IA|BE Data Science Certificate

Module 1 on Foundations of machine learning in actuarial sciences
Linear and Generalized Linear Models

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

October 19, 2021

P&C insurance pricing models

- In a pricing model, identify for each insured i:

  N_i: number of claims during (period of) exposure d_i

  L_i: aggregate loss over N_i claims.

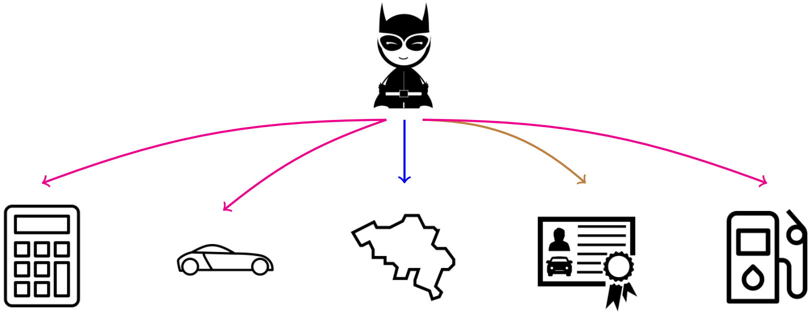
- The pure premium or risk premium π_i:

$$\pi_i \quad = \quad E\left(\frac{N_i}{d_i}\right) \cdot E\left(\frac{L_i}{N_i} | N_i > 0\right) = \underbrace{E(F_i)}_{\text{frequency}} \cdot \underbrace{E(\text{Sev}_i)}_{\text{severity}}.$$

- Classify risks using a priori measurable characteristics:

  risk classification or segmentation.

Motivation



Claim frequency and claim severity

as function of

nominal / numeric ~ ordinal / spatial

features


EXAMPLE

Risk classification: an example with claim counts

- This example is from the book by Denuit et al., Actuarial Modelling of Claim Counts, starting on p.52.
- Data are from a Belgian motor third party liability insurance portfolio, observed during the year 1997.
- 14,505 policies with an observed mean claim frequency of 14.6%.
- The observed claim number distribution is given below:

Number of claims	Number of policies	Total exposure (in years)
0	12,962	10,545.94
1	1,369	1,187.13
2	157	134.66
3	14	11.08
4	3	2.52
Total	14,505	11,881.35

Risk classification: an example with claim counts

- Available risk characteristics:
  - Age: policyholder's age with 4 categories (1='between 18 and 24'; 2='between 25 and 30'; 3='between 31 and 60'; 4='more than 60')
  - Gender: policyholder's gender
  - District: kind of district where the policyholder lives (1='urban'; 2='rural')
  - Use: use of the car (1='private use'; 2='professional use')
  - Split: premium split (1='premium paid once a year'; 2= 'premium split up').
- All the explanatory variables listed above are categorical (or factor variables).

Risk classification: an example with claim counts

What about exposure–to–risk?

- Majority of policies are in force during the whole year.
- Some policies do not have an observation period of a full year:
  - new policyholders entering the portfolio during the observation period
  - when changes occur in the observable characteristics of the policies (e.g. a move).

Risk classification: an example with claim counts

- Preliminary: actuary considers the marginal impact of each rating factor (disregarding the possible effect of other explanatory variables).

- Assume for instance $N_i \sim \text{Poi}(d_i \lambda_{\text{Age}(i)})$ (for insured i)

  [Poisson regression $\in$ Generalized Linear Models (GLMs)]

  - N_i the number of claims
  - d_i the exposure to risk
  - $\text{Age}(i)$ the age category to which insured i belongs.

- λ_js the annual expected claim frequencies for the 4 age classes.

Risk classification: an example with claim counts

- Recall: $N \sim \text{POI}(\lambda)$ implies $P(N=k) = \exp(-\lambda)\frac{\lambda^k}{k!}$.
- Therefore:

  for $N_i \sim \text{Poi}(d_i \lambda_{\text{Age}(i)})$

$$P(N_i = k) \quad = \quad \exp\left(-d_i \lambda_{\text{Age}(i)}\right) \frac{\left(d_i \lambda_{\text{Age}(i)}\right)^k}{k!},$$

  with $k = 0, 1, 2, \ldots$.

