Generalized linear models (GLMs)

**Linear regression**
The model is $Y_i \sim N(\mu_i, \sigma^2)$ where $E(Y_i) = \mu_i = X_i \beta$.

**Logistic regression**
The model is $Y_i \sim Bin(1, p_i)$ where $E(Y_i) = \mu_i = p_i = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}}$.
We thus have $\ln \frac{\mu_i}{1 - \mu_i} = X_i \beta$.

- In both cases either $E(Y_i)$ or a function of $E(Y_i)$ is a linear model.
- Can we extend this to other models?
Generalized linear models (GLMs)

In a generalized linear model we have:

- $Y_i$ all independently distributed from the same member of the exponential family (see later).
- $E(Y_i) = \mu_i$;
- $\eta_i = X_i \beta$ is the “linear predictor”.
- $g(\mu_i) = X_i \beta = \eta_i$ where $g(\cdot)$ is some monotonous and differentiable function called link function.

Examples:

- Linear regression: $g(\mu_i) = \eta_i = \mu_i$, $g(x) = x$ is the identity function.
- Logistic regression: $g(\mu_i) = \ln \frac{\mu_i}{1-\mu_i}$, $g(x) = \ln \frac{x}{1-x}$ is the logit transformation.
The Exponential Family

The ”exponential family” (EF) is a large class of probability distributions (both continuous and discrete). There are several definitions, here follows a specific one:

One-parameter (or ”Natural”) EF

- Let $Y_i$ be a continuous (discrete) random variable with density function (probability mass function) $f(\cdot)$ as given below.
- (one-parameter) EF:

$$f(Y_i; \theta) = h(Y_i) \cdot \exp(\eta_i(\theta) \cdot T(Y_i) - A(\theta))$$

where $h(y)$, $\eta_i(\theta)$, $T(y)$ and $A(\theta)$ are known functions.
The Gaussian distribution is a member of the exponential family. So are exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, Binomial, Poisson, Negative binomial, Wishart, Inverse Wishart and many others.

**Example: Poisson distribution**

If \( Y_i \sim Po(\mu_i) \) we can write the probability mass function as

\[
f(Y_i; \beta) = e^{-\mu_i} \frac{\mu_i^{Y_i}}{Y_i!} = \frac{1}{Y_i!} \cdot \exp(\ln \mu_i \cdot Y_i - \mu_i) \cdot h(Y_i)
\]

where \( \eta_i = \ln \mu_i = x_i \beta \) and thus \( \mu_i = e^{x_i \beta} \) is a natural parametrization.
Advantage of working with GLMs: the generality of methods. For whatever distribution from the EF:

- we can write the likelihood function as \( \prod_i f(Y_i; \beta) \) (for independent \( Y_i \)).
- maximize the (log)likelihood by Newton-Raphson.
- invoke asymptotic normality of maximum likelihood estimates for inferences.
- use likelihood-ratios and deviances for testing.
- Because of the generality of GLMs+EF we do not need to reintroduce specific calculations for each possible distribution for our responses.
As mentioned in Lecture 8, everything is based on the construction of the likelihood function and consequent inferences.

\[ \hat{\beta} \approx N(\beta, H^{-1}), \quad (n \to \infty) \]

where \( H \) is the Hessian matrix of \(-\ln L(\beta)\) evaluated at \( \hat{\beta} \).

We are going to look in detail into 2 additional members (Gaussian and Binomial have been considered already):

- Poisson distribution \( \rightarrow \) Poisson regression;
- Negative Binomial distribution \( \rightarrow \) Negative Binomial regression;
- Poisson regression: just need to change the family argument into \( \text{glm()} \) accordingly.
Poisson regression

We want to investigate the relationship between a variable taking non-negative integer values and some covariates.

This type of response often represents a count (though not exclusively).

Examples:

- Response: The number of people in line at a certain time in the grocery store. Predictors: the number of items currently offered at a special discounted price and whether a special event (e.g., a holiday, a big sporting event) is incoming.

- Response: The number of awards earned by students at a high school. Predictors: the type of program in which the students were enrolled (e.g., vocational, general or academic) and the score on their final exam in math.

