# Generalized Linear Model

Siyu Wang,Yu Zheng

School of Management
University of Science and Technology of China

2021.11.10

# Table of Contents

1. Generalized Linear Model
   - Introduction
   - Model Specification
   - Estimation
   - Practical Aspects
2. Overdispersion
   - Overdispersion for Binomial Model
   - Overdispersion for Poisson Model
3. Quasi-likelihood
   - Iteratively Reweighted Least Squares Algorithm
   - Conclusion

# Table of Contents

1. Generalized Linear Model
   - Introduction
   - Model Specification
   - Estimation
   - Practical Aspects
2. Overdispersion
   - Overdispersion for Binomial Model
   - Overdispersion for Poisson Model
3. Quasi-likelihood
   - Iteratively Reweighted Least Squares Algorithm
   - Conclusion

# Introduction

- Generalized linear models (GLM) extend the concept of the well understood linear regression model.
- Ordinary linear regression predicts the expected value of the response variable $Y$, as a linear combination of a set of predictors $\boldsymbol{X}$,i.e.

$$E(Y \mid \boldsymbol{X}) = \boldsymbol{X}^\top \beta.$$

# Introduction

- Unfortunately, the restriction to linearity cannot take into account the variety of practical situations. For example, a continuous distribution of the error $\varepsilon$ term implies that the response $Y$ must have a continuous distribution as well. Hence, the linear regression model may fail when dealing with binary $Y$ or with counts.
- Generalized linear models cover all these situations by allowing that the response variables follow arbitrary distributions (rather than simply normal distributions).
- Through a link function, the transformed value of the conditional expectation for response variables could vary linearly with the predictors.

# Example 1

- **Example 1**(Bernoulli responses)
  Let us illustrate a binary response model (Bernoulli $Y$ ) using a sample on credit worthiness. For each individual in the sample we know if the granted loan has defaulted or not. The responses are coded as

$$Y = \begin{cases} 1 & \text{loan defaults,} \\ 0 & \text{otherwise .} \end{cases}$$

  The term of interest is how credit worthiness depends on observable individual characteristics $\boldsymbol{X}$ (age, amount and duration of loan, employment, purpose of loan, etc.).

# Example 1

Recall that for a Bernoulli variable $P(Y = 1 \mid \boldsymbol{X}) = E(Y \mid \boldsymbol{X})$ holds. Hence, the default probability $P(Y = 1 \mid \boldsymbol{X})$ equals a regression of $Y$ on $\boldsymbol{X}$. A useful approach is the following logit model:

$$P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1+\exp\left(-\boldsymbol{x}^{\top}\boldsymbol{\beta}\right)}.$$

Here the function of interest $E(Y \mid \boldsymbol{X})$ is linked to a linear function of the explanatory variables.

# Introduction

- Nelder and Wedderburn (1972) and McCullagh and Nelder (1989) show that if the distribution of the dependent variable $Y$ is a member of the exponential family, then the class of models which connects the expectation of $Y$ to a linear combination of the variables $\boldsymbol{X}^\top\boldsymbol{\beta}$ can be treated in a unified way.
- In the following sections we denote the function which relates $\mu = E(Y \mid \boldsymbol{X})$ and $\boldsymbol{X}^\top\boldsymbol{\beta}$ by

  $$G(\mu) = \boldsymbol{X}^\top\boldsymbol{\beta}.$$

  This function $G$ is called link function.

# Model Characteristics

We assume that the distribution of Y is a member of the exponential family.

- The exponential family covers a large number of distributions, for example discrete distributions as the Bernoulli, binomial and Poisson which can handle binary and count data, or continuous distributions as the normal, Gamma.

# Exponential Family

- We say that a distribution is a member of the exponential family if its probability mass function (if $Y$ discrete) or its density function (if Y continuous) has the following form:

$$f(y, \theta, \psi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\psi)} + c(y, \psi) \right\}. \tag{1}$$

- The functions $a(\cdot), b(\cdot)$ and $c(\cdot)$ will vary for different $Y$ distributions.
- Our parameter of interest is $\theta$, which is also called the canonical parameter.
- The additional parameter $\psi$, that is only relevant for some of the distributions, is considered as a nuisance parameter.

# Exponential Family

- **Example 2** (Normal distribution)

$$f(y) = \exp\left\{-(y-\mu)^2 / \left(2\sigma^2\right)\right\} / (\sqrt{2\pi}\sigma)$$

can be written as in the form of exponential family by setting $\theta = \mu$ and $\psi = \sigma$ and $a(\psi) = \psi^2, b(\theta) = \theta^2/2$ and

$$c(y, \psi) = -y^2 / \left(2\psi^2\right) - \log(\sqrt{2\pi}\psi).$$

# Exponential Family

- **Example 3** (Bernoulli distribution)

$$P(Y=y)=\mu^{y}(1-\mu)^{1-y}=\begin{cases}\mu & \text{if } y=1,\\ 1-\mu & \text{if } y=0.\end{cases}$$

using the logit transformation $\theta=\log\{\mu/(1-\mu)\}$, it can be transformed into $P(Y=y)=\exp(y\theta)/\left(1+e^{\theta}\right)$. Thus we obtain an exponential family with $a(\psi)=1, b(\theta)=-\log(1-\mu)=\log\left(1+e^{\theta}\right)$, and $c(y,\psi)=0$.

# Exponential Family

- **Example 4** (Poisson distribution)

  $$P(Y = y) = \frac{\mu^y \exp(-\mu)}{y!} = \exp(y \log(\mu) - \mu - \log(y!))$$

  .

  Here, the canonical parameter $\theta = \log(\mu)$ and dispersion parameter $\psi = 1$, taking $a(\psi) = 1$, $b(\theta) = \mu = \exp(\theta)$ and $c(y, \phi) = \log(y!)$. The Poisson distribution is useful for modelling count data. It is particularly useful for situations in which the outcome is a counting variable with approximately the same mean and variance.

# Properties of the Exponential Family

- Noting that $\int f(y,\theta,\psi)dy = 1$,this implies

$$\begin{aligned}0 &= \frac{\partial}{\partial\theta}\int f(y,\theta,\psi)dy = \int \frac{\partial}{\partial\theta} f(y,\theta,\psi)dy \\ &= \int \left\{\frac{\partial}{\partial\theta}\log f(y,\theta,\psi)\right\} f(y,\theta,\psi)dy = E\left\{\frac{\partial}{\partial\theta}\ell(y,\theta,\psi)\right\},\end{aligned}$$

  where $\ell(y,\theta,\psi) = \log f(y,\theta,\psi)$ denotes the log-likelihood function.

# Properties of the Exponential Family

- The derivative of $\ell$ with respect to $\theta$ is called the score function. It is known that

$$
\begin{aligned}
E\frac{\partial}{\partial\theta}\ell(y,\theta,\psi) &= 0\\
E\left\{\frac{\partial^2}{\partial\theta^2}\ell(y,\theta,\psi)\right\} &= -E\left\{\frac{\partial}{\partial\theta}\ell(y,\theta,\psi)\right\}^2.
\end{aligned}
$$

- Therefore,

$$0 = E\left\{\frac{Y-b'(\theta)}{a(\psi)}\right\}, \text{ and } \frac{-b''(\theta)}{a(\psi)} = -E\left\{\frac{Y-b'(\theta)}{a(\psi)}\right\}^2,$$

- We also assume that the factor $a(\psi)$ is identical over all observations.

# Properties of the Exponential Family

- Such that we can conclude
$$\begin{aligned} E(Y) &= \mu = b'(\theta), \\ \mathrm{Var}(Y) &= b''(\theta)a(\psi). \end{aligned}$$
- Since $b''(\theta) > 0$, we define
$$V(\mu) = b''((b')^{-1}(\mu)) = b''(\theta).$$
The variance of $Y$ is also determined by $\mu$.

# Link function

- After specified the distribution of $Y$, the link function $G$ is the second component to choose for the GLM.
- The link function provides the relationship between the linear predictor and the mean of the distribution function.
- Recall the model notation $G(\mu) = \boldsymbol{X}^\top \boldsymbol{\beta}$. For Bernoulli $Y$,the canonical link is given by the logit transformation $\boldsymbol{X}^\top \boldsymbol{\beta} = \log\{\mu/(1-\mu)\}$.