Risk classification: an example with claim counts

Let's start with a one–way analysis using the Age covariate:

- assuming independence between policy holders: construct Poisson likelihood

$$
\begin{aligned}
\mathcal{L}(\lambda_1,\lambda_2,\lambda_3,\lambda_4) &= \prod_{i=1}^{n} P(N_i = k_i) = \prod_{i=1}^{n} \exp\left(-d_i\lambda_{\text{Age(i)}}\right)\frac{\left(d_i\lambda_{\text{Age}(i)}\right)^{k_i}}{k_i!} \\
&\propto \prod_{j=1}^{4} \exp\left(-\lambda_j \sum_{i|\text{Age}(i)=j} d_i\right)\lambda_j^{\sum_{i|\text{Age}(i)=j} k_i},
\end{aligned}
$$

where k_i denotes the observed number of claims for policyholder i.

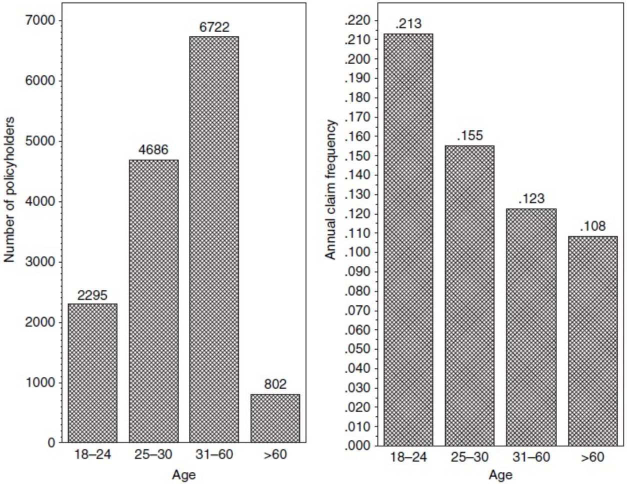
Risk classification: an example with claim counts

Let's start with a one–way analysis using the Age covariate:

- differentiate $L(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \ln \mathcal{L}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ with respect to λ_j;
- set the derivative to 0

$$
\begin{aligned}
-\sum_{i|\text{Age}(i)=j} d_i + \frac{1}{\lambda_j} \sum_{i|\text{Age}(i)=j} k_i &= 0 \\
&\Updownarrow \\
\hat{\lambda}_j &= \frac{\sum_{i|\text{Age}(i)=j} k_i}{\sum_{i|\text{Age}(i)=j} d_i}.
\end{aligned}
$$

Risk classification: an example with claim counts



Risk classification: an example with claim counts

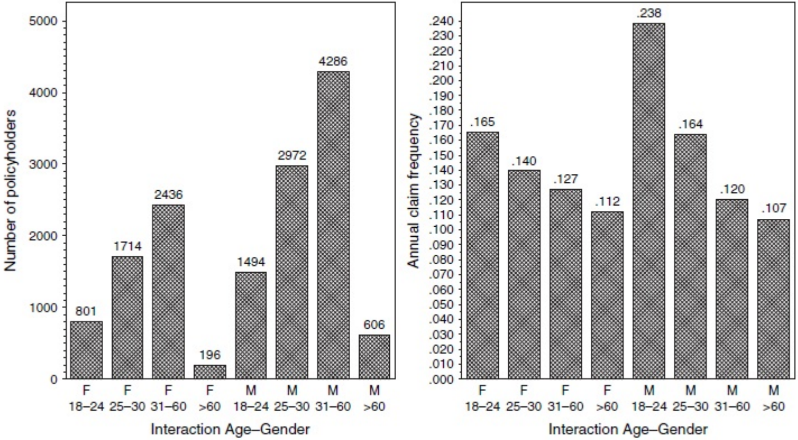
- When two explanatory variables interact:

  the effect of *one* factor varies depending on the levels of the *other* factor.

- Example: gender–age interaction

  the effect of age on the average claim frequency is different for males than for females.