- Response: the number customers calling some technical support by phone during an hour. Predictors: time of the day; day of the week (Monday morning should be especially busy).
Poisson regression or loglinear models?

As previously mentioned, $Y$ does not have to represent a count. However this is most often the case of interest in practical applications.

Convention:

- All covariates categorical: responses can be grouped in a (contingency) table with counts in the cells. In literature, these type of models are called **loglinear models**.

- Numerical/continuous covariates: in literature convention is to call them **Poisson regression** models.

For the rest of this lecture we use the term ”Poisson regression” for all cases.
### Poisson regression

We observe $Y_i = \text{"number of events in experiment } i\text{"} \sim Po(\mu_i)$ with $E(Y_i) = V(Y_i) = \mu_i$. Since $\mu_i$ must be positive a suitable function could be $\mu_i = e^{X_i \beta}$ and the link $\ln \mu_i = X_i \beta$ (log-link). Since the probabilities are given by

$$P(Y_i = y_i) = \frac{e^{-\mu_i} \cdot \mu_i^{y_i}}{y_i!} = \frac{e^{-e^{X_i \beta}} (e^{X_i \beta})^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

![Poisson distributions](image-url)
As previously remarked: with GLMs we need to derive the likelihood function.

\[
L(\beta) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \frac{e^{-e^{x_i \beta}} (e^{x_i \beta})^{y_i}}{y_i!},
\]

\[
\ln L(\beta) = \sum_{i=1}^{n} (-e^{x_i \beta}) + \sum_{i=1}^{n} y_i x_i \beta - \sum_{i=1}^{n} \ln(y_i!)
\]

\[
\frac{\partial \ln L(\beta)}{\partial \beta_0} = - \sum_{i=1}^{n} e^{x_i \beta} + \sum_{i=1}^{n} y_i = 0
\]

\[
\frac{\partial \ln L(\beta)}{\partial \beta_1} = - \sum_{i=1}^{n} x_{1i} \cdot e^{x_i \beta} + \sum_{i=1}^{n} x_{1i} y_i = 0
\]

\[
\vdots
\]

The non-linear system \(X'M = X'Y\) is solved by Newton-Raphson.
The rate ratio $e^{\beta_j}$ is the relative increase in the expected value when $X_j$ is increased by 1 (and all other predictors are kept fixed): $\frac{\mu(X_j+1)}{\mu(X_j)} = e^{\beta_j}$, $j = 1, \ldots, p$.

Model comparison: as we know, for GLM models we can test hypotheses about several $\beta_j$ (i.e. larger vs smaller models) using likelihood ratio (deviance) tests.

Use $\hat{\beta}_j \approx N(\beta_j, V(\hat{\beta}_j))$ and $\ln \hat{\mu}_i = X_i \hat{\beta} \approx N(X_i \beta, V(X_i \hat{\beta}))$ for large $n$ to construct intervals for $\beta_j$ and $\ln \mu_i$.

Here, and for all GLMs, $V(\hat{\beta}_j)$ is obtained from the diagonal elements of the inverted Hessian matrix at convergence of Newton-Raphson.

A confidence interval for $\mu_i$ is then given by $I_{\mu_i} = e^{I X_i \hat{\beta}}$. 

Anna Lindgren - anna@maths.lth.se

Linear and Logistic Regression, L10 12/19
Example: number of awards

Response: the number of awards earned by students at a high school. $\bar{Y} = 0.63 < s^2 = 1.11$. Larger variance explained by covariates?

Predictors of the number of awards earned include the type of program in which the student was enrolled (e.g., vocational, general or academic) and the score on their final exam in math. Model: $Y_i \sim Po(\mu_i)$ where

$$\ln \mu_i = \beta_0 + \beta_1 X_{\text{Academic},i} + \beta_2 X_{\text{Vocational},i} + \beta_3 X_{\text{math},i} = X_i \beta.$$  