# Link Function

- There exists a number of specific link functions for most of the models exists.
- For Bernoulli $Y$, for example, any smooth cdf can be used. Typical links are the logit and the inverse of standard normal (Gaussian) cdfs, which lead to logit and probit models, respectively. The probit model assume a latent variable $Y^* = \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon$ with standard Gaussian noise $\varepsilon$. $Y^* > 0$ iff $Y = 1$.

# Link Function

- For poisson distribution,the link function is given by $\log(\mu)$.
- For normal distribution,the link function is given by $\mu$.
- For positive $Y$ observations, a flexible class of link functions is the class of power functions. These links are given by the Box-Cox transformation, i.e. $\left(\mu^{\lambda} - 1\right)/\lambda$ for the case $\lambda \neq 0$, and $\log(\mu)$ for $\lambda = 0$. The parameter $\lambda$ is estimated using the profile likelihood function.

# Estimation

- By assuming that the distribution of $Y$ belongs to the exponential family it is possible to derive maximum-likelihood estimates for the coefficients of a GLM.
- We denote now the vector of all response observations by $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and their conditional expectations (given $\boldsymbol{X}_i$) by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$.
- The sample log-likelihood of the vector $\mathbf{Y}$ is given by

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} \ell\left(Y_i, \theta_i, \psi\right). \tag{2}$$

# Estimation

- We now plug-in the exponential family form (1) into (2) and obtain $$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} \left\{ \frac{Y_i\theta_i - b(\theta_i)}{a(\psi)} + c\left(Y_i, \psi\right) \right\}.$$
- Neither $a(\psi)$ nor $c\left(Y_i, \psi\right)$ depend on the unknown parameter vector $\boldsymbol{\beta}$, it is sufficient to consider

$$\sum_{i=1}^{n} \left\{ Y_i\theta_i - b\left(\theta_i\right) \right\} \tag{3}$$

for the maximization.

# Iteratively Reweighted Least Squares Algorithm

- We denote the gradient by

$$\nabla(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \left[ -\sum_{i=1}^{n} \left\{ Y_i \theta_i - b\left(\theta_i\right) \right\} \right] = -\sum_{i=1}^{n} \left\{ Y_i - b'\left(\theta_i\right) \right\} \frac{\partial}{\partial \boldsymbol{\beta}} \theta_i. \tag{4}$$

- Our optimization problem consists in solving

$$\nabla(\boldsymbol{\beta}) = 0. \tag{5}$$

The smoothness of the link function allows us to compute the Hessian of likelihood, which is denoted by $\mathcal{H}(\boldsymbol{\beta})$.

# Iteratively Reweighted Least Squares Algorithm

- A Newton–Raphson algorithm can be applied which determines the optimal $\widehat{\beta}$ using the following iteration steps:
$$\widehat{\boldsymbol{\beta}}^{\text{new}} = \widehat{\boldsymbol{\beta}}^{\text{old}} - \left\{\mathcal{H}\left(\widehat{\boldsymbol{\beta}}^{\text{old}}\right)\right\}^{-1} \nabla\left(\widehat{\boldsymbol{\beta}}^{\text{old}}\right).$$
- A variant of the Newton–Raphson is the Fisher scoring algorithm that replaces the Hessian by its expectation with respect to the observations $Y_i$ :
$$\widehat{\boldsymbol{\beta}}^{\text{new}} = \widehat{\boldsymbol{\beta}}^{\text{old}} - \left\{E\mathcal{H}\left(\widehat{\boldsymbol{\beta}}^{\text{old}}\right)\right\}^{-1} \nabla\left(\widehat{\boldsymbol{\beta}}^{\text{old}}\right).$$

# Iteratively Reweighted Least Squares Algorithm

- Notice that

$$\boldsymbol{X}_i^\top \boldsymbol{\beta} = G(\mu_i) = G(b'(\theta_i)).$$

$b''(\theta_i) = V(\mu_i)$ which implies

$$\boldsymbol{X}_i = G'(\mu_i)V(\mu_i)\tfrac{\partial \theta_i}{\partial \boldsymbol{\beta}},$$

That is

$$\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \frac{1}{G'(\mu_i)V(\mu_i)}\boldsymbol{X}_i.$$

# Iteratively Reweighted Least Squares Algorithm

- From this we can express the gradient and the Hessian of the likelihood by

$$\nabla(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \left(Y_i - \mu_i\right) \frac{1}{G'(\mu_i)V(\mu_i)} \boldsymbol{X}_i,$$

$$\mathcal{H}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \left\{ \left(Y_i - \mu_i\right) A(\boldsymbol{\beta}) - \frac{1}{\left(G'(\mu_i)\right)^2 V(\mu_i)} \right\} \boldsymbol{X}_i \boldsymbol{X}_i^\top.$$

where $A(\boldsymbol{\beta})\boldsymbol{X}_i$ is the derivative of $\frac{1}{G'(\mu_i)V(\mu_i)}$.

# Iteratively Reweighted Least Squares Algorithm

- The expectation of $\mathcal{H}(\beta)$ in the Fisher scoring algorithm equals
$$E\mathcal{H}(\beta) = \sum_{i=1}^{n} \left\{ \frac{1}{(G'(\mu_i))^2 V(\mu_i)} \right\} \boldsymbol{X}_i \boldsymbol{X}_i^\top .$$
- Let us consider only the Fisher scoring algorithm for the moment. We define the weight matrix
$$\mathbf{W} = \operatorname{diag}\left( \frac{1}{(G'(\mu_1))^2 V(\mu_i)}, \ldots, \frac{1}{(G'(\mu_n))^2 V(\mu_n)} \right)$$
and the vectors $\widetilde{\boldsymbol{Y}} = \left( \tilde{Y}_1, \ldots, \tilde{Y}_n \right)^\top, \boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$ by
$$\tilde{Y}_i = G'(\mu_i)(Y_i - \mu_i),\ Z_i = \boldsymbol{X}_i^\top \beta^{\text{old}} + \tilde{Y}_i .$$
- Denote further by $\boldsymbol{X}$ the design matrix given by the rows $x_i^\top$.

# Iteratively Reweighted Least Squares Algorithm

- Then, the Fisher scoring iteration step for $\boldsymbol{\beta}$ can be rewritten as
$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} + \left(\mathbf{X}^\top \mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \widetilde{\boldsymbol{Y}} = \left(\mathbf{X}^\top \mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \boldsymbol{Z}.$$
- This immediately shows that each Fisher scoring iteration step is the result of a weighted least squares regression of the adjusted dependent variables $Z_i$ on the explanatory variables $\boldsymbol{X}_i$.
- Since the weights are recalculated in each step we speak of the iteratively reweighted least squares (IRLS) algorithm.
- The iteration will be stopped when the parameter estimate do not change significantly anymore. We denote the final parameter estimate by $\hat{\boldsymbol{\beta}}$.

# Deviance

- Without putting any model restriction on $\theta_i$, we would have to maximize
$$\sum_{i=1}^{n} \{Y_i\theta_i - b(\theta_i)\}$$
with respect to $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T$. So the unrestricted maximizer is $\tilde{\theta}_i = (b')^{-1}(Y_i)$, by taking the derivative and setting it to zero.
- The deviance is defined as [1]
$$D(\mathbf{Y}; \widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} \left\{ \left( Y_i\tilde{\theta}_i - b\left(\tilde{\theta}_i\right)\right) - \left( Y_i\hat{\theta}_i - b\left(\hat{\theta}_i\right)\right)\right\}$$
- $D^*(\mathbf{Y}; \widehat{\boldsymbol{\mu}}) = D(\mathbf{Y}; \widehat{\boldsymbol{\mu}})/a(\psi)$is called the scaled deviance. The deviance measures the goodness of fitting for given model.

