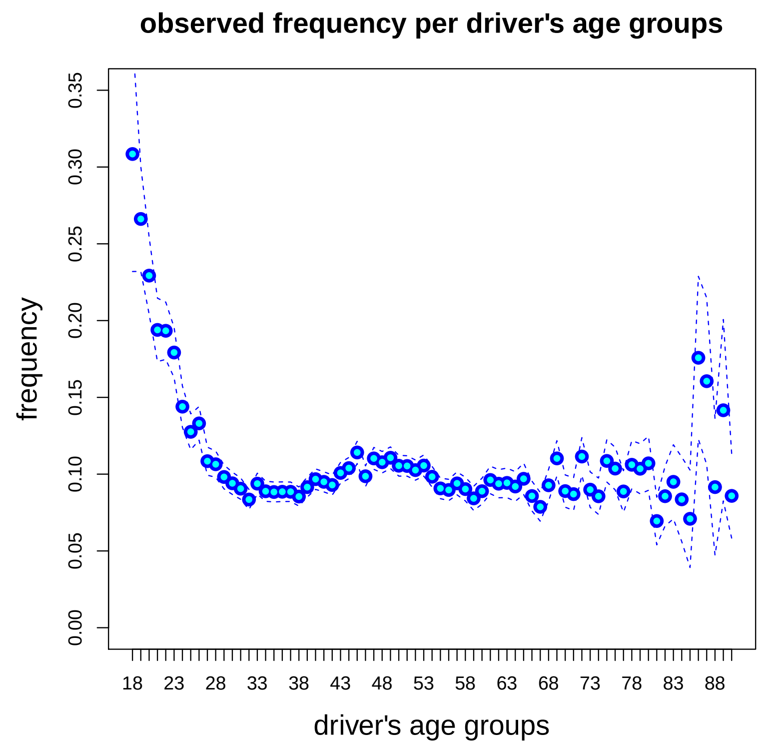




- Most exposures are between 0 and 1 year.
- Exposures bigger than 1 are considered to be data error and are capped at 1.
- Most insurance policies do not suffer any claim (class imbalance problem).