Risk classification: an example with claim counts



Risk classification: an example with claim counts

- All explanatory variables presented above are categorical (or factor).
- Defining binary (or: dummy) variables:
  - a categoric variable with k levels partitions into k classes
  - coded with $k-1$ binary variables
  - all zero for the reference level.

Risk classification: an example with claim counts

- Example: our portfolio has reference levels '31-60' for Age, 'Male' for Gender, 'Urban' for District, 'Premium paid once a year' for Split and 'Private' for Use.
- Policyholder i is then represented by a vector of dummies with values of:

$$
\begin{aligned}
x_{i1} &= \begin{cases} 1 \text{ if policyholder } i \text{ less than } 24 \\ 0 \text{ otherwise,} \end{cases} \\
x_{i2} &= \begin{cases} 1 \text{ if policyholder } i \text{ is } 25\text{--}30 \\ 0 \text{ otherwise,} \end{cases} \\
x_{i3} &= \begin{cases} 1 \text{ if policyholder } i \text{ over } 60 \\ 0 \text{ otherwise,} \end{cases}
\end{aligned}
$$

Risk classification: an example with claim counts

- Example: our portfolio has reference levels '31-60' for Age, 'Male' for Gender, 'Urban' for District, 'Premium paid once a year' for Split and 'Private' for Use.

- ... and:

$$
\begin{aligned}
x_{i4} &= \begin{cases} 1 \text{ if policyholder } i \text{ female} \\ 0 \text{ otherwise,} \end{cases} \\
x_{i5} &= \begin{cases} 1 \text{ if policyholder } i \text{ lives in a rural area} \\ 0 \text{ otherwise,} \end{cases} \\
x_{i6} &= \begin{cases} 1 \text{ if policyholder } i \text{ splits his premium payment} \\ 0 \text{ otherwise,} \end{cases} \\
x_{i7} &= \begin{cases} 1 \text{ if policyholder } i \text{ uses his car for professional reasons} \\ 0 \text{ otherwise.} \end{cases}
\end{aligned}
$$

Risk classification: an example with claim counts

- Reference class:

  all x_{ij}s are equal to 0

  $=$ a man aged between 31 and 60, living in an urban area, paying premium once a year and using the car for private purposes only.

- Example:

  sequence (1,0,0,0,0,0,1)

  $=$ a man aged less than 24, living in an urban area, paying premium once a year and using the car for professional reasons.

Risk classification: an example with claim counts

- Consider the interaction between Age and Gender in our portfolio.
- To reflect the situation accurately:

  create an interaction variable 'Gender · Age' $\Rightarrow$ 8 levels, coded by 7 dummies.
- Replace $x_{i1}, \ldots, x_{i4}$ with $x'_{i1}, \ldots, x'_{i7}$.

Risk classification: an example with claim counts

- Policyholder i is then represented by a vector of dummies with values of:

$$
\begin{aligned}
x'_{i1} &= \begin{cases} 1 \text{ if policyholder } i \text{ female } \leq 24 \\ 0 \text{ otherwise,} \end{cases} \\
x'_{i2} &= \begin{cases} 1 \text{ if policyholder } i \text{ female aged } 25\text{–}30 \\ 0 \text{ otherwise,} \end{cases} \\
x'_{i3} &= \begin{cases} 1 \text{ if policyholder } i \text{ female aged } 31\text{–}60 \\ 0 \text{ otherwise,} \end{cases} \\
x'_{i4} &= \begin{cases} 1 \text{ if policyholder } i \text{ female over } 60 \\ 0 \text{ otherwise.} \end{cases}
\end{aligned}
$$

Risk classification: an example with claim counts

- Policyholder i is then represented by a vector of dummies with values of:

$$
\begin{aligned}
x'_{i5} &= \begin{cases} 1 \text{ if policyholder } i \text{ male } \leq 24 \\ 0 \text{ otherwise,} \end{cases} \\
x'_{i6} &= \begin{cases} 1 \text{ if policyholder } i \text{ male aged } 25\text{–}30 \\ 0 \text{ otherwise,} \end{cases} \\
x'_{i7} &= \begin{cases} 1 \text{ if policyholder } i \text{ male over } 60 \\ 0 \text{ otherwise.} \end{cases}
\end{aligned}
$$

- Reference class: 'Male 31-60'.