# Model selection

- For model choice between two nested models, a likelihood ratio test (LR test) is used. Assume that $\mathcal{M}_0$ ($p_0$ parameters) is a submodel of the model $\mathcal{M}$ ( $p$ parameters) and that we have estimated them as $\widehat{\boldsymbol{\mu}}_0$ and $\widehat{\boldsymbol{\mu}}$. For one-parameter exponential families (without a nuisance parameter $\psi$ ) we use that asymptotically

$$D\left(\boldsymbol{Y};\hat{\boldsymbol{\mu}}_0\right)-D(\boldsymbol{Y};\hat{\boldsymbol{\mu}})\sim\chi^2_{p-p_0}. \tag{6}$$

- Model selection procedures for possibly non-nested models can be based on Akaike's information criterion

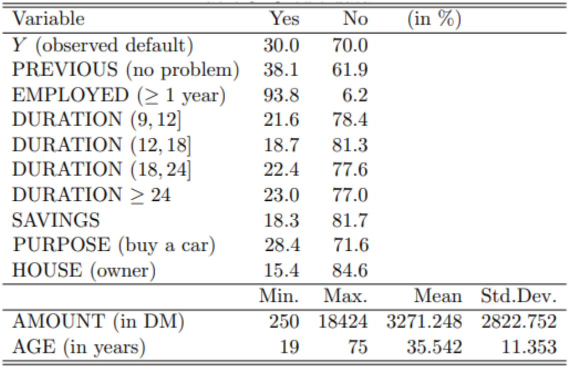
$$AIC=D(\boldsymbol{Y};\widehat{\boldsymbol{\mu}})+2p,$$

or Schwarz' Bayes information criterion $BIC=D(\boldsymbol{Y};\widehat{\boldsymbol{\mu}})+\log(n)p$, where $p$ denotes the number of estimated parameters.

# Practical Aspects

- To illustrate the GLM in practice we recall Example 1 on credit worthiness.The credit data set that we use contains n = 1000 observations.
- The model for credit worthiness is based on the idea that default can be predicted from the individual and loan characteristics.We consider criteria as **age, information on previous loans, savings, purpose,employment and house ownership** to characterize the credit applicants.**Amount and duration** of the loan are prominent features of the granted loans.We remark that we have categorized the durations (months) into intervals since most of the realizations are multiples of 3 or 6 months.Some descriptive statistics can be found in Figure 1.

# Practical Aspects



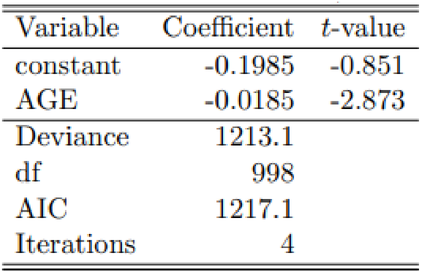
| Variable | Yes | No | (in %) | |
|---|---|---|---|---|
| $Y$ (observed default) | 30.0 | 70.0 | | |
| PREVIOUS (no problem) | 38.1 | 61.9 | | |
| EMPLOYED ($\geq 1$ year) | 93.8 | 6.2 | | |
| DURATION $(9, 12]$ | 21.6 | 78.4 | | |
| DURATION $(12, 18]$ | 18.7 | 81.3 | | |
| DURATION $(18, 24]$ | 22.4 | 77.6 | | |
| DURATION $\geq 24$ | 23.0 | 77.0 | | |
| SAVINGS | 18.3 | 81.7 | | |
| PURPOSE (buy a car) | 28.4 | 71.6 | | |
| HOUSE (owner) | 15.4 | 84.6 | | |
| | Min. | Max. | Mean | Std.Dev. |
| AMOUNT (in DM) | 250 | 18424 | 3271.248 | 2822.752 |
| AGE (in years) | 19 | 75 | 35.542 | 11.353 |

Figure 1: Credit data.

- Recall that for binary $Y$ it holds $P(Y=1 \mid \boldsymbol{X}) = E(Y \mid \boldsymbol{X})$,Our first approach is a GLM with logit link such that $P(Y=1 \mid \boldsymbol{X}) = \exp\left(\boldsymbol{X}^\top \beta\right) / \left\{1 + \exp\left(\boldsymbol{X}^\top \beta\right)\right\}$.

# Practical Aspects

- **Example 5** (Credit default on AGE)
  We initially estimate the default probability solely related to age, i.e. the model
  $$P(Y=1\mid AGE)=\frac{\exp(\beta_0+\beta_1 AGE)}{1+\exp(\beta_0+\beta_1 AGE)}$$
  or equivalently $\mathrm{logit}\{P(Y=1\mid AGE)\}=\beta_0+\beta_1 AGE$.



| Variable | Coefficient | $t$-value |
|---|---|---|
| constant | -0.1985 | -0.851 |
| AGE | -0.0185 | -2.873 |
| Deviance | 1213.1 | |
| df | 998 | |
| AIC | 1217.1 | |
| Iterations | 4 | |

Figure 2: Credit default on AGE (logit model).

# Practical Aspects

- From the table we see that the estimated coefficient of AGE has a negative sign.
- Since the link function and its inverse are strictly monotone increasing, we can conclude that the probability of default must thus be decreasing with increasing AGE.
- The following figure 3 shows on the left frequency barplots of AGE separately for $Y = 1$ and $Y = 0$.
- we can recognize clearly the decreasing propensity to default. The right graph in Figure 3 displays the estimated probabilities $P(Y = 1 \mid AGE)$ using the fitted logit model which are indeed decreasing.

# Practical Aspects

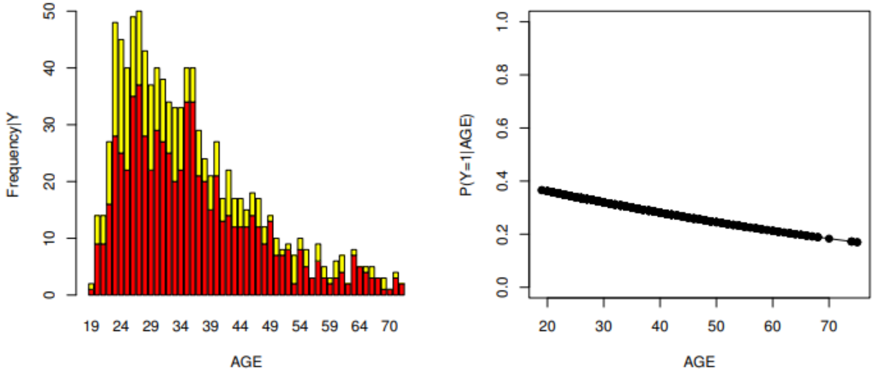


Figure 3: Credit default on AGE, left: frequency barplots of AGE for $Y = 1$ (yellow) and $Y = 0$ (red), right: estimated probabilities.

# Practical Aspects

- Models using different link functions cannot be directly compared as the link functions might be differently scaled.
- In our binary response model for example a logit or a probit link function may be reasonable. However, the variance parameter of the standard logistic distribution is $\pi^2/3$ whereas that of the standard normal is 1.
- We therefore need to rescale one of the link functions in order to compare the resulting model fits.
- The following figure 3 shows the standard logistic cdf (the inverse logit link) against the cdf of $N\left(0, \pi^2/3\right)$.

# Practical Aspects

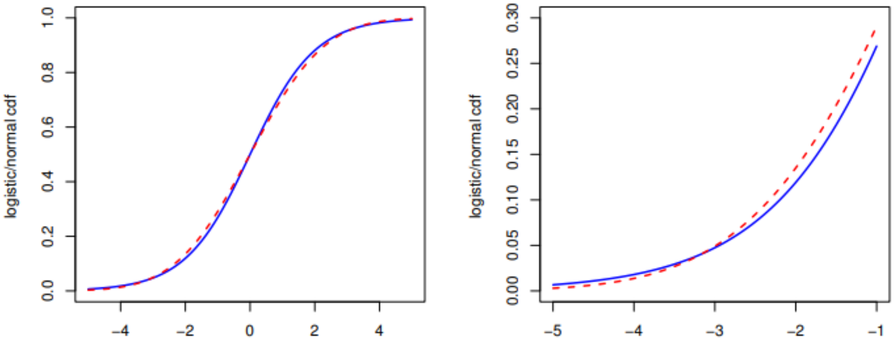


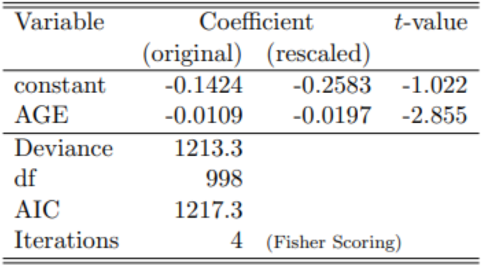
Figure 4: Logit (solid blue) versus appropriately rescaled probit link (dashed red), left: on the range $[-5, 5]$, right: on the range of $[-5, -1]$.