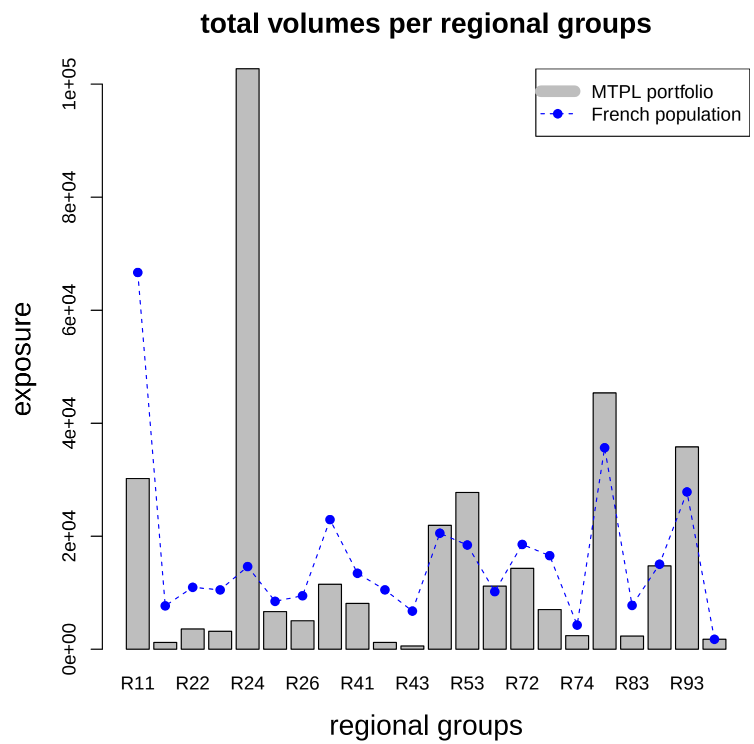
Continuous Covariates: Age of Driver

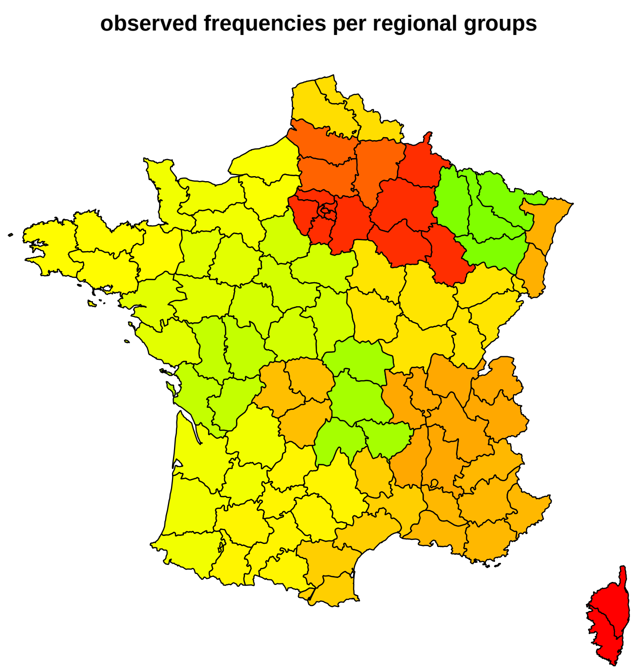




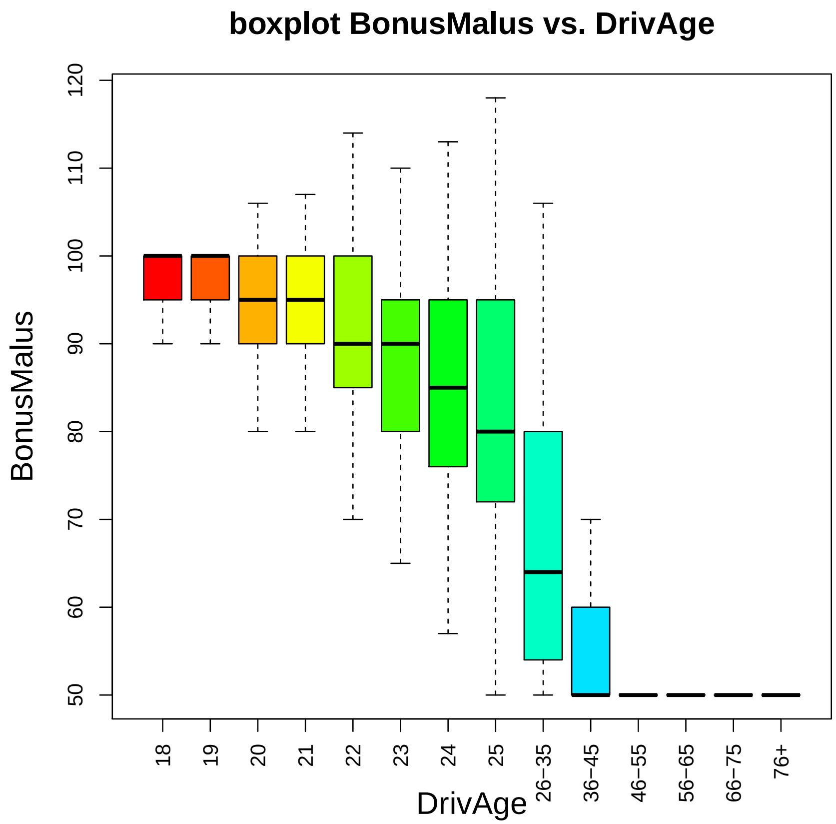
- Systematic effects of continuous covariates are not necessarily monotone.

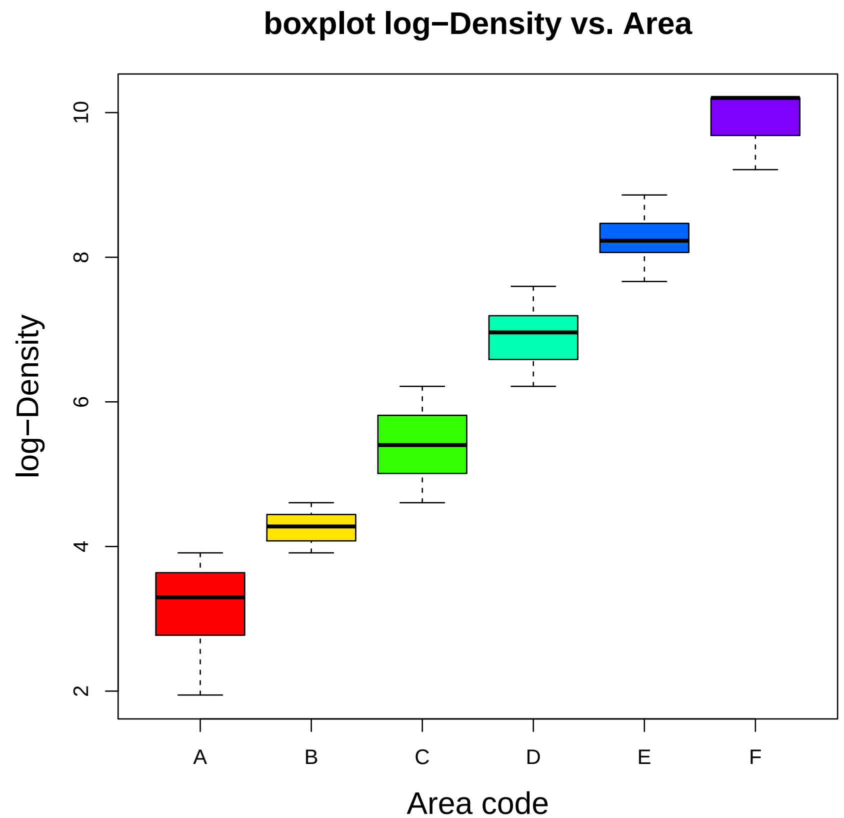
Categorical Covariates: French Region





Covariates: Dependence





- These covariates show strong dependence/collinearity.