Poisson regression model

- Let N_i $(i = 1, 2, \ldots, n)$ be the number of claims reported by policyholder i with corresponding risk exposure d_i.
- Assume the N_is are independent.
- $\boldsymbol{x}_i^{'} = (x_{i1}, \ldots, x_{ip})$: observable characteristics of this policyholder.
- Assumptions in the Poisson regression model:

$$\begin{aligned} E[N_i|\boldsymbol{x}_i] &= d_i \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right) \; i = 1, 2, \ldots, n, \\ N_i &\sim \text{Poi}\left(d_i \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)\right) \; i = 1, 2, \ldots, n. \end{aligned}$$

Poisson regression model

- The score or linear predictor:

$$\text{score}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

- The expected claim frequency for policyholder i is: $d_i \exp(\text{score}_i)$.
- With $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_p$ the estimates of the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$:

$$\begin{aligned} \hat{\lambda}_i &= d_i \exp\left(\widehat{\text{score}}_i\right) \\ &= d_i \exp\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}\right), \end{aligned}$$

the predicted expected number of claims for policyholder i.

Poisson regression model

- With explanatory variables x_{ij}s coded by means of binary variables:

  β_0 represents the risk associated to the reference class.

- Annual claim frequency λ_i associated with $\boldsymbol{x}_i$ is of multiplicative form:

$$\lambda_i = \exp(\beta_0) \prod_{j|x_{ij}=1} \exp(\beta_j),$$

  - $\exp(\beta_0)$ is the annual claim frequency corresponding with the reference class.
  - $\exp(\beta_j)$ models the impact of the jth ratemaking variable.

Poisson regression model

- Example: say in our portfolio the score for policyholder i is

$$\text{score}_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_7 x_{i7}.$$

  Here:

  - $\exp(\beta_0)$ = annual claim frequency for men aged 31-60, living in an urban area, paying the premium once a year, using the car for private purposes;
  - $\exp(\beta_0 + \beta_1)$ = annual claim frequency for men less than 24, living in an urban area, paying the premium once a year, using the care for private purposes;
  - and so on.

Poisson regression model: Maximum Likelihood

- Let k_i be the number of claims filed by policyholder i during the observation period, with exposure d_i.

- The associated likelihood:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}) &= \prod_{i=1}^{n} P[N_i = k_i | \boldsymbol{x}_i] \\ &= \prod_{i=1}^{n} \exp(-\lambda_i) \cdot \frac{\lambda_i^{k_i}}{k_i!}, \end{aligned}$$

  where $\lambda_i = d_i \exp(\text{score}_i) = \exp(\ln d_i + \text{score}_i)$.

- ML estimator $\hat{\boldsymbol{\beta}}$ maximizes $\mathcal{L}(\boldsymbol{\beta})$.

Poisson regression model: Wald CIs

- Asymptotic variance–covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ of MLE $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \left(\sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i' \hat{\lambda}_i\right)^{-1},$$

  where $\hat{\lambda}_i = d_i \exp\left(\widehat{\text{score}}_i\right)$.

- Provided the sample size is large enough $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is $\approx N(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$.

- A confidence interval at level $1-\alpha$ for each β_j:

$$[\hat{\beta}_j - z_{\alpha/2}\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + z_{\alpha/2}\hat{\sigma}_{\hat{\beta}_j}],$$

  where $\hat{\sigma}^2_{\hat{\beta}_j}$ is the element (j,j) of $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$, and $\Phi(-z_{\alpha/2}) = \alpha/2$.

Poisson regression model: hypothesis test on single parm.

- Say we want to test $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$.
- Failing to reject H_0 suggests that this variable is not relevant to explaining the expected number of claims.
- Test statistic for H_0 against H_1:

$$T = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}},$$

  which is approximately $N(0,1)$ distributed under H_0.
- Alternatively, T^2 is approximately χ^2_1.
- Rejection of H_0 occurs when T is large in absolute value, or when T^2 is large.