# Practical Aspects

- The functions in the left graph of Figure 4 are hardly distinguishable.
- If we zoom in (right graph) we see that the logistic cdf vanishes to zero at the left boundary at a lower rate.
- This holds similarly for the right boundary and explains the ability of logit models to (slightly) better handle the case of extremal observations.

# Practical Aspects

- **Example 6** (Probit versus logit)
  If we want to compare the estimated coefficients from a probit to that of the logit model we need to rescale the probit coefficients by $\pi/\sqrt{3}$.



| Variable | Coefficient | | $t$-value |
|---|---|---|---|
| | (original) | (rescaled) | |
| constant | -0.1424 | -0.2583 | -1.022 |
| AGE | -0.0109 | -0.0197 | -2.855 |
| Deviance | 1213.3 | | |
| df | 998 | | |
| AIC | 1217.3 | | |
| Iterations | 4 | (Fisher Scoring) | |

Figure 5: Credit default on AGE (probit model), original and rescaled coefficients for comparison with logit.

The resulting rescaled coefficient for AGE in is of similar size as that for the logit model.

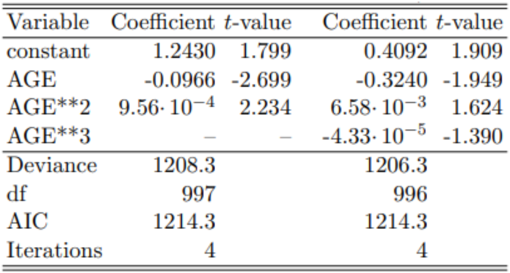
# Practical Aspects

The next two examples intend to analyze if the fit could be improved by using a nonlinear function on AGE instead of $\theta = \beta_0 + \beta_1 \mathrm{AGE}$.Two principally different approaches are possible:

- include higher order terms of AGE into $\theta$,
- categorize AGE in order to fit a stepwise constant $\theta$ function.

# Practical Aspects

- **Example 7** (Credit default on polynomial AGE)
  We fit two logit models using second and third order terms in AGE.The estimated coefficients are presented as follows:



| Variable | Coefficient | $t$-value | Coefficient | $t$-value |
|---|---|---|---|---|
| constant | 1.2430 | 1.799 | 0.4092 | 1.909 |
| AGE | -0.0966 | -2.699 | -0.3240 | -1.949 |
| AGE**2 | $9.56 \cdot 10^{-4}$ | 2.234 | $6.58 \cdot 10^{-3}$ | 1.624 |
| AGE**3 | – | – | $-4.33 \cdot 10^{-5}$ | -1.390 |
| Deviance | 1208.3 | | 1206.3 | |
| df | 997 | | 996 | |
| AIC | 1214.3 | | 1214.3 | |
| Iterations | 4 | | 4 | |

Figure 6: Credit default on polynomial AGE (logit model).

# Practical Aspects

- A comparison of the quadratic fit and the linear fit from Example 5 using the LR test statistic (6) shows that the linear fit is rejected at a significance level of 3%.
- A subsequent comparison of the quadratic against the cubic fit no significant improvement by the latter model.
- Thus, the quadratic term for AGE improves the fit whereas the cubic term does not show any further statistically significant improvement.
- This result is confirmed when we compare the AIC values of both models which are practically identical.
- The following figure 7 shows the estimated default probabilities for the quadratic (left) and cubic AGE fits.

# Practical Aspects

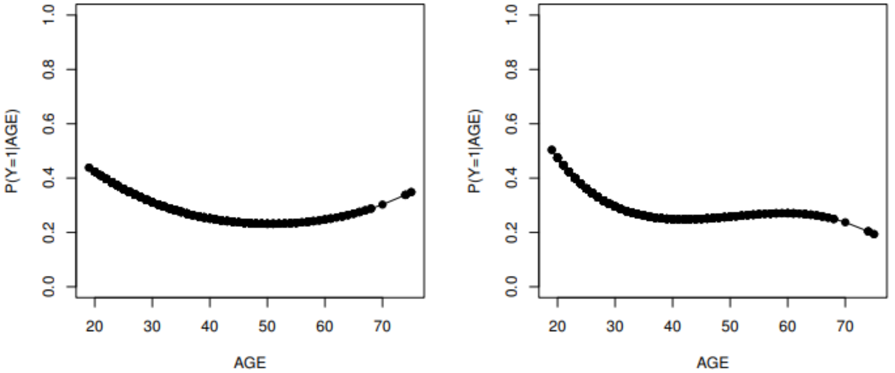


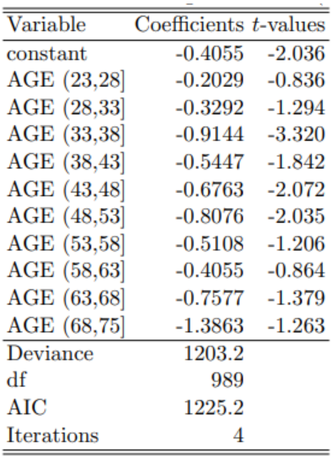
Figure 7: Credit default on polynomial AGE, left: estimated probabilities from quadratic function,right: estimated probabilities from cubic function.

We find that the curves are of similar shape.

# Practical Aspects

- **Example 8**(Credit default on categorized AGE)
  We have chosen the intervals $(18, 23], (23, 28], \ldots, (68, 75]$ as categories.The first interval (18, 23] is chosen for the reference such that we will estimate coefficients only for the remaining 10 intervals.The resulting coefficients for this model are showed in Figure 8.Frequency barplots for the intervals and estimated default probabilities are displayed in Figure 9.

# Practical Aspects



| Variable | Coefficients | $t$-values |
|---|---|---|
| constant | -0.4055 | -2.036 |
| AGE (23,28] | -0.2029 | -0.836 |
| AGE (28,33] | -0.3292 | -1.294 |
| AGE (33,38] | -0.9144 | -3.320 |
| AGE (38,43] | -0.5447 | -1.842 |
| AGE (43,48] | -0.6763 | -2.072 |
| AGE (48,53] | -0.8076 | -2.035 |
| AGE (53,58] | -0.5108 | -1.206 |
| AGE (58,63] | -0.4055 | -0.864 |
| AGE (63,68] | -0.7577 | -1.379 |
| AGE (68,75] | -1.3863 | -1.263 |
| Deviance | 1203.2 | |
| df | 989 | |
| AIC | 1225.2 | |
| Iterations | 4 | |

Figure 8: Credit default on categorized AGE (logit model).

# Practical Aspects

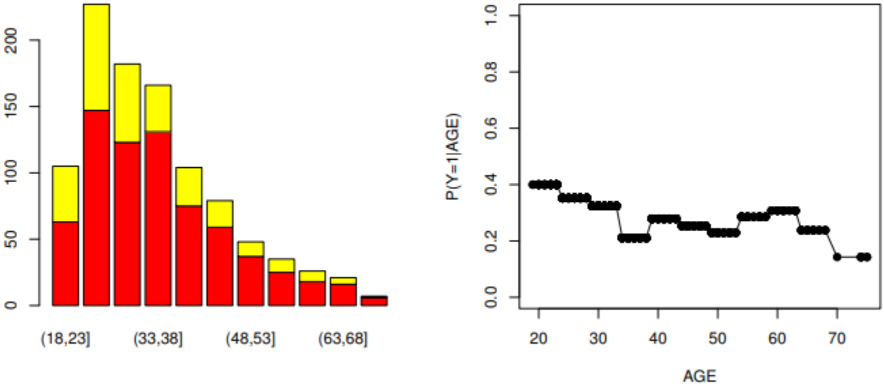


Figure 9: Credit default on categorized AGE, left: frequency barplots of categorized AGE for $Y = 1$ (yellow) and $Y = 0$ (red),right: estimated probabilities.

# Practical Aspects

- We see here that all coefficient estimates are negative.
- This means, keeping in mind that the group of youngest credit applicants is the reference, that all applicants from other age groups have an (estimated) lower default probability.
- However, we do not have a true decrease in the default probabilities with AGE since the coefficients do not form a decreasing sequence.
- In the range from age 33 to 63 we find two local minima and maxima for the estimated default probabilities.