Goal: Regression Modeling

- Denote by $\boldsymbol{x}_i$ the covariates of insurance policy $1 \le i \le n$.
- **Goal:** Find regression function μ:

$$\boldsymbol{x}_i \;\mapsto \mu(\boldsymbol{x}_i),$$

  such that for all insurance policies $1 \le i \le n$ we have

$$\mathbb{E}[N_i] = \mu(\boldsymbol{x}_i) v_i,$$

  where N_i denotes the number of claims and $v_i > 0$ is the time exposure of insurance policy $1 \le i \le n$ (pro-rata temporis).

- μ extracts the systematic effects from information $\boldsymbol{x}_i$ to explain N_i.

- **Exponential Dispersion Family (EDF)**

Exponential Dispersion Family (EDF)

- Sir Fisher (1934), Barndorff-Nielsen (2014), Jørgensen (1986, 1987).
- Exponential dispersion family (EDF) gives a unified notational framework of a large family of distribution functions.
- The parametrization of this family is chosen such that it is particularly suitable for maximum likelihood estimation (MLE).
- The EDF is the base statistical model for generalized linear modeling (GLM) and for neural network regressions.
- Examples: Gaussian, Poisson, gamma, binomial, categorical, Tweedie's, inverse Gaussian models.

- Remark: This first chapter on GLMs gives us the basic understanding and tools for neural network regression modeling.

Exponential Dispersion Family (EDF)

- Assume $(Y_i)_i$ are independent with density

$$Y_i \;\sim\; f(y;\theta_i, v_i/\varphi) = \exp\left\{\frac{y\theta_i - \kappa(\theta_i)}{\varphi/v_i} + a(y; v_i/\varphi)\right\},$$

with

$v_i > 0$	(known) exposure of risk i,
$\varphi > 0$	dispersion parameter,
$\theta_i \in \boldsymbol{\Theta}$	canonical parameter of risk i in the effective domain $\boldsymbol{\Theta}$,
$\kappa : \boldsymbol{\Theta} \to \mathbb{R}$	cumulant function (type of distribution),
$a(\cdot;\cdot)$	normalization, *not* depending on the canonical parameter θ_i.

Cumulant Function

- Assume $(Y_i)_i$ are independent with density

$$Y_i \;\sim\; f(y;\theta_i, v_i/\varphi) = \exp\left\{\frac{y\theta_i - \kappa(\theta_i)}{\varphi/v_i} + a(y; v_i/\varphi)\right\}.$$

- Cumulant function $\kappa : \boldsymbol{\Theta} \to \mathbb{R}$ is convex and smooth in the interior of $\boldsymbol{\Theta}$.

- Examples:

$$\kappa(\theta) = \begin{cases} \theta^2/2 & \text{Gauss,} \\ \exp(\theta) & \text{Poisson,} \\ -\log(-\theta) & \text{gamma,} \\ \log(1+e^\theta) & \text{Bernoulli/binomial,} \\ -(-2\theta)^{1/2} & \text{inverse Gaussian,} \\ ((1-p)\theta)^{\frac{2-p}{1-p}}/(2-p) & \text{Tweedie with } p>1,\ p\neq 2. \end{cases}$$

Mean and Variance Function

- The mean is given by

$$\mu_i = \mathbb{E}[Y_i] = \kappa'(\theta_i).$$

- The variance is given by

$$\mathrm{Var}(Y_i) = \frac{\varphi}{v_i}\kappa''(\theta_i) = \frac{\varphi}{v_i}V(\mu_i) > 0,$$

where $\mu \mapsto V(\mu) = \kappa''((\kappa')^{-1}(\mu))$ is the so-called variance function.

