MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: . . . with Data Gathering, 9 hp

Lecture 10, spring 2023
Generalized linear models

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

8/5-23

Generalized linear models (GLMs)

Linear regression

The model: $Y_i \sim N(\mu_i,\, \sigma^2)$ where $E(Y_i) = \mu_i = \mathbf{x}_i\boldsymbol{\beta}$.

Logistic regression

The model: $Y_i \sim Bin(1,\, p_i)$ where $E(Y_i) = \mu_i = p_i = \dfrac{\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}$.

We thus have $\text{logit}(\mu_i) = \ln \dfrac{\mu_i}{1-\mu_i} = \mathbf{x}_i\boldsymbol{\beta}$.

- In both cases either $E(Y_i)$ or a (monotonous) 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 next slide).
- $E(Y_i) = \mu_i$
- $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$ is the **linear predictor**.
- $g(\mu_i) = \mathbf{x}_i\boldsymbol{\beta} = \eta_i$ where $g(\cdot)$ is some monotonous and differentiable function called a **link function**.

Examples:

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

(*) 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") Exponential Family

- Let Y_i be a continuous (discrete) random variable with density function (probability mass function) $p(\cdot;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter (vector) in the distribution.
- The distribution belongs to the (one-parameter) Exponential Family if $p()$ can be written as

$$p(Y_i;\boldsymbol{\theta}) = h(Y_i)\cdot\exp\left(\eta_i(\boldsymbol{\theta})\cdot T(Y_i) - A(\boldsymbol{\theta})\right)$$

  where $h(Y_i)$, $\eta_i(\boldsymbol{\theta})$, $T(Y_i)$ and $A(\boldsymbol{\theta})$ are known functions
- and $h()$ and $T()$ do not contain $\boldsymbol{\theta}$
- while $\eta_i()$ and $A()$ do not contain Y_i.

GLM, cont'd

- The Normal 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

$$p(Y_i;\boldsymbol{\beta}) = \mathrm{e}^{-\mu_i}\,\frac{\mu_i^{Y_i}}{Y_i!} = \underbrace{\frac{1}{Y_i!}}_{h(Y_i)} \cdot \exp(\underbrace{\ln\mu_i}_{\eta_i(\boldsymbol{\beta})} \cdot \underbrace{Y_i}_{T(Y_i)} - \underbrace{\mu_i}_{A(\boldsymbol{\beta})})$$

where $\eta_i = \ln\mu_i = \mathbf{x}_i\boldsymbol{\beta}$ and thus $\mu_i = \mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}$ is a natural parametrization.

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

- we can write the likelihood function as $\prod_{i=1}^{n} p(Y_i; \boldsymbol{\beta})$ (for independent Y_i).
- maximize the (log)likelihood by Newton-Raphson.
- invoke asymptotic normality of maximum likelihood estimates for confidence intervals and tests..
- 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.

Remember that the Deviance is the distance from the saturated model, $\hat{\boldsymbol{\mu}}_{\mathsf{sat}} = \mathbf{Y}$:

$$D = 2\left(\ln L(\hat{\boldsymbol{\mu}}_{\mathsf{sat}}) - \ln L(\hat{\boldsymbol{\beta}})\right)$$

In logistic regression we had $\ln L(\hat{\boldsymbol{\mu}}_{\mathsf{sat}}) = 0$. This is not generally the case.

- Everything is based on the construction of the likelihood function and consequent inferences.

We have

$$\hat{\boldsymbol{\beta}} \approx N_{p+1}(\boldsymbol{\beta}, \mathbf{H}^{-1}), \qquad (n \to \infty)$$

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

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

- Poisson distribution $\rightarrow$ Poisson regression;
- Negative Binomial distribution $\rightarrow$ Negative Binomial regression;
- Poisson regression: just change the `family` argument in `glm()` from `"binomial"` to `"poisson"`.

Poisson regression

We want to investigate the relationship between a variable, Y, taking non-negative integer values, and some covariates, $x_1, \dots, x_p$.

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

Examples:

- 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.
- 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.
- The number of 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 in practical applications.