# Practical Aspects

- It is interesting to note that the deviance of the categorized AGE fit is the smallest that we obtained up to now.
- This is explained by the fact that we have fitted the most flexible model here. Unfortunately, this flexibility pays with the number of parameters.
- The AIC criterion as a compromise between goodness-of-fit and number of parameters states that all previous fitted models are preferable.
- Nevertheless, categorization is a valuable tool to explore if there are nonlinear effects.

# Practical Aspects

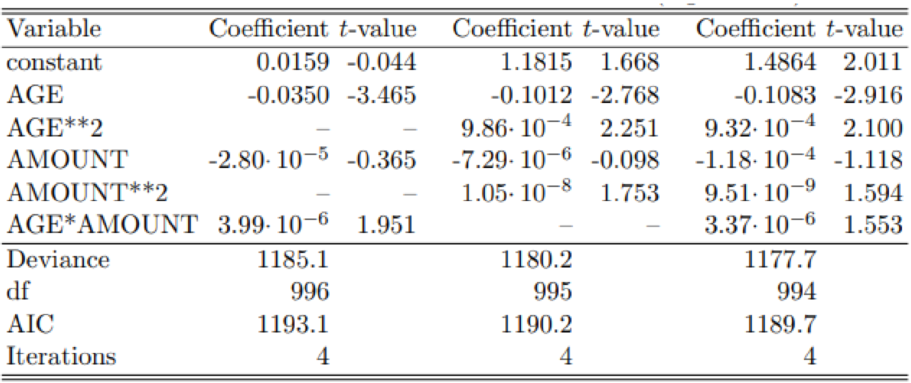
Before fitting the full model with all available information, we discuss the modeling of interaction effects.

- **Example 9** (Credit default on AGE and AMOUNT)
  The variable AMOUNT is the second continuous explanatory variable in the credit data set.We will therefore use AGE and AMOUNT to illustrate the effects of the simultaneous use of two explanatory variables.A very simple model is of course $\text{logit}\{P(Y=1 \mid AGE, AMOUNT)\} = \beta_0 + \beta_1 AGE + \beta_2$ AMOUNT.This model, however, separates the impact of AGE and AMOUNT into additive components.The effect of having both characteristics simultaneously is modeled by adding the multiplicative interaction term AGE$*$AMOUNT.

# Practical Aspects

On the other hand we have seen that at least AGE should be complemented by a quadratic term. For that reason we compare the linear interaction model $\text{logit}\{P(Y = 1 \mid AGE, AMOUNT)\} = \beta_0 + \beta_1 AGE + \beta_2 AMOUNT + \beta_3 AGE*$ AMOUNT with a specification using quadratic terms and a third model specification using both, quadratic and interaction terms.the results for all three fitted models are showed in Figure 10.

# Practical Aspects



| Variable | Coefficient | $t$-value | Coefficient | $t$-value | Coefficient | $t$-value |
|---|---|---|---|---|---|---|
| constant | 0.0159 | -0.044 | 1.1815 | 1.668 | 1.4864 | 2.011 |
| AGE | -0.0350 | -3.465 | -0.1012 | -2.768 | -0.1083 | -2.916 |
| AGE**2 | – | – | $9.86\cdot 10^{-4}$ | 2.251 | $9.32\cdot 10^{-4}$ | 2.100 |
| AMOUNT | $-2.80\cdot 10^{-5}$ | -0.365 | $-7.29\cdot 10^{-6}$ | -0.098 | $-1.18\cdot 10^{-4}$ | -1.118 |
| AMOUNT**2 | – | – | $1.05\cdot 10^{-8}$ | 1.753 | $9.51\cdot 10^{-9}$ | 1.594 |
| AGE*AMOUNT | $3.99\cdot 10^{-6}$ | 1.951 | – | – | $3.37\cdot 10^{-6}$ | 1.553 |
| Deviance | 1185.1 | | 1180.2 | | 1177.7 | |
| df | 996 | | 995 | | 994 | |
| AIC | 1193.1 | | 1190.2 | | 1189.7 | |
| Iterations | 4 | | 4 | | 4 | |

Figure 10: Credit default on AGE and AMOUNT (logit model).

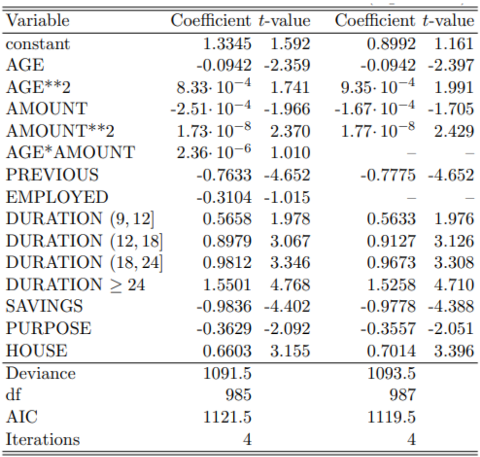
The model with quadratic and interaction terms has the smallest AIC of the three fits.

# Practical Aspects

In a final analysis we present now the results for the full set of variables.

- **Example 10** (Credit default on the full set of explanatory variables) We first estimated a logit model using all variables (AGE and AMOUNT also with quadratic and interaction terms).Most of the estimated coefficients in the second column of Figure 11 have the expected sign.

# Practical Aspects



| Variable | Coefficient | $t$-value | Coefficient | $t$-value |
|---|---|---|---|---|
| constant | 1.3345 | 1.592 | 0.8992 | 1.161 |
| AGE | -0.0942 | -2.359 | -0.0942 | -2.397 |
| AGE**2 | $8.33 \cdot 10^{-4}$ | 1.741 | $9.35 \cdot 10^{-4}$ | 1.991 |
| AMOUNT | $-2.51 \cdot 10^{-4}$ | -1.966 | $-1.67 \cdot 10^{-4}$ | -1.705 |
| AMOUNT**2 | $1.73 \cdot 10^{-8}$ | 2.370 | $1.77 \cdot 10^{-8}$ | 2.429 |
| AGE*AMOUNT | $2.36 \cdot 10^{-6}$ | 1.010 | – | – |
| PREVIOUS | -0.7633 | -4.652 | -0.7775 | -4.652 |
| EMPLOYED | -0.3104 | -1.015 | – | – |
| DURATION $(9, 12]$ | 0.5658 | 1.978 | 0.5633 | 1.976 |
| DURATION $(12, 18]$ | 0.8979 | 3.067 | 0.9127 | 3.126 |
| DURATION $(18, 24]$ | 0.9812 | 3.346 | 0.9673 | 3.308 |
| DURATION $\geq 24$ | 1.5501 | 4.768 | 1.5258 | 4.710 |
| SAVINGS | -0.9836 | -4.402 | -0.9778 | -4.388 |
| PURPOSE | -0.3629 | -2.092 | -0.3557 | -2.051 |
| HOUSE | 0.6603 | 3.155 | 0.7014 | 3.396 |
| Deviance | 1091.5 | | 1093.5 | |
| df | 985 | | 987 | |
| AIC | 1121.5 | | 1119.5 | |
| Iterations | 4 | | 4 | |

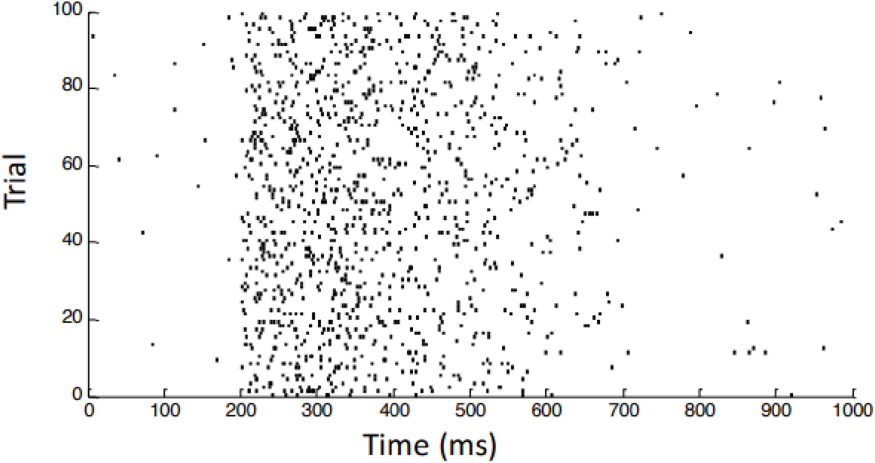
Figure 11: Credit default on full set of variables (logit model).

