Generalized linear models (GLMs)

**Linear regression**

The model is \( Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi} + \epsilon_i = X_i \beta + \epsilon_i \)

where \( \epsilon_i \sim N(0, \sigma^2) \). This gives \( Y_i \sim N(\mu_i, \sigma^2) \) where

\[ E(Y_i) = \mu_i = X_i \beta. \]

**Logistic regression**

The model is \( Y_i \sim Bin(1, p_i) \) where

\[ \ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi} = X_i \beta. \]

We thus have

\[ E(Y_i) = \mu_i = p_i = \frac{e^{X_i \beta}}{1+e^{X_i \beta}} \quad \text{and} \quad \ln \frac{\mu_i}{1-\mu_i} = X_i \beta. \]
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: $g(\mu_i) = \eta_i = \mu_i$, $g(\cdot)$ is the identity function.
- Logistic regression: $g(\mu_i) = \ln \frac{\mu_i}{1-\mu_i}$, $g(\cdot)$ is the logit transformation.
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:

\[
 f(Y_i; \eta_i, \phi, w_i) = h(\phi, Y_i, w_i) \exp\left( \frac{w_i}{\phi}(\eta_i Y_i - r(\eta_i)) \right)
\]

with scalars \( \eta_i, \phi, w_i \)
At the board we proved that the Gaussian distribution is a member of the exponential family (EF).

Other members: exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, Poisson, Wishart, Inverse Wishart and many others.

Advantage of working with GLMs: the generality of methods. For whatever distribution from the EF:

- we can write the likelihood function as \( \prod_i f(Y_i; \cdot) \) (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 all specific calculations for each possible distribution for our responses.

As mentioned before, 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 \( \to \) Poisson regression;
- Negative Binomial distribution \( \to \) Negative Binomial regression;
- Poisson regression: just need to change the \texttt{family} argument into \texttt{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 one 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 of minutes you have to wait when calling some technical support by phone. Predictors: time of the day; day of the week (Monday morning should be especially busy).
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 $Y_i = "\text{number of events in experiment } i" \sim Po(\mu_i)$ 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

$$P(Y_i = y_i) = \frac{e^{-\mu_i} \cdot \mu_i^{y_i}}{y_i!} = \frac{e^{-e^{X_i \beta}} (e^{X_i \beta})^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$
As previously remarked: with GLMs we need to derive the likelihood function.

\[ L(\beta_0, \ldots, \beta_q) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \frac{e^{-e^{X_i \beta}} (e^{X_i \beta})^{y_i}}{y_i!}, \]

\[ \ln L(\beta_0, \ldots, \beta_q) = \sum_{i=1}^{n} (-e^{X_i \beta}) + \sum_{i=1}^{n} y_i X_i \beta - \sum_{i=1}^{n} \ln(y_i!) \]

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

\[ \frac{\partial \ln L(\beta_0, \ldots, \beta_q)}{\partial \beta_1} = - \sum_{i=1}^{n} X_{1i} \cdot e^{X_i \beta} + \sum_{i=1}^{n} X_{1i} y_i = 0 \]

\[ \vdots \]

Solved by Newton-Raphson.
- $e^{\beta_j}$ is the relative increase in the expected value when $X_j$ is increased by 1 (and all other predictors are kept fixed):
\[
\frac{\mu(X_j + 1)}{\mu(X_j)} = e^{\beta_j}, \quad j = 1, \ldots, q.
\]

- 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 $I_{\mu_i} = e^{I_{X_i \hat{\beta}}}$. 
Example

\[ Y_i \sim Po(\mu_i) \text{ where } \ln \mu_i = \beta_0 + \beta_1 X_i = X_i \beta. \]

Estimates:

<table>
<thead>
<tr>
<th>\hat{\beta}_j</th>
<th>S.E.(\hat{\beta}_j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>\beta_0</td>
<td>0.998</td>
</tr>
<tr>
<td>\beta_1</td>
<td>0.010</td>
</tr>
</tbody>
</table>

Estimated line:

\[ \hat{\mu}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot X_i} = e^{0.998 + 0.010 \cdot X_i} = e^{0.998} \cdot (e^{0.010})^{X_i} = 2.7 \cdot 1.01^{X_i}. \]

Confidence interval for line:

\[ I_{\mu_i} = e^{I_{\beta_0 + \beta_1 X_i}} = e^{(\hat{\beta}_0 + \hat{\beta}_1 X_i \pm \lambda_{\alpha/2} \cdot S.E.(\hat{\beta}_0 + \hat{\beta}_1 X_i))} \]

\[ S.E.(\hat{\beta}_0 + \hat{\beta}_1 X_i) = \sqrt{S.E.^2(\hat{\beta}_0) + X_i^2 \cdot S.E.^2(\hat{\beta}_1) + 2X_i \cdot Cov(\hat{\beta}_0, \hat{\beta}_1)} = \]

\[ = \sqrt{0.105^2 + X_i^2 \cdot 0.0016^2 + 2X_i \cdot (-0.00015)} \]
A few outliers. However no observation has a strong influence on the estimates.