Poisson regression model: deviance

- Let $\mathcal{L}(\hat{\boldsymbol{\lambda}})$ be the model likelihood, i.e.

$$\mathcal{L}(\hat{\boldsymbol{\lambda}}) = \prod_{i=1}^{n} \exp\left(-\hat{\lambda}_i\right) \frac{\hat{\lambda}_i^{k_i}}{k_i!}.$$

  Note: maximal value of $\lambda \mapsto \exp(-\lambda)\lambda^k/k!$ is obtained for $\lambda = k$.

- Maximum likelihood possible under the Poisson assumption:

$$\mathcal{L}(\boldsymbol{k}) = \prod_{i=1}^{n} \exp\left(-k_i\right) \frac{k_i^{k_i}}{k_i!}.$$

  $\Rightarrow$ likelihood of the saturated model.

Poisson regression model: deviance

- The deviance $D(\boldsymbol{k}, \hat{\boldsymbol{\lambda}})$ is defined as the likelihood ratio test statistic (LRT) for the current model against the saturated model:

$$
\begin{aligned}
D(\boldsymbol{k}, \hat{\boldsymbol{\lambda}}) &= -2 \ln \frac{\mathcal{L}(\hat{\boldsymbol{\lambda}})}{\mathcal{L}(\boldsymbol{k})} = 2 \left(\ln \mathcal{L}(\boldsymbol{k}) - \ln \mathcal{L}(\hat{\boldsymbol{\lambda}}) \right) \\
&= 2 \ln \left(\prod_{i=1}^{n} \exp\left(-k_i\right) \frac{k_i^{k_i}}{k_i!} \right) - 2 \ln \left(\prod_{i=1}^{n} \exp\left(-\hat{\lambda}_i\right) \frac{\hat{\lambda}_i^{k_i}}{k_i!} \right) \\
&= 2 \sum_{i=1}^{n} \left(k_i \ln \frac{k_i}{\hat{\lambda}_i} - (k_i - \hat{\lambda}_i) \right).
\end{aligned}
$$

The smaller the deviance, the bigger is the current model!

Poisson regression model: deviance

- Testing a hypothesis on a set of parameters:

$$\begin{cases} H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-q})' \\ H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = (\beta_0, \beta_1, \ldots, \beta_{p-q}, \beta_{p-q+1}, \ldots, \beta_p)'. \end{cases}$$

- Say D_0 the deviance of the Poisson model under H_0 and D_1 the deviance under H_1.
- Use the test statistic: (drop in deviance)

$$\begin{aligned} \triangle &= D_0 - D_1 = 2\left(L(\boldsymbol{k}) - L(\hat{\boldsymbol{\beta}}_0)\right) - 2\left(L(\boldsymbol{k}) - L(\hat{\boldsymbol{\beta}}_1)\right) \\ &= 2\left(L(\hat{\boldsymbol{\beta}}_1) - L(\hat{\boldsymbol{\beta}}_0)\right) \approx_d \chi^2_q. \end{aligned}$$

Poisson regression model: deviance

- $\triangle$ is a likelihood ratio test statistic.
- The null hypothesis H_0 is rejected in favor of H_1

  if the observed value of the test statistic, $\triangle_{\text{obs}}$, is 'too large', i.e.

$$\triangle_{\text{obs}} > \chi^2_{q,1-\alpha}.$$

Poisson regression model: AIC and BIC

- χ^2 approximation to the distribution of the LRT statistic is valid only when considering nested hypotheses.
- Information criteria such as AIC or BIC are useful with non-nested models: (with $\boldsymbol{\xi}$ the parameter vector used by the model)

$$\begin{aligned} AIC &= -2L(\hat{\boldsymbol{\xi}}) + 2\dim(\boldsymbol{\xi}); \\ BIC &= -2L(\hat{\boldsymbol{\xi}}) + \ln(n)\dim(\boldsymbol{\xi}). \end{aligned}$$

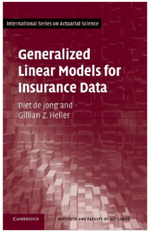
  Both criteria are:

  - minus 2 times the maximum log–likelihood;
  - penalized by a function of the number of parameters and sample size.