# Practical Aspects

- In addition to binary data, we next consider countable data, such as Poisson data.The range of $y$ is $\{0, 1, 2, \dots\}$,the probability mass $f(y) = \frac{\mu^y}{y!} e^{-\mu}$,where $\mu$ is the expectation of y.
- the Canonical link function for Poisson data is $\log(\mu)$, i.e, $\log(\mu) = \boldsymbol{X}^\top \boldsymbol{\beta}$.
- We next show the example of a inhomogeneous Poisson model.

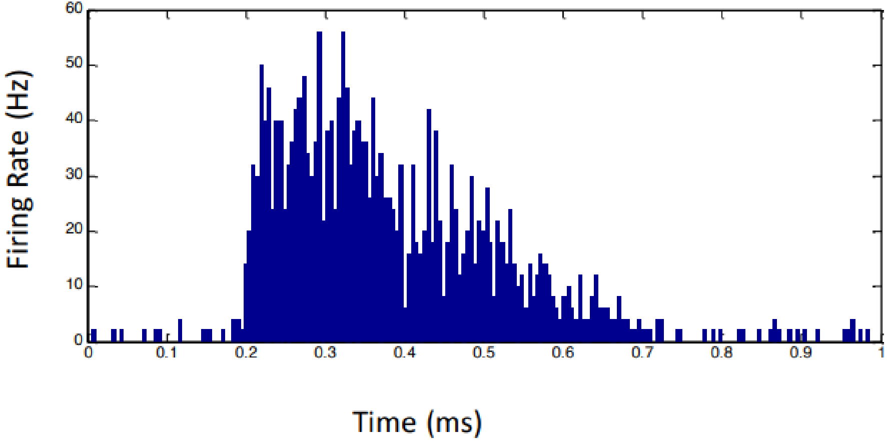
# Example: Inhomogeneous Poisson Model

- Construct an inhomogeneous Poisson spiking model for repeated trial data as a function of time.



# Example: Inhomogeneous Poisson Model

- Construct an inhomogeneous Poisson spiking model for repeated trial data as a function of time.



# Example: inhomogeneous Poisson Model

- For an inhomogeneous Poisson model for repeated trial data as a function of time

  Polynomial model:

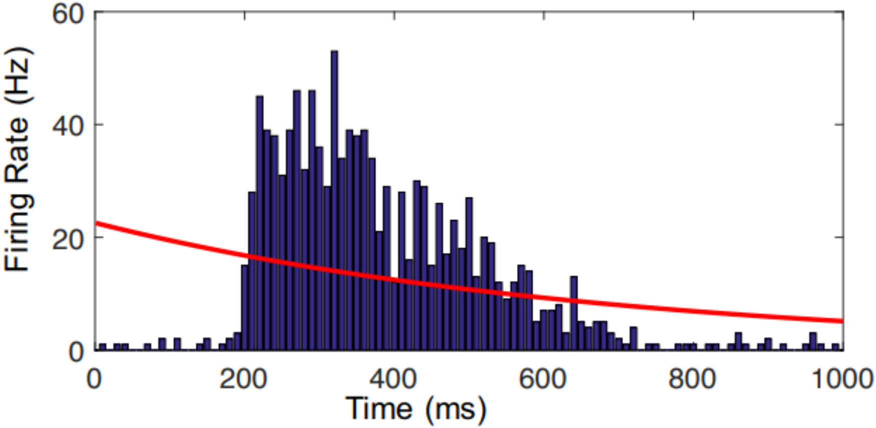
$$\log\left(\mu_t\right) = \beta_0 + \sum_{j=1}^{p} \beta_j t^j,$$

  or:

$$\mu_t = e^{\beta_0 + \sum_{j=1}^{p} \beta_j t^j}.$$

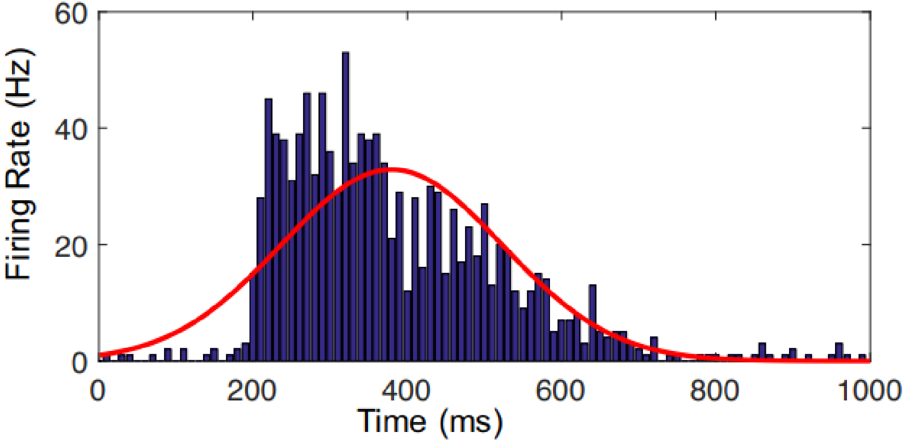
# Example: Inhomogeneous Poisson Model

- Inhomogeneous Poisson GLM using $1^{st}$ order polynomial in time.



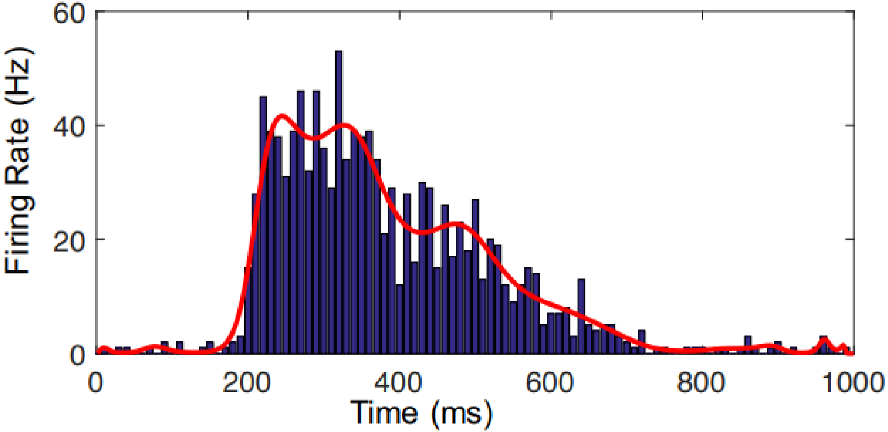
# Example: Inhomogeneous Poisson Model

- Inhomogeneous Poisson GLM using $2^{nd}$ order polynomial in time.



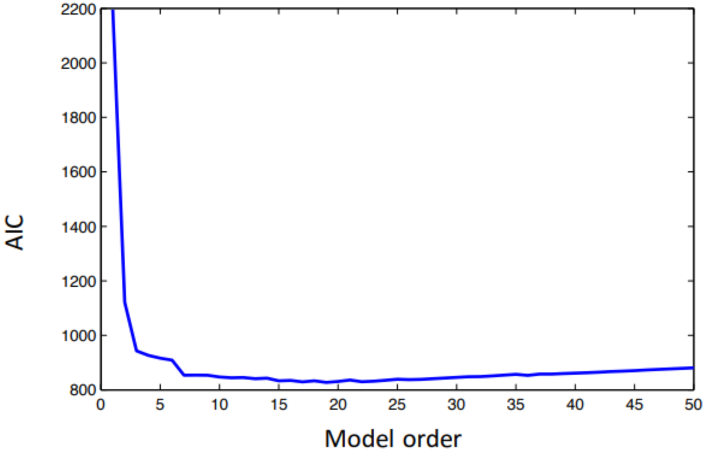
# Example: Inhomogeneous Poisson Model

- Inhomogeneous Poisson GLM using $50^{th}$ order polynomial in time.