- Examples:

$$V(\mu) = \begin{cases} 1 & \text{Gauss,} \\ \mu & \text{Poisson,} \\ \mu^2 & \text{gamma,} \\ \mu^3 & \text{inverse Gaussian,} \\ \mu^p & \text{Tweedie with } p \geq 1. \end{cases}$$

Maximum Likelihood Estimation (MLE)

- MLE homogeneous θ case: log-likelihood of independent observations $(Y_i)_{i=1}^n$ is

$$\ell_{\mathbf{Y}}(\theta) \;=\; \log\left(\prod_{i=1}^n f(Y_i;\theta, v_i/\varphi)\right) \;=\; \sum_{i=1}^n \frac{Y_i\theta - \kappa(\theta)}{\varphi/v_i} + a(Y_i; v_i/\varphi).$$

- This provides score equations

$$\frac{\partial}{\partial\theta}\ell_{\mathbf{Y}}(\theta) \;=\; \sum_{i=1}^n \frac{v_i}{\varphi}\left[Y_i - \kappa'(\theta)\right] = 0,$$

and MLE $\widehat{\theta}$

$$\widehat{\theta} = (\kappa')^{-1}\left(\frac{\sum_{i=1}^n v_i Y_i}{\sum_{i=1}^n v_i}\right).$$

- MLE is straightforward within the EDF!

Canonical Link and Unbiasedness

- Canonical link $h(\cdot) = (\kappa')^{-1}(\cdot)$

$$\mu = \mathbb{E}[Y] = \kappa'(\theta) \qquad \text{or} \qquad h(\mu) = h(\mathbb{E}[Y]) = \theta.$$

- This provides for the MLE

$$\widehat{\theta} \ = \ (\kappa')^{-1}\left(\frac{\sum_{i=1}^{n} v_i Y_i}{\sum_{i=1}^{n} v_i}\right) \ = \ h\left(\frac{\sum_{i=1}^{n} v_i Y_i}{\sum_{i=1}^{n} v_i}\right).$$

The latter gives a sufficient statistics.

- Unbiasedness of estimated means in the homogeneous case

$$\mathbb{E}\left[\widehat{\mathbb{E}}[Y]\right] = \mathbb{E}\left[\kappa'(\widehat{\theta})\right] = \kappa'(\theta).$$

▷ Unbiasedness emphasizes that we receive the right price level in pricing.

- **Generalized Linear Models (GLMs)**

Generalized Linear Models (GLMs)

- Nelder–Wedderburn (1972) and McCullagh–Nelder (1983).

- Assume we have heterogeneity between $(Y_i)_{i=1}^n$ which manifests in systematic effects modeled through covariates/features $\boldsymbol{x}_i \in \mathbb{R}^q$.

- Assume for link function choice g and regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$

$$\boldsymbol{x}_i \ \mapsto \ g(\mu_i) = g(\mathbb{E}[Y_i]) = g(\kappa'(\theta_i)) = \beta_0 + \sum_{j=1}^{q} \beta_j x_{i,j}.$$

  This gives a GLM with link function g. Parameter β_0 is called intercept/bias.

- Link g should be monotone and smooth.

- The choice $g = h = (\kappa')^{-1}$ is called canonical link.

Design Matrix

- Assume for link function choice g and regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$

$$\boldsymbol{x}_i \;\mapsto\; g(\mu_i) = g(\mathbb{E}[Y_i]) = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle = \beta_0 + \sum_{j=1}^{q} \beta_j x_{i,j}.$$

- The design matrix is

$$\mathfrak{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,q} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,q} \end{pmatrix} \in \mathbb{R}^{n \times (q+1)}.$$

- The design matrix $\mathfrak{X}$ is assumed to have full rank $q + 1 \leq n$.

- Full rank property is important for uniqueness of MLE of $\boldsymbol{\beta}$.

Maximum Likelihood Estimation of GLMs

- The log-likelihood of independent observations $(Y_i)_{i=1}^n$ is given by

$$\boldsymbol{\beta} \;\mapsto\; \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} \frac{Y_i h(\mu_i) - \kappa(h(\mu_i))}{\varphi/v_i} + a(Y_i; v_i/\varphi),$$

  with mean $\mu_i = \mu_i(\boldsymbol{\beta}) = g^{-1}\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle$ and canonical parameter $\theta_i = h(\mu_i)$.

- This provides score equations for MLE

$$\nabla_{\boldsymbol{\beta}}\, \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) \;=\; 0.$$

- Score equations are solved numerically with Fisher's scoring method or the iterated re-weighted least squares (IRLS) algorithm.

MLE and Deviance Loss Functions

- The log-likelihood of independent observations $(Y_i)_{i=1}^n$ is given by

$$\ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) \;=\; \sum_{i=1}^n \frac{Y_i h(\mu_i) - \kappa(h(\mu_i))}{\varphi/v_i} + a(Y_i; v_i/\varphi),$$

  with mean $\mu_i = \mu_i(\boldsymbol{\beta}) = g^{-1}\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle$.