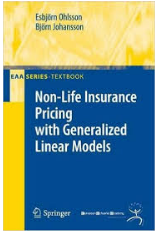
Theoretical fundamentals on GLMs

GLMs: references

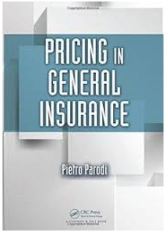
- Non–life insurance pricing with GLMs:



de Jong & Heller



Ohlsson & Johansson



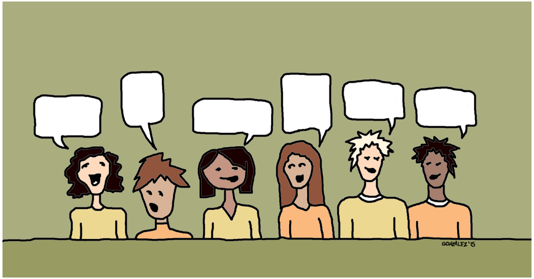
Parodi

- See also the lecture sheets by prof. Claudia Czado on GLMs.

Class discussion

Applications of GLMs in other insurance applications?

- Life insurance
- Health or disability insurance



GLMs

- Consider independent response variables with covariates.
- In a GLM a transformation of the mean is modelled with a linear predictor, i.e. $\boldsymbol{x}_i^{'}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$.
- GLM's allow to include nonnormal errors such as binomial, Poisson and Gamma errors.
- Regression parameters are estimated using Maximum Likelihood Estimation (MLE).

Exponential family

- Class of distributions for which the theory of Generalized Linear Models has been developed.

- Density from the exponential family can be written as:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right),$$

for certain $a(.)$, $b(.)$ en $c(.)$.

Exponential family

- A (general) expression for

  the mean and the variance

  of a distribution from the exponential family.

$$\begin{aligned} \mu &= EY = b^{'}(\theta) \\ \text{Var}(Y) &= b^{''}(\theta) \cdot a(\phi). \end{aligned}$$

Exponential family

- The variance of Y is the product of two functions:
  - $b^{''}(\theta)$: only depends on θ and thus μ (since $\mu = b^{'}(\theta)$)

    this is the variance function $V(\mu)$

    (with $b^{''}((b^{'})^{-1}(\mu)) := V(\mu)$)
  - $a(\phi)$ is often of the form ϕ/w

    with w a weight and ϕ the dispersion parameter.

Model components

- Response Y_i and independent variables $x_i^{'} = (x_{i1}, \ldots, x_{ip})$ for $i = 1, \ldots, n$.

  (1) Random component: Y_i with $1 \leq i \leq n$ independent with density from the exponential family, i.e.

  $$f(y; \theta, \phi) \quad = \quad \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right).$$

  Here ϕ is a dispersion parameter and functions $b(.)$, $a(.)$ and $c(.,.)$ are known.

  (2) Systematic component: $\eta_i(\boldsymbol{x}_i^{'}\boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ the linear predictor, with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ regression parameters.

  (3) Parametric link function: the link function $g(\mu_i) = \eta_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta}$ combines the linear predictor with the mean $\mu_i = E[Y_i]$.

Link functions

- Recall: $\eta = \boldsymbol{x}'\boldsymbol{\beta}$, $\eta = g(\mu)$ and $E[Y] = \mu$.
- g is monotone, differentiable: the link function.
- Normal, linear regression: $\mu \in \mathbb{R}$, $\eta \in \mathbb{R}$, thus $g : \mathbb{R} \to \mathbb{R}$.
  Often, we use $g(\mu) = \mu = \boldsymbol{x}'\boldsymbol{\beta}$; other possibility

$$g_\alpha(\mu) = \begin{cases} \frac{\mu^\alpha - 1}{\alpha}, & \alpha \neq 0 \\ \log(\mu), & \alpha = 0. \end{cases}$$

  This is the Box–Cox class of transformations.
- Poisson regression: $\mu > 0$, $g : \mathbb{R}^+ \to \mathbb{R}$

$$g(\mu) = \log(\mu).$$

Link functions

- Binomial (proportion): $\mu = p \in [0,1]$, we need a monotone $g : [0,1] \to \mathbb{R}$. Use $g(\mu) := F^{-1}(\mu)$, with $F(.)$ a cdf.