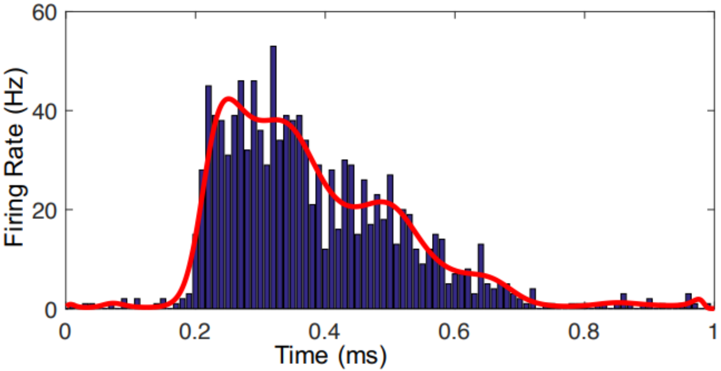
# Model order selection

- AIC plot of polynomial model order.



# Example:Inhomogeneous Poisson Model

- AIC minimized for order 19 polynomial.



# Table of Contents

1. Generalized Linear Model
   - Introduction
   - Model Specification
   - Estimation
   - Practical Aspects
2. Overdispersion
   - Overdispersion for Binomial Model
   - Overdispersion for Poisson Model
3. Quasi-likelihood
   - Iteratively Reweighted Least Squares Algorithm
   - Conclusion

# Overdispersion

What

- In statistics, overdispersion is the presence of greater variability (statistical dispersion) in a data set than would be expected based on a given statistical model.

# Overdispersion

Why

- A common task in applied statistics is choosing a parametric model to fit a given set of empirical observations.
- However, especially for simple models with few parameters, theoretical predictions may not match empirical observations for higher moments.
- When the observed variance is higher than the variance of a proposed model, **overdispersion** has occurred. Conversely, **underdispersion** means that there was less variation in the data than predicted.

# Overdispersion

When

- Correlation between individual responses. For example, in cancer studies involving litters of rats we may expect to see some correlation between rats in the same litter.
- Omitted unobserved variables.
- Variability of experimental material. This can be thought of as individual variability of the experimental units and may give an additional component of variability which is not accounted for by the basic model.

# Model for Overdispersion-Binomial

**Beta-Binomial**

- As a more concrete example, it has been observed that the number of boys born to families does not conform faithfully to a binomial distribution as might be expected. Instead, the sex ratios of families seem to skew toward either boys or girls i.e. there are more all-boy families, more all-girl families and not enough families close to the population 51:49 boy-to-girl mean ratio than expected from a binomial distribution, and the resulting empirical variance is larger than specified by a binomial model.
- One can think of the probability parameter of the binomial model (say, probability of being a boy) as itself a random variable (i.e. random effects model) drawn for each family from a beta distribution as the mixing distribution.[2]

# Model for Overdispersion-Binomial

- $Y_i$ is the "positive" number of $i$th family, which following

$$Y_i|P_i \sim Bin(m_i, P_i).$$

- Writing $Y_i = \sum_{j=1}^{m_i} R_{ij}$ , where $R_{ij}$ are Bernoulli random variables with

$$E(R_{ij}) = \pi_i \; and \; Var(R_{ij}) = \pi_i(1 - \pi_i).$$

# Model for Overdispersion-Binomial

- Assuming a constant correlation $\rho$ between the $R_{ij}$ 's for $j \neq k$, we have
$$Cov(R_{ij}, R_{ik}) = \rho\pi_i(1 - \pi_i)$$
- So
$$E(Y_i) = m_i\pi_i,$$
$$\begin{aligned} Var(Y_i) &= \sum_{j=1}^{m_i} Var(R_{ij}) + \sum_{j=1}^{m_i}\sum_{k\neq j} Cov(R_{ij}, R_{ik}) \\ &= m_i\pi_i(1 - \pi_i)[1 + \rho(m_i - 1)]. \end{aligned}$$
- The variance of $Y_i$ would increase with $\rho$.

# Model for Overdispersion-Binomial

A special case of this is the beta-binomial distribution, which is obtained by assuming that $P_i \sim Beta(\alpha_i, \beta_i)$, with $\alpha_i + \beta_i$ is fixed.

**Beta-Binomial**

- Assume

$$Y_i|P_i \sim Bin(m_i, P_i)$$

$$P_i \sim Beta(\alpha_i, \beta_i)$$

- So

$$E(P_i) = \frac{\alpha}{\alpha + \beta} := \pi_i,$$

$$Var(P_i) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{1}{\alpha+\beta+1}\pi_i(1-\pi_i).$$

# Model for Overdispersion-Binomial

**Beta-Binomial**

- Finally we can have

$$
\begin{aligned}
E[Y_i] &= E_{P_i}(E[Y_i|P_i]) \\
&= E[m_i P_i] \\
&= m_i \pi_i, \\
Var(Y_i) &= E_{P_i}(Var[Y_i|P_i]) + Var_{P_i}(E[Y_i|P_i]) \\
&= E(m_i P_i(1-P_i)) + Var(m_i P_i) \\
&= m_i E(P_i) - m_i(Var(P_i) + (E(P_i)^2)) + {m_i}^2 Var(P_i) \\
&= m_i \pi_i (1-\pi_i)[1 + \frac{1}{\alpha+\beta+1}(m_i - 1)].
\end{aligned}
$$

# Logistic-normal and related models

- **Random effect** in the linear predictor

$$\eta_i = \beta' x_i + \sigma z_i$$

where $z_i \sim N(0,1)$ is the latent random variable, and

$$Y_i | P_i \sim Bin(m_i, P_i)$$

with $logit(P_i) = \eta_i$.

- Writing that

$$P_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}.$$

# Logistic-normal and related models

- Using Taylor series for $\eta_i$ on $x_i^T\beta$, we have

$$P_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} + \frac{e^{x_i^T\beta}}{(1+e^{x_i^T\beta})^2}(\eta_i - x_i^T\beta) + o(\eta_i - x_i^T\beta).$$

- Then

$$E(P_i) \approx \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} := \pi_i,$$

$$Var(P_i) \approx [\frac{e^{x_i^T\beta}}{(1+e^{x_i^T\beta})^2}]^2 Var(\sigma z_i) = \sigma^2\pi_i^2(1-\pi_i)^2.$$

# Logistic-normal and related models

- Finally we can get

$$\begin{aligned} E[Y_i] &= E_{P_i}(E[Y_i|P_i]) \\ &= E[m_i P_i] \\ &\approx m_i \pi_i , \end{aligned}$$

$$\begin{aligned} Var(Y_i) &= E_{P_i}(Var[Y_i|P_i]) + Var_{P_i}(E[Y_i|P_i]) \\ &= E(m_i P_i(1-P_i)) + Var(m_i P_i) \\ &= m_i E(P_i) - m_i(Var(P_i) + (E(P_i)^2)) + {m_i}^2 Var(P_i) \\ &\approx m_i \pi_i (1-\pi_i)[1+\sigma^2(m_i-1)\pi_i(1-\pi_i)]. \end{aligned}$$

# Overdispersion for Poisson Model

- Poisson regression analysis is commonly used to model **count data**.
- In the case of count data, a Poisson mixture model like the **negative binomial distribution** can be proposed instead, in which the **mean** of the Poisson distribution can itself be thought of as a **random variable** drawn – in this case – from the gamma distribution thereby introducing an additional free parameter (note the resulting negative binomial distribution is completely characterized by two parameters).

# Model for Overdispersion-Poisson

- Negative Binomial Type Variance

$$Y_i|\theta_i \sim Pois(\theta_i),\ \ \theta_i \sim \Gamma(k, \lambda_i)$$

leads to negative binomial distribution with

$$E(Y_i) = \mu_i = k/\lambda_i$$

and

$$\begin{aligned} Var(Y_i) &= E_{\theta_i}[Var(Y_i|\theta_i)] + Var_{\theta_i}[E(Y_i|\theta_i)] \\ &= E(\theta_i) + Var(\theta_i) \\ &= \frac{k}{\lambda_i} + \frac{k}{\lambda_i^2} \\ &= \mu_i + \frac{\mu_i^2}{k} \\ &= \mu_i(1 + \frac{\mu_i}{k}). \end{aligned}$$

# Poisson-normal and related models

- Individual level **random effect** in the linear predictor

$$\eta_i = \beta' x_i + \sigma Z_i.$$

- Assume $Z_i \sim N(0, 1)$ so

$$Y_i | Z_i \sim Pois(\lambda_i)$$

with $log(\lambda_i) = x_i^T \beta + \sigma Z_i$.