Estimated model:

$$\hat{\mu}_i = \begin{cases} 
  e^{-5.25 + 0.07 \cdot X_{\text{math},i}} = 0.005 \cdot 1.07^{X_i}, & \text{General} \\
  e^{-5.25 + 1.08 + 0.07 \cdot X_{\text{math},i}} = 0.016 \cdot 1.07^{X_i}, & \text{Academic} \\
  e^{-5.25 + 0.37 + 0.07 \cdot X_{\text{math},i}} = 0.008 \cdot 1.07^{X_i}, & \text{Vocational}
\end{cases}$$

The expected number of awards increases by 7% for each extra math score and Academic students get $e^{1.08} = 2.96 \approx 3$ times more awards than General, for the same math score.
A few outliers. However no observation has a strong influence on the estimates. Note that the residuals have constant variance, even though the original observations have increasing variance.
Negative binomial regression

Count data often vary more than the Poisson distribution allows. That is the fact that for Poisson variables the mean equals the variability is a rather strong assumptions that in many situations does not hold\(^1\).

- Let each individual have their own poisson mean, \( z_i \cdot \mu_i \), randomly distributed around some common value, \( \mu_i = e^{X_i\beta} \).

- Then let \( Y_i, \, i = 1, \ldots, n \) be independent observations with

\[
(Y_i \mid Z_i = z_i) \sim \text{Po}(z_i \mu_i)
\]

and

\[
Z_i \sim \Gamma(\theta, 1/\theta) \quad \text{with} \quad E(Z_i) = 1 \text{ and } V(Z_i) = 1/\theta.
\]

\(^1\)See the nice example at http://stats.idre.ucla.edu/r/dae/negative-binomial-regression/
Now $Y_i, i = 1, \ldots, n$ are independent observations from the “Negative Binomial” distribution below

$$P(Y_i = y_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(y + 1)\Gamma(\theta)} \cdot \frac{(\mu_i/\theta)^{y_i}}{(1 + \mu_i/\theta)^{\theta+y_i}}$$

$$= \frac{\Gamma(\theta + y_i)}{\Gamma(y + 1)\Gamma(\theta)} \cdot \frac{(e^{x_i\beta}/\theta)^{y_i}}{(1 + e^{x_i\beta}/\theta)^{\theta+y_i}}, \quad y_i = 0, 1, 2 \ldots$$

$$E(Y_i) = \mu_i$$

$$V(Y_i) = \mu_i + \mu_i^2/\theta > \mu_i \quad (\theta > 0)$$

$$(\Gamma(x) = (x - 1)! \text{ if } x \text{ is an integer, else } \Gamma(x) = \int_0^\infty t^{x-1}e^{-t} dt.)$$
Variances: $5 + 5^2/0.05 = 505$; $5 + 5^2/0.5 = 55$; $5 + 5^2/5 = 10$; $5 + 5^2/50 = 5.5$:

Estimate $\beta$ and $\theta$ by Maximum Likelihood. Solved by Newton-Raphson.

R: negative-binomial regression is not implemented in the basic R library. Need to load the MASS package via library(MASS)
notice again, maximization of the likelihood will also return a \( \hat{\theta} \), see the bottom of the summary(model).

Test hypotheses about several \( \beta_j \) using likelihood ratio (deviance) tests.

Use \( \hat{\beta}_j \approx N(\beta_j, V(\hat{\beta}_j)) \) and \( \ln \hat{\mu}_i = X_i\hat{\beta} \approx N(X_i\beta, V(X_i\hat{\beta})) \) for large \( n \) to construct intervals for \( \beta_j \) and \( \ln \mu_i \).

A confidence interval for \( \mu_i \) is then given by \( I_{\mu_i} = e^{lx_i\beta} \).

Test if it is worth to use a negative-binomial instead of the (simpler) Poisson using a likelihood ratio test:

\[-2 \ln L_{\text{poisson}} - (-2 \ln L_{\text{negbin}}) > \chi^2_{1-\alpha}(1).\]
Warning

A little remark: in order to execute a likelihood ratio test to compare a Poisson vs Negative-Binomial model do NOT use `anova()`.

`anova()` is built to compare (nested) models where observations follow the same distribution.

Compare “by hand” using

```
-2*(logLik(model.pois)[1]-logLik(model.nb)[1])
```

For a nice additional example see http://stats.idre.ucla.edu/r/dae/negative-binomial-regression/