  (a) With $F(.)$ the cdf of the logistic distribution:

  $$\begin{aligned} F(z) &= \frac{e^z}{1+e^z} \\ &\Rightarrow \quad g(\mu) := F^{-1}(\mu) = \log\left(\frac{\mu}{1-\mu}\right). \end{aligned}$$

  This is the logit link, used in logistic regression.

  (b) With $F(.)$ the cdf of the standard normal distribution:

  $$\begin{aligned} F(z) &= \Phi(z) \\ &\Rightarrow \quad g(\mu) = \Phi^{-1}(\mu). \end{aligned}$$

  This is the probit link, used in probit regression.

Maximum Likelihood Estimation

- For an illustration of MLE equations in a Poisson model, cfr infra.
- To solve the ML equations numerically, use:
  - Newton–Raphson
  - or Fisher scoring
  - which can be written as a weighted iterative least squares algorithm.

Statistical inference

- Start from a regression model with p explanatory variables and test

$$
\begin{aligned}
H_0 &: \quad \beta_{p-q+1} = \ldots = \beta_p = 0 \\
H_1 &: \quad \text{not } H_0.
\end{aligned}
$$

- Drop–in–deviance test statistic: ($\sim$ extra–sum–of–squares F–test)
  - scaled(!) deviance from the reduced model - scaled(!) deviance from the larger model
  - (sum of squared residuals from the reduced model) - (sum of squared residuals from the larger model), appropriately scaled.
- With GLMs, a new type of residual is used: deviance residual.

Statistical inference

- Denote with $L(H_0)$, resp. $L(H_1)$, the log–likelihood under H_0, resp. H_1.
- $L(S)$ is the log–likelihood under the saturated model.

$$
\begin{aligned}
-2\log\frac{\mathcal{L}(H_0)}{\mathcal{L}(H_1)} &= -2[L(H_0)-L(H_1)] \\
&= -2[(L(H_0)-L(S))-(L(H_1)-L(S))] \\
&= \frac{\text{DEV}(X_1,\ldots,X_{p-q})}{\phi}-\frac{\text{DEV}(X_1,\ldots,X_p)}{\phi},
\end{aligned}
$$

where

- $\text{DEV}(X_1,\ldots,X_{p-q})$ is the deviance of the reduced model
- $\text{DEV}(X_1,\ldots,X_p)$ is the deviance of the larger model.

Statistical inference

- Small: the reduced model does about as good a job as the larger model at explaining the responses.
- Large: the reduced model is relatively inadequate.

Drop–in–deviance χ^2 test: to test the significance of a set of q predictor variables, use the difference in scaled deviances

$$\text{DEV}(X_1, \ldots, X_{p-q})/\phi - \text{DEV}(X_1, \ldots, X_p)/\phi,$$

when ϕ is known.

Under H_0 this test statistic has a χ^2 distribution with q degrees of freedom.

Statistical inference

- Estimate ϕ when it is unknown.
- Possible estimator: (from the larger model under H_1)

$$\hat{\phi} = \frac{\text{Deviance}}{\text{Degrees of freedom}},$$

with the degrees of freedom $= n - (p + 1)$.

- The test statistic then becomes

$$F = \frac{\text{Drop-in-deviance}/q}{\hat{\phi}},$$

with $q =$ difference in number of parameters between the model under H_1 and H_0.

Statistical inference

- The corresponding p–value is then given by

$$P(F_{q,n-(p+1)} > F - \text{statistic}),$$

with $p+1$ the number of parameters in the model under H_1.

- Notice the analogy with the partial F–test for general linear models.

Wrap–up

After this class you are able to:

- explain what a GLM is: identification of the distributional framework, the main concepts (distribution, link function and linear predictor), identify differences with LMs, ...
- specify typical examples of GLMs: Poisson, gamma, logistic, probit regression
- understand and interpret output from a GLM analysis
- explain and apply main inferential methods in GLMs: Wald test, deviance statistic, drop–in–deviance test
- start working with GLM analyses in a statistical software package.