Convention:

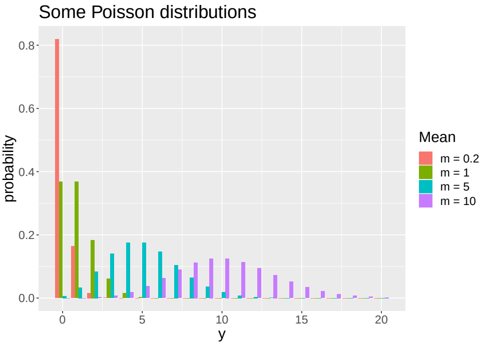
- All covariates categorical: responses can be grouped in a (contingency) table with counts in the cells. In literature, these types of models are called **loglinear models**.
- Numerical/continuous covariates: in literature convention is to call them **Poisson regression** models.

We use the term "Poisson regression" for all cases.

Poisson regression

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

$$Pr(Y_i = y_i) = \frac{\mathrm{e}^{-\mu_i}\cdot\mu_i^{y_i}}{y_i!} = \frac{\mathrm{e}^{-\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}(\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}})^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$



With GLMs we need to derive the likelihood function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} Pr(Y_i = y_i) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}}(\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}})^{Y_i}}{Y_i!},$$

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}(-\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}) + \sum_{i=1}^{n} Y_i\mathbf{x}_i\boldsymbol{\beta} - \sum_{i=1}^{n}\ln(Y_i!)$$

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} = -\sum_{i=1}^{n} x_{ij}\cdot \mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}} + \sum_{i=1}^{n} x_{ij}Y_i = 0 \quad \text{for } j = 1, \ldots, p$$

where $x_{i0} = 1$ for $i = 1, \ldots, n$. The non-linear system $\mathbf{X}'\mathbf{M} = \mathbf{X}'\mathbf{Y}$ is solved by Newton-Raphson giving $\hat{\boldsymbol{\beta}}$ with

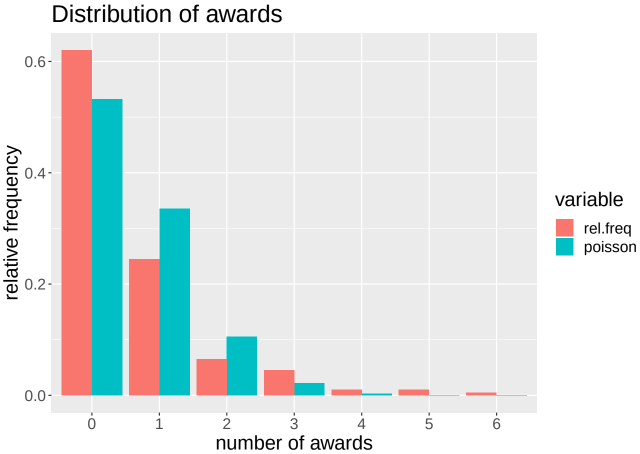
$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta},\, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}) \qquad \text{asymptotically}$$

where $\mathbf{W}$ is a diagonal matrix with elements $w_{ii} = \mathrm{e}^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}$, i.e., an estimate of $\mathrm{Var}(\mathbf{Y})$.

- The rate ratio (RR) e^{β_j} is the relative increase in the expected value when x_j is increased by 1 (and all other predictors are kept fixed): $\dfrac{\mu(x_j+1)}{\mu(x_j)} = \mathrm{e}^{\beta_j}$, $\qquad j = 1, ..., p$.
- Model comparison: as we know, for GLM models we can test hypotheses about several β_j (i.e. larger vs smaller models) using likelihood ratio (deviance) tests.
- Use profile likelihood to construct confidence intervals for β_j.
- Use $\ln \hat{\mu}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}} \approx N(\mathbf{x}_0\boldsymbol{\beta}, \mathbf{x}_0\mathrm{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0')$ for large n to construct confidence intervals for $\ln \mu_0 = \mathbf{x}_0\boldsymbol{\beta}$.
- 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 μ_0 is then given by $I_{\mu_0} = \mathrm{e}^{I_{\mathbf{x}_0\boldsymbol{\beta}}}$.
- Prediction intervals for $\hat{Y}_{\mathsf{pred}_0}$ requires, e.g., bootstrap.