- Maximizing log-likelihoods is equivalent to minimizing deviance losses

$$\begin{aligned} D^*(\boldsymbol{Y}, \boldsymbol{\beta}) &= 2\left[\ell_{\boldsymbol{Y}}(\boldsymbol{Y}) - \ell_{\boldsymbol{Y}}(\boldsymbol{\beta})\right] \\ &= 2\sum_{i=1}^n \frac{v_i}{\varphi}\Big[Y_i h(Y_i) - \kappa(h(Y_i)) - Y_i h(\mu_i) + \kappa(h(\mu_i))\Big] \;\geq\; 0. \end{aligned}$$

- The deviance loss of the Gaussian model is the square loss function, other examples of the EDF have deviance losses *different* from square losses.

Examples of Deviance Loss Functions

- Gaussian case:

$$D^*(\boldsymbol{Y},\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{v_i}{\varphi} \left(Y_i - \mu_i\right)^2 \;\; \geq \;\; 0.$$

- Gamma case:

$$D^*(\boldsymbol{Y},\boldsymbol{\beta}) = 2\sum_{i=1}^{n} \frac{v_i}{\varphi} \left(\frac{Y_i}{\mu_i} - 1 + \log\left(\frac{\mu_i}{Y_i}\right)\right) \;\; \geq \;\; 0.$$

- Inverse Gaussian case:

$$D^*(\boldsymbol{Y},\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{v_i}{\varphi} \frac{(Y_i - \mu_i)^2}{\mu_i^2 Y_i} \;\; \geq \;\; 0.$$

- Poisson case:

$$D^*(\boldsymbol{Y},\boldsymbol{\beta}) = 2\sum_{i=1}^{n} \frac{v_i}{\varphi} \left(\mu_i - Y_i - Y_i \log\left(\frac{\mu_i}{Y_i}\right)\right) \;\; \geq \;\; 0.$$

Balance Property under Canonical Link

- Under the canonical link $g = h = (\kappa')^{-1}$ we have balance property for the MLE

$$\sum_{i=1}^{n} v_i \widehat{\mathbb{E}}[Y_i] \;=\; \sum_{i=1}^{n} v_i \kappa' \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{x}_i \rangle = \sum_{i=1}^{n} v_i Y_i.$$

  ▷ The estimated model mean over the entire portfolio is unbiased.

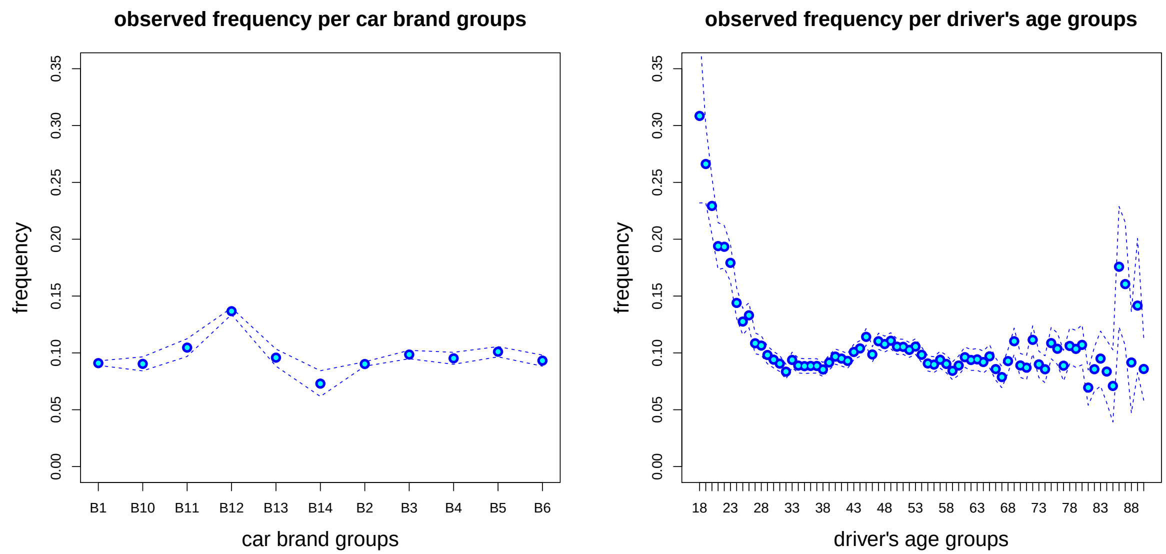
- If one does not work with the canonical link, one should correct in $\widehat{\beta}_0$ for the bias.

- **Feature Engineering / Covariate Pre-Processing**

Feature Engineering

- Assume monotone link function choice g

$$\boldsymbol{x}_i \;\mapsto\; \mu_i = \mathbb{E}[Y_i] \;=\; g^{-1}\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle = g^{-1}\left(\beta_0 + \sum_{j=1}^{q} \beta_j x_{i,j}\right).$$



- What about categorical covariates and non-monotone covariates?