# Poisson-normal and related models

- Which gives that

$$
\begin{aligned}
E[Y_i] = E_{Z_i}(E[Y_i|Z_i]) &= E_{Z_i}[e^{x_i^T\beta+\sigma Z_i}] \\
&= e^{x_i^T\beta+\frac{1}{2}\sigma^2} := \mu_i
\end{aligned}
$$

$$
\begin{aligned}
Var(Y_i) &= E_{Z_i}(Var[Y_i|Z_i]) + Var_{Z_i}(E[Y_i|Z_i]) \\
&= e^{x_i^T\beta+\frac{1}{2}\sigma^2} + Var_{Z_i}(e^{x_i^T\beta+\sigma Z_i}) \\
&= e^{x_i^T\beta+\frac{1}{2}\sigma^2} + e^{2x_i^T\beta+\sigma^2}(e^{\sigma^2}-1)
\end{aligned}
$$

i.e. a variance function of the form

$$Var(Y_i) = \mu_i + k'\mu_i^2 = \mu_i(1+k'\mu_i).$$

# Overdispersion

- For a Poisson model, the variance function is $\mu$. To account for overdispersion, we will include another factor $\alpha$ called the "scale parameter" so that

  $$V = \alpha\mu$$

  $\alpha > 1$ is called "overdispersion" and $\alpha < 1$ is called "underdispersion."

# Table of Contents

1. Generalized Linear Model
   - Introduction
   - Model Specification
   - Estimation
   - Practical Aspects
2. Overdispersion
   - Overdispersion for Binomial Model
   - Overdispersion for Poisson Model
3. Quasi-likelihood
   - Iteratively Reweighted Least Squares Algorithm
   - Conclusion

# Why we use quasi-likelihood

- The most popular method for adjusting for overdispersion comes from the theory of quasilikelihood.
- Instead of specifying a probability distribution for the data, only a **relationship between the mean and the variance** is specified in the form of a variance function giving the variance as a function of the mean.
- Generally, this function is allowed to include a multiplicative factor known as the **overdispersion parameter** or **scale parameter** that is estimated from the data. This means we still assume that

$$E(Y) = \mu$$

$$Var(Y) = a(\phi)V(\mu)$$

# Definition of the Quasi-likelihood Function

- Suppose we have independent observations $z_i, \; i = 1, .., n$, with expectations $\mu_i$ and variances $V(\mu_i)$, where $V$ is some known function. We canl relax this specification and say $var(z_i) \propto V(\mu_i)$.
- We suppose that for each observation's $\mu_i$ is some known functions of a set of parameters $\beta_1, ..., \beta_r$.
- Then for each observation we define the quasi-likelihood function $K(z_i, \mu_i)$ by the relation [3]

$$\frac{\partial K(z_i, \mu_i)}{\partial \mu_i} = \frac{z_i - \mu_i}{V(\mu_i)}.$$

# Properties of Quasi-likelihood

- $E(\frac{\partial K}{\partial \mu}) = E(\frac{z-\mu}{V(\mu)}) = 0.$
- $E(\frac{\partial K}{\partial \beta_j}) = E(\frac{\partial K}{\partial \mu}\frac{\partial \mu}{\partial \beta_j}) = 0.$
- $E(\frac{\partial K}{\partial \mu})^2 = Var(\frac{\partial K}{\partial \mu}) + (E(\frac{\partial K}{\partial \mu}))^2 = Var(\frac{\partial K}{\partial \mu}) = \frac{1}{V(\mu)},$
  $E(\frac{\partial^2 K}{\partial \mu^2}) = E(\frac{-V(\mu)+(z-\mu)V'(\mu)}{V^2(\mu)}) = -\frac{1}{V(\mu)},$
  $\Rightarrow E(\frac{\partial K}{\partial \mu})^2 = -E(\frac{\partial^2 K}{\partial \mu^2}) = \frac{1}{V(\mu)}.$
- $E(\frac{\partial K}{\partial \beta_j}\frac{\partial K}{\partial \beta_k}) = E(\frac{\partial K}{\partial \mu})^2\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k} = \frac{1}{V(\mu)}\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k},$
  $-E(\frac{\partial^2 K}{\partial \beta_j \partial \beta_k}) = -E(\frac{\partial}{\partial \beta_k}\{\frac{z-\mu}{V(\mu)}\frac{\partial \mu}{\partial \beta_j}\}) =$
  $-E[(z-\mu)\frac{\partial}{\partial \beta_k}\{\frac{1}{V(\mu)}\frac{\partial \mu}{\partial \beta_j}\} - \frac{1}{V(\mu)}\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k}] = \frac{1}{V(\mu)}\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k},$
  $\Rightarrow E(\frac{\partial K}{\partial \beta_j}\frac{\partial K}{\partial \beta_k}) = -E(\frac{\partial^2 K}{\partial \beta_j \partial \beta_k}) = \frac{1}{V(\mu)}\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k}.$

# Estimation using quasi-likelihoods

- **Estimation using likelihoods**
  For the observation $z$, the log likelihood function is $L$. Then denote $S(\cdot)$ as summation over the observations, maximum likelihood estimates are to solve
  $$\frac{\partial S(L)}{\partial \beta_j} = 0$$
- **Estimation using quasi-likelihoods**
  Like maximum likelihood estimates, maximum quasi-likelihood estimate are to solve
  $$\frac{\partial S(K)}{\partial \beta_j} = 0$$

# Estimation using quasi-likelihoods

- Let $u$ be the vector whose components are $\frac{\partial S(K)}{\partial \beta_j}$. With the properties of quasi-likelihood, we can get

$$E(u) = 0, D := Var(u) = \left(-E(\frac{\partial^2 S(K)}{\partial \beta_j \partial \beta_k})\right)_{1 \leq j,k \leq r}.$$

- Approximating the heissen by its expectation $-D$, hence $\hat{\beta}$ by the Newton-Raphson method is given iteratively as

$$\hat{\beta} \leftarrow \hat{\beta} + D^{-1} u.$$

# Iteratively re-weighted least square

- Let $h_j = \frac{\partial \mu}{\partial \beta_j}$ and $r_i = z_i - \mu_i$. $h_j, r$ and $V(\mu) \in \mathbb{R}^n$. We have

$$u_j = \frac{\partial S(K)}{\partial \beta_j} = S\left(\frac{rh_j}{V(\mu)}\right).$$

- For $D$, we have

$$D_{jk} = -E\frac{\partial^2 S(K)}{\partial \beta_j \partial \beta_k} = S\left(\frac{h_j h_k}{V(\mu)}\right).$$

- Therefore,

$$\begin{aligned}\hat{\beta} + D^{-1}S(u) &= \hat{\beta} + \left(H^T W(\mu) H\right)^{-1} H^T W(\mu) r \\ &= \left(H^T W(\mu) H\right)^{-1} H^T W(\mu)(H\hat{\beta} + r)\end{aligned}$$

where $W(\mu) = \text{diag}(V(\mu_1), \cdots, V(\mu_n))^{-1}$, $H = (h_1, ..., h_r)$.

# Conclusion

In conclusion, there are some comparisons among linear model, generalized linear model and quasi-likelihood model:

| Linear Model | Generalized Linear Model | Quasi-Model |
|---|---|---|
| iid | iid | iid |
| $y \sim N(\mu, \sigma^2)$ | $y \sim$ Exponential family | $y \sim f,\ \ V(\mu), a(\phi)$ |
| $\mu = x^T\beta$ | $g(\mu) = x^T\beta$ | $g(\mu) = x^T\beta$ |

It may be difficult to decide what distribution one's observations follow, but the form of the mean-variance relationship is often much easier to postulate; this is what makes quasi-likelihoods useful.

# Reference

[1] Jianqing Fan et al. *Statistical foundations of data science*. Chapman and Hall/CRC, 2020.

[2] John Hinde and Clarice GB Demétrio. “Overdispersion: models and estimation”. In: *Computational statistics & data analysis* 27.2 (1998), pp. 151–170.

[3] Robert WM Wedderburn. “Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method”. In: *Biometrika* 61.3 (1974), pp. 439–447.