Example: number of awards

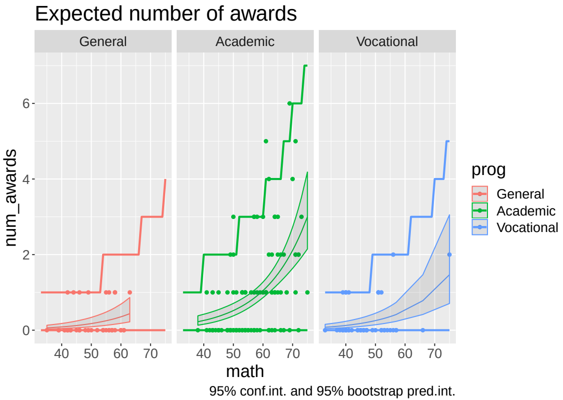
Response: the number of awards earned by a student at a high school. $\bar{Y} = 0.63 < s^2 = 1.11$. Larger variance explained by covariates?



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
$\ln \mu_i = \beta_0 + \beta_1 x_{i,\text{Aca}} + \beta_2 x_{i,\text{Voc}} + \beta_3 x_{i,\text{math}} = \mathbf{x}_i \boldsymbol{\beta}$.
Estimated model:

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

The expected number of awards increases by 7 % for each extra math score and Academic students get RR $= \mathrm{e}^{1.08} = 2.96 \approx 3$ times more awards than General, for the same math score.

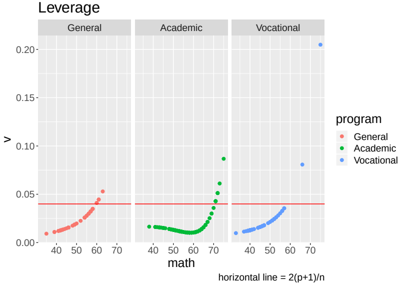


Parameter	β	est.	P-value	95 % C.I.	RR (e^{β})	95 % C.I.
(Intercept)	β_0	-5.2	< 0.001	$(-6.6\ , -4.0)$	0.005	(0.001, 0.02)
Academic	β_1	1.08	0.002	$(0.44\ ,\ 1.86)$	2.96	$(1.54\ ,\ 6.4)$
Vocational	β_2	0.37	0.40	$(-0.49, 1.27)$	1.45	$(0.61\ , 3.55)$
math	β_3	0.07	< 0.001	$(0.05\ ,\ 0.09)$	1.07	$(1.05\ , 1.10)$

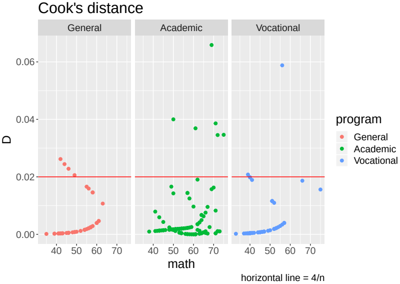
LR-test against null model $D_0 - D = 98.2 > \chi^2_{0.05}(3) = 7.81$.

LR-test against model with only math: $D_{\text{red}} - D_{\text{full}} = 14.6 > \chi^2_{0.05}(2) = 5.99$

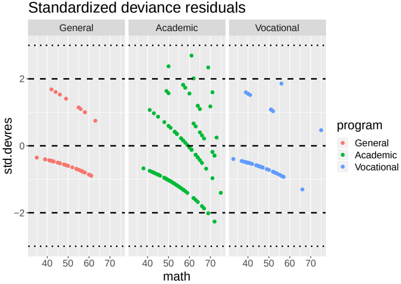
Residuals and influential observations


Leverage

horizontal line = 2(p+1)/n


Cook's distance

horizontal line = 4/n


Standardized deviance residuals

- Few and small outliers.
- Some observations have a large influence on the estimates.
- The residuals have constant variance, even though the original observations have increasing variance.

Offset variables in Poisson

- When we are counting the number of events during different lengths of time, on areas of different sizes, or in populations of different sizes, the expected value of Y_i will be proportional to the length of the interval, the area, or the population size.
- In these situations it is often convenient to express the expected value as events per hour, or per m^2, or per 1000 inhabitants.
- However, we can not model, Y_i/area_i since this has no suitable distribution. Also, the variance becomes larger for smaller areas than for larger areas!
- Instead, use the area as an **offset** variable. This is an x-variable where the β-parameter is known to be 1:

$$\mu_i = \text{area}_i \cdot \mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}} \qquad \Leftrightarrow \qquad \ln \mu_i = \ln \text{area}_i + \mathbf{x}_i\boldsymbol{\beta}$$

```
glm(y ~ x1 + x2, offset(log(area)), family = "poisson", data = my.df)
```