- What about different interactions?

One-Hot Encoding of Categorical Covariates

B1 $\mapsto \boldsymbol{e}_1 =$	1	0	0	0	0	0	0	0	0	0	0
B10 $\mapsto \boldsymbol{e}_2 =$	0	1	0	0	0	0	0	0	0	0	0
B11 $\mapsto \boldsymbol{e}_3 =$	0	0	1	0	0	0	0	0	0	0	0
B12 $\mapsto \boldsymbol{e}_4 =$	0	0	0	1	0	0	0	0	0	0	0
B13 $\mapsto \boldsymbol{e}_5 =$	0	0	0	0	1	0	0	0	0	0	0
B14 $\mapsto \boldsymbol{e}_6 =$	0	0	0	0	0	1	0	0	0	0	0
B2 $\mapsto \boldsymbol{e}_7 =$	0	0	0	0	0	0	1	0	0	0	0
B3 $\mapsto \boldsymbol{e}_8 =$	0	0	0	0	0	0	0	1	0	0	0
B4 $\mapsto \boldsymbol{e}_9 =$	0	0	0	0	0	0	0	0	1	0	0
B5 $\mapsto \boldsymbol{e}_{10} =$	0	0	0	0	0	0	0	0	0	1	0
B6 $\mapsto \boldsymbol{e}_{11} =$	0	0	0	0	0	0	0	0	0	0	1

- One-hot encoding for the 11 car brands: $\texttt{brand} \mapsto \boldsymbol{e}_{\mathrm{j}} \in \mathbb{R}^{11}$.

- One-hot encoding does not lead to full rank design matrices $\mathfrak{X}$, because we have a redundancy.

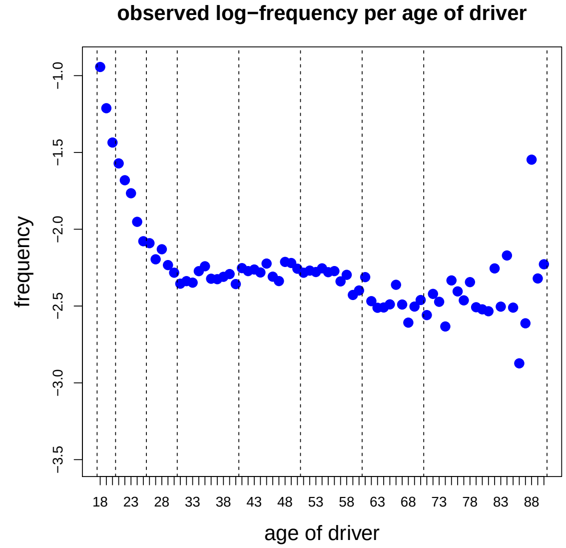
Dummy Coding of Categorical Covariates

B1	0	0	0	0	0	0	0	0	0	0
B10	1	0	0	0	0	0	0	0	0	0
B11	0	1	0	0	0	0	0	0	0	0
B12	0	0	1	0	0	0	0	0	0	0
B13	0	0	0	1	0	0	0	0	0	0
B14	0	0	0	0	1	0	0	0	0	0
B2	0	0	0	0	0	1	0	0	0	0
B3	0	0	0	0	0	0	1	0	0	0
B4	0	0	0	0	0	0	0	1	0	0
B5	0	0	0	0	0	0	0	0	1	0
B6	0	0	0	0	0	0	0	0	0	1

- Declare one label as reference level and drop the corresponding column.
- Dummy coding for the 11 car brands: $\texttt{brand} \mapsto \boldsymbol{x}_{\mathtt{j}} \in \mathbb{R}^{10}$.
- Dummy coding leads to full rank design matrices $\mathfrak{X}$.
- There are other full rank codings like Helmert's contrast coding.

Pre-Processing of Continuous Covariates (1/2)

age class 1:	18-20
age class 2:	21-25
age class 3:	26-30
age class 4:	31-40
age class 5:	41-50
age class 6:	51-60
age class 7:	61-70
age class 8:	71-90



- Continuous features need feature engineering, too, to bring them into the right functional form for GLM. Assume we have log-link for g

$$\boldsymbol{x} \;\mapsto\; \log\left(\mathbb{E}\left[Y\right]\right) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle = \beta_0 + \sum_{j=1}^{q} \beta_j x_j.$$

- We build homogeneous categorical classes, and then apply dummy coding.