(a) Standardised residuals $r_i$

(b) Cook’s distance
Response: the number of awards earned by students at one high school.

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.

See poissregr-awards.R.
summary(m1 <- glm(num_awards ~ prog + math, family="poisson"))

Coefficients:                   Estimate Std. Error z value Pr(>|z|)
(Intercept)           -5.24712    0.65845  -7.969 1.60e-15 ***
progAcademic          1.08386    0.35825   3.025  0.00248 **
progVocational        0.36981    0.44107   0.838   0.40179
math                  0.07015    0.01060   6.619 3.63e-11 ***

Therefore, for a fixed math grade, students from the academic program get on average $e^{1.08} = 2.96$ more awards than those from the general program (baseline).
You can experiment and read more on this example at http://stats.idre.ucla.edu/r/dae/poisson-regression/
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 \( Y_i, \ i = 1, \ldots, n \) be independent observations from the “Negative Binomial” distribution below. Again, choose \( \mu_i = e^{\beta_0 + \beta_1 X_{1i} + \ldots + \beta_q X_{qi}} \), which gives

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

\[
= \frac{\Gamma(\theta + y_i)}{\Gamma(y + 1)\Gamma(\theta)} \cdot \frac{(e^{\mathbf{x}_i\beta}/\theta)^y_i}{(1 + e^{\mathbf{x}_i\beta}/\theta)^{\theta+y_i}}, \quad y_i = 0, 1, 2 \ldots
\]

\[
E(Y_i) = \mu_i
\]

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

\[
(\Gamma(x) = (x − 1)! \text{ if } x \text{ is an integer, else } \Gamma(x) = \int_0^\infty t^{x−1}e^{−t} \, dt.)
\]

\(^1\)See the nice example at

http://stats.idre.ucla.edu/r/dae/negative-binomial-regression/
Variance: \( 5 + 5^2/0.05 = 505; \ 5 + 5^2/0.5 = 55; \ 5 + 5^2/5 = 10; \ 5 + 5^2/50 = 5.5 \):

![Negative binomial distributions (μ=5)](image)

Estimate \( \beta_0, \ldots, \beta_p \) 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 $\ln \hat{\mu}_i = \mathbf{X}_i \hat{\beta} \approx N(\mathbf{X}_i \beta, V(\mathbf{X}_i \hat{\beta}))$ for large $n$ to construct intervals for $\beta_j$ and $\ln \mu_i$.
- A confidence interval for $\mu_i$ is then given by $I_{\mu_i} = e^{I_x \beta}$.
- Test if it is worth to use a negative-binomial instead of the (simpler) Poisson using a likelihood ratio test: $-2 \ln L_{\text{poisson}} - (-2 \ln L_{\text{negbin}}) > \chi^2_{1-\alpha}(1)$. 

![Poisson regression](image1.png) ![Negative binomial regression](image2.png)
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

\[-2\cdot (\text{logLik(model.pois)[1]} - \text{logLik(model.nb)[1]})\]
The negative binomial distribution is suitable as a robust alternative to the Poisson. It approaches the Poisson for large $\theta$.

Proof of the convergence: take the probability mass function of a NB random variable $Y$:

$$P(Y = y) = \frac{\Gamma(\theta + y)}{\Gamma(y + 1)\Gamma(\theta)} \cdot \frac{(\mu/\theta)^y}{(1 + \mu/\theta)^{\theta+y}}, \quad y = 0, 1, 2 \ldots$$

$$= (\text{use } \Gamma(y) = (y - 1)! \text{ for integer } y)$$

$$= \frac{\mu^y}{y!} \cdot \frac{\Gamma(\theta + y)}{\theta^y\Gamma(\theta)} \cdot \frac{1}{(1 + \mu/\theta)^{\theta+y}}$$

$$= \frac{\mu^y}{y!} \cdot \frac{\Gamma(\theta + y)}{\theta^y\Gamma(\theta)} \cdot \frac{1}{(1 + \mu/\theta)^{\theta}(1 + \mu/\theta)^y}$$
Now as $\theta \to \infty$ we have the (famous) limit result $(1 + \frac{\mu}{\theta})^\theta \to e^\mu$

Also it can be seen that $\frac{\Gamma(\theta+y)}{\theta y \Gamma(\theta)} \to 1$.

In conclusion as $\theta \to \infty$ we have

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, ...$$

which is the probability mass function of $Po(\mu)$. 

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