Pseudo R^2 using the deviance

We can define a pseudo R^2 for Poisson regression comparing the deviance $D = 2(\ln L(\hat{\boldsymbol{\mu}}) - \ln L(\hat{\boldsymbol{\beta}}))$ with the null deviance $D_0 = 2(\ln L(\hat{\boldsymbol{\mu}}) - \ln L(\hat{\beta}_0))$. We can also adjust it with the number of parameters:

$$R^2_{\mathsf{D}} = 1 - \frac{D}{D_0} = 1 - \frac{2(\ln L(\hat{\boldsymbol{\mu}}) - \ln L(\hat{\boldsymbol{\beta}}))}{2(\ln L(\hat{\boldsymbol{\mu}}) - \ln L(\hat{\beta}_0))} = 1 - \frac{\ln L(\hat{\boldsymbol{\beta}}) - \ln L(\hat{\boldsymbol{\mu}})}{\ln L(\hat{\beta}_0) - \ln L(\hat{\boldsymbol{\mu}})}$$

$$R^2_{\mathsf{D,adj}} = 1 - \frac{D + p}{D_0} = 1 - \frac{\ln L(\hat{\boldsymbol{\beta}}) - \ln L(\hat{\boldsymbol{\mu}}) - p/2}{\ln L(\hat{\beta}_0) - \ln L(\hat{\boldsymbol{\mu}})}$$

Note: McFadden R^2_{McF} for logistic regression was a special case of this where $\ln L(\hat{\boldsymbol{\mu}}) = 0$.

AIC and BIC

As before: $AIC = 2(p + 1) + D$ and $BIC = \ln n \cdot (p + 1) + D$

Negative binomial regression

Count data often vary more than the Poisson distribution allows. That the mean equals the variance is a rather strong assumption that in many situations does not hold[1].

- Let each observation have their own poisson mean, $z_i \cdot \mu_i$, randomly distributed around some common value determined by $\mathbf{x}_i$: $\mu_i = \mathrm{e}^{\mathbf{x}_i \boldsymbol{\beta}}$.
- Then let Y_i, $i = 1, \ldots, n$ be independent observations with

  $$(Y_i \mid Z_i = z_i) \sim Po(z_i \mu_i)$$

  and

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

  We then get the distribution of Y_i as (Law of Total Probability)

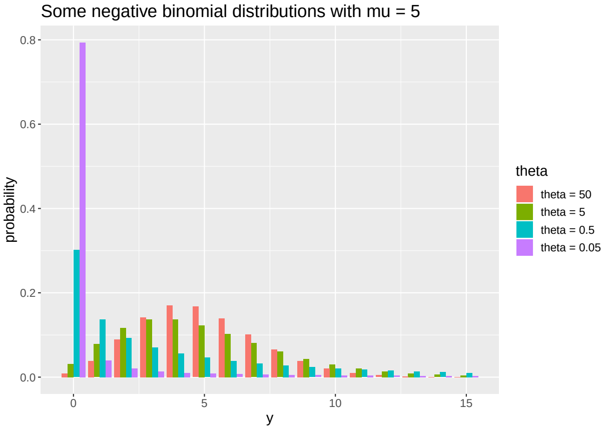
  $$Pr(Y_i = y_i) = \int_0^\infty Pr(Y_i = y_i \mid Z_i = z) \cdot f_{Z_i}(z)\, dz$$

[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 with

$$\begin{aligned} Pr(Y_i = y_i) &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} \cdot \frac{(\mu_i/\theta)^{y_i}}{(1 + \mu_i/\theta)^{\theta + y_i}} \\ &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} \cdot \frac{(\mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}/\theta)^{y_i}}{(1 + \mathrm{e}^{\mathbf{x}_i\boldsymbol{\beta}}/\theta)^{\theta + y_i}}, \quad y_i = 0, 1, 2 \ldots \\ E(Y_i) &= \mu_i \\ V(Y_i) &= \mu_i + \mu_i^2/\theta > \mu_i \qquad (\theta > 0) \end{aligned}$$

($\Gamma(x) = (x-1)!$ if x is an integer, else $\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}\,dt$.)