Pre-Processing of Continuous Covariates (2/2)

- Categorical coding of continuous covariates has some disadvantages.
- By changing continuous features to categorical dummies we lose adjacency relationships between neighboring classes.
- The number of parameters can grow very large if we have many classes.
- Balance property holds true on every categorical level.
  Caution: if we have very rare categorical levels this will lead to over-fitting; and it will also lead to high correlations with the intercept β_0.
- One may also consider other functional forms for continuous covariates, e.g.,

$$\texttt{age} \;\mapsto\; \beta_1\texttt{age} + \beta_2\texttt{age}^2 + \beta_3 \log(\texttt{age}).$$

- Similarly, we can model interactions between covariate components

$$(\texttt{age}, \texttt{weight}) \;\mapsto\; \beta_1\texttt{age} + \beta_2\texttt{weight} + \beta_3\texttt{age}/\texttt{weight}.$$

- **Variable Selection**

Variable Selection: Likelihood Ratio Test (LRT)

- **Null hypothesis** H_0: $\beta_1 = \ldots = \beta_p = 0$ for given $1 \leq p \leq q$.

- **Likelihood ratio test (LRT).** Calculate test statistics (nested models)

$$\chi^2_{\boldsymbol{Y}} = D^*(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}_{H_0}) - D^*(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}_{\text{full}}) \geq 0.$$

Under H_0, test statistics $\chi^2_{\boldsymbol{Y}}$ is approximately χ^2-distributed with p df.

Variable Selection: Wald Test

- **Null hypothesis** H_0: $\boldsymbol{\beta}_p = (\beta_1, \ldots, \beta_p)^\top = 0$ for given $1 \leq p \leq q$.

- **Wald test.** Choose matrix I_p such that $I_p \boldsymbol{\beta}_{\text{full}} = \boldsymbol{\beta}_p$. Consider Wald statistics

$$W = (I_p \widehat{\boldsymbol{\beta}}_{\text{full}} - 0)^\top \left(I_p \, \mathcal{I}(\widehat{\boldsymbol{\beta}}_{\text{full}})^{-1} \, I_p^\top \right)^{-1} (I_p \widehat{\boldsymbol{\beta}}_{\text{full}} - 0).$$

  Under H_0, test statistics W is approximately χ^2-distributed with p df.

- $\mathcal{I}(\widehat{\boldsymbol{\beta}}_{\text{full}})$ is Fisher's information matrix; the above test is based on asymptotic normality of the MLE $\widehat{\boldsymbol{\beta}}_{\text{full}}$.

- Model only needs to be fitted once.

Model Selection: AIC

- **Akaike's information criterion (AIC)** is useful for non-nested models

$$\mathrm{AIC} = -2\ell_{\boldsymbol{Y}}(\widehat{\boldsymbol{\beta}}) + 2\mathrm{dim}(\boldsymbol{\beta}).$$

- Models do not need to be nested.
- Models can have different distributions.
- AIC considers all terms of the log-likelihood (also normalizing constants).
- Models need to be estimated with MLE.
- Different models need to consider the same data on the same scale (log-normal vs. gamma).

Example: Poisson Frequency GLM

```
Call:
glm(formula = claims ~ powerCAT + area + log(dens) + gas + ageCAT +
                       acCAT + brand + ct, family = poisson(), data = dat, offset =

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1373  -0.3820  -0.2838  -0.1624   4.3856

Coefficients:
               Estimate Std. Error z value Pr(>!z!)
(Intercept)  -1.903e+00  4.699e-02 -40.509  < 2e-16 ***
powerCAT2     2.681e-01  2.121e-02  12.637  < 2e-16 ***
.
.
powerCAT9    -1.044e-01  4.708e-02  -2.218 0.026564 *
area          4.333e-02  1.927e-02   2.248 0.024561 *
log(dens)     3.224e-02  1.432e-02   2.251 0.024385 *
gasRegular    6.868e-02  1.339e-02   5.129 2.92e-07 ***
.
.
ctZG         -8.123e-02  4.638e-02  -1.751 0.079900 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

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

```

    Null deviance: 145532  on 499999  degrees of freedom
Residual deviance: 140641  on 499943  degrees of freedom
AIC: 191132
```

Forward Parameter Selection: ANOVA

```
Analysis of Deviance Table

Model: poisson, link: log

Response: claims

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev
NULL                      499999     145532
acCAT       3  2927.32    499996     142605
ageCAT      7   850.00    499989     141755
ct         25   363.29    499964     141392
brand      10   124.37    499954     141267
powerCAT    8   315.48    499946     140952
gas         1    50.53    499945     140901
area        1   255.20    499944     140646
log(dens)   1     5.07    499943     140641
```

Pay attention: order of covariates inclusion is important.

Backward Parameter Reduction: Drop1

```
Single term deletions

Model:
claims ~ acCAT + ageCAT + ct + brand + powerCAT + gas + area + log(dens)

          Df Deviance    AIC     LRT  Pr(>Chi)
<none>         140641 191132
acCAT      3   142942 193426 2300.61 < 2.2e-16 ***
ageCAT     7   141485 191962  843.91 < 2.2e-16 ***
ct        25   140966 191406  324.86 < 2.2e-16 ***
brand     10   140791 191261  149.70 < 2.2e-16 ***
powerCAT   8   140969 191443  327.68 < 2.2e-16 ***
gas        1   140667 191156   26.32 2.891e-07 ***
area       1   140646 191135    5.06   0.02453 *
log(dens)  1   140646 191135    5.07   0.02434 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We should keep the full model according to AIC and according to the LRT on a 5% significance level.