Variances:

$5 + 5^2/50 = 5.5$

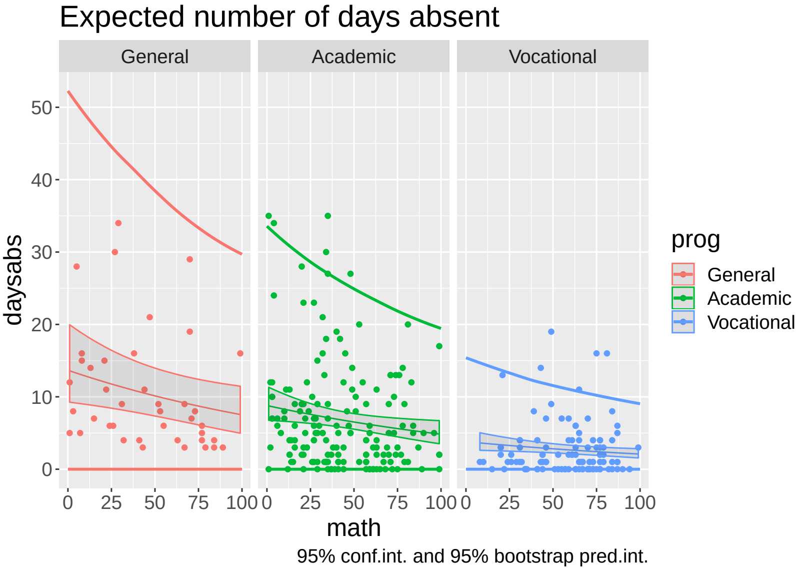
$5 + 5^2/5 = 10$

$5 + 5^2/0.5 = 55$

$5+5^2/0.05 = 505$

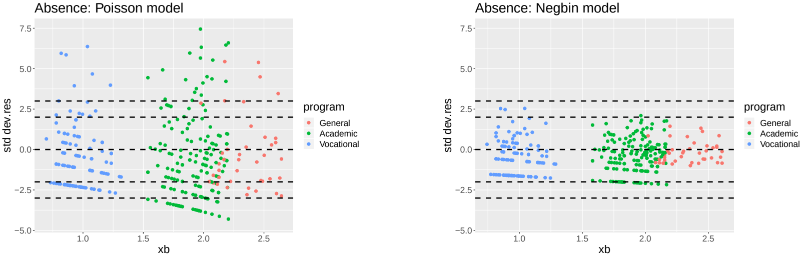
Estimate β and θ by Maximum Likelihood. Solved by Newton-Raphson.

- R: negative-binomial regression is not implemented in the basic R library. Need to load the `MASS` package via `library(MASS)`
- Maximization of the likelihood will also return a $\hat{\theta}$, see the bottom of the `summary(model)`.
- Test hypotheses about several β_j using likelihood ratio (deviance) tests.
- Use profile likelihood to construct confidence intervals for β_j.
- Use $\ln\hat{\mu}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}} \approx N(\mathbf{x}_0\boldsymbol{\beta},\, \mathbf{x}_0\mathrm{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0')$ for large n to construct confidence intervals for $\ln\mu_0 = \mathbf{x}_0\boldsymbol{\beta}$.
- A confidence interval for μ_0 is then given by $I_{\mu_0} = \mathrm{e}^{I_{\mathbf{x}_0\boldsymbol{\beta}}}$.
- Prediction intervals with bootstrap.
- Offset variables in the same way as for Poisson.
- AIC, BIC and pseudo R^2_{D} and $R^2_{\mathsf{D,adj}}$ can be used for ranking models.


Expected number of days absent
General
Academic
Vocational
daysabs
0
10
20
30
40
50
0
25
50
75
100
math
prog
General
Academic
Vocational
95% conf.int. and 95% bootstrap pred.int.

Negative binomial or Poisson?

- Compare the standardized deviance residuals for the two models. If the residuals for the Poisson model are larger than for the Negative binomial model, use Negative binomial.



- Compare the models using a likelihood ratio test. Reject H$_0$: $1/\theta = 0$ if $-2\ln L_{\text{poisson}} - (-2\ln L_{\text{negbin}}) > \chi^2_\alpha(1)$. Since $2657.3 - 1731.3 = 926.0 > \chi^2_{0.05}(1) = 3.84$ (P-value $= 2 \cdot 10^{-203} \ll 0.05$) we should reject the Poisson model and use the Negative binomial.