- **Car Insurance Frequency Example**

Example: Poisson Frequency Model (1/2)

- The Poisson model has dispersion $\varphi = 1$.

- The Poisson model has cumulant function

$$\theta \;\mapsto\; \kappa(\theta) = \exp(\theta).$$

- Mean and variance of EDFs are given by

$$\mu_i \;=\; \mathbb{E}\left[Y_i\right] \;=\; \kappa'(\theta_i) \;=\; \exp(\theta_i),$$

$$\mathrm{Var}\left(Y_i\right) \;=\; \frac{\varphi}{v_i}\kappa''(\theta_i) \;=\; \frac{1}{v_i}\exp(\theta_i) \;=\; \frac{1}{v_i}\mu_i.$$

$\triangleright$ $N_i = v_i Y_i$ has a Poisson distribution with mean $v_i \mu_i$.

Example: Poisson Frequency Model (2/2)

- Mean of the Poisson model for $N_i = v_i Y_i$

$$v_i \mu_i \;=\; \mathbb{E}\left[N_i\right] \;=\; v_i \kappa'(\theta_i) \;=\; v_i \exp(\theta_i) \;=\; \exp(\log v_i + \theta_i).$$

The term $\log v_i$ is called offset.

- The Poisson GLM with canonical link $g = h = \log$ is given by

$$\boldsymbol{x}_i \;\mapsto\; \log\left(\mathbb{E}\left[N_i\right]\right) = \log v_i + \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle = \log v_i + \beta_0 + \sum_{j=1}^{q} \beta_j x_{i,j}.$$

	run time	# param. $q+1$	AIC	in-sample loss	out-of-sample loss
homogeneous model	–	1	263’143	32.935	33.861
Model GLM1	20s	49	253’062	31.267	32.171

Losses are in 10^{-2}.

Further Points

- To prevent from over-fitting: regularization can be used.
- Ridge regression is based on an L^2-penalization and generally reduces regression parameter components in $\boldsymbol{\beta}$ (exclude the intercept β_0).
- LASSO (least absolute shrinkage and selection operator) regression is based on an L^1-penalization and can set regression parameter components exactly to zero.
- LASSO has difficulties with collinearity in covariate components, therefore, sometimes an elastic net regularization is used which combines ridge and LASSO.
- Regularization has a Bayesian interpretation.
- Generalized additive models (GAMs) allow for more flexibility than GLMs in marginal covariate component modeling. But they often suffer from computational complexity.

References

- Barndorff-Nielsen (2014). Information and Exponential Families: In Statistical Theory. Wiley
- Charpentier (2015). Computational Actuarial Science with R. CRC Press.
- Efron, Hastie (2016). Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge UP.
- Fahrmeir, Tutz (1994). Multivariate Statistical Modelling Based on Generalized Linear Models. Springer.
- Fisher (1934). Two new properties of mathematical likelihood. Proceeding of the Royal Society A 144, 285-307.
- Hastie, Tibshirani, Friedman (2009). The Elements of Statistical Learning. Springer.
- Jørgensen (1986). Some properties of exponential dispersion models. Scandinavian Journal of Statistics 13/3, 187-197.
- Jørgensen (1987). Exponential dispersion models. Journal of the Royal Statistical Society. Series B (Methodological) 49/2, 127-145.
- Jørgensen (1997). The Theory of Dispersion Models. Chapman & Hall.
- Lehmann (1983). Theory of Point Estimation. Wiley.
- Lorentzen, Mayer (2020). Peeking into the black box: an actuarial case study for interpretable machine learning. SSRN 3595944.
- McCullagh, Nelder (1983). Generalized Linear Models. Chapman & Hall.
- Nelder, Wedderburn (1972). Generalized linear models. Journal of the Royal Statistical Society. Series A (General) 135/3, 370-384.
- Noll, Salzmann, Wüthrich (2018). Case study: French motor third-party liability claims. SSRN 3164764.
- Ohlsson, Johansson (2010). Non-Life Insurance Pricing with Generalized Linear Models. Springer.
- Wüthrich, Buser (2016). Data Analytics for Non-Life Insurance Pricing. SSRN 2870308, Version September 10, 2020.
- Wüthrich, Merz (2021). Statistical Foundations of Actuarial Learning and its Applications. SSRN 3822407.