CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

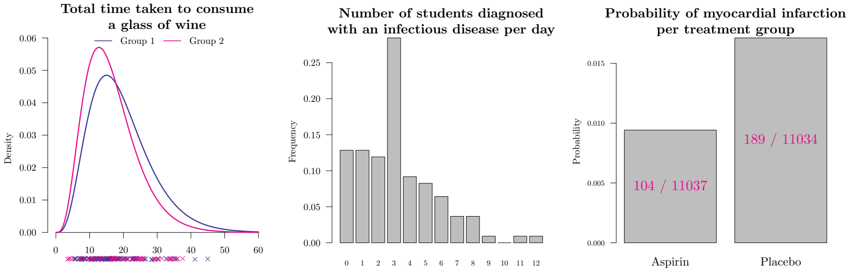
MRC | Laboratory of Molecular Biology

Generalised Linear Models (GLM)

Cancer Research UK – 21^{st} of November 2022

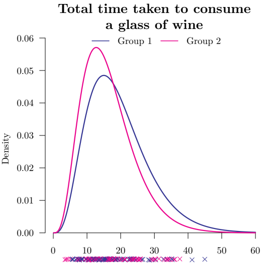
D.-L. Couturier / R. Nicholls / M. Fernandes

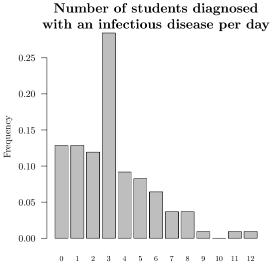
Examples of data with non-normal conditional distributions

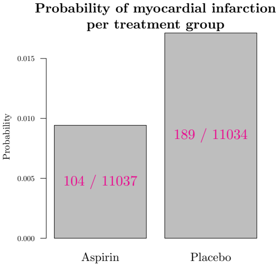


UNIVERSITY OF CAMBRIDGE

Examples of data with non-normal conditional distributions







Linear model not suitable:

- Assumed model:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2),$$

$$Y_i | (\mathbf{x}_i, \boldsymbol{\beta}) \sim N(\mu_i, \sigma^2).$$

  - theoretical range of $\epsilon_i = [-\infty, +\infty]$,
  - $\mathbf{x}_i^T \boldsymbol{\beta}$ not bounded to $[0, \infty]$ or $[0, 1]$,
  - $\text{Var}[Y_i]$ independent of $\text{E}[Y_i]$.

- Solution:

$$Y_i | (\mathbf{x}_i, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{x}_i^T \boldsymbol{\beta}), \phi),$$

where *distribution* belongs to the exponential family and *function* is monotonically increasing.

GLM: conditional distributions

$$Y_i|(\mathbf{x}_i, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{x}_i^T\boldsymbol{\beta}), \phi),$$

- Some possible conditional *distributions* :
  statistical probability mass functions & density functions
  - Within the exponential family ['classical' GLM framework]

    normal
    exponential
    gamma
    chi-squared
    beta
    Dirichlet
    Bernoulli
    Poisson
    Wishart
    Inverse Wishart
    ...

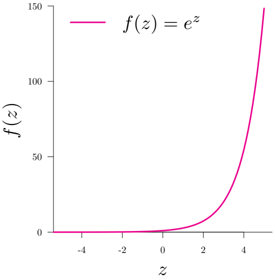
  - Outside the exponential family ['extended' GLM framework]

    Box-Cox power
    exponential
    exponential Gaussian
    generalized beta
    generalized gamma
    generalized inverse
    Gaussian
    inverse Gaussian
    logistic
    power exponential
    reverse Gumbel
    skew power exponential
    Weibull
    Pareto type I, II, III
    Poisson inverse Gaussian
    ...

GLM: link functions

$$Y_i|(\mathbf{x}_i, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{x}_i^T \boldsymbol{\beta}), \phi),$$

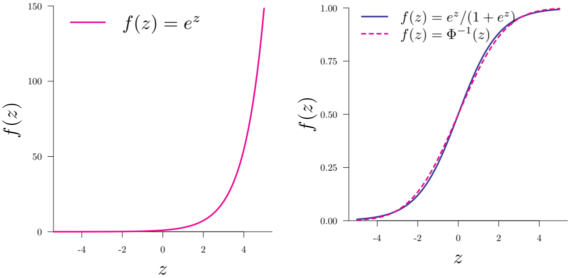
- Most used link *functions* :
  connection between Y_i and $\mathbf{x}_i^T \boldsymbol{\beta}$
  - to restrict $f(\mathbf{x}_i^T \boldsymbol{\beta})$ to belong to $[0, \infty[$:
    - ▷ log link: $f(z) = e^z$



GLM: link functions

$$Y_i|(\mathbf{x}_i, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{x}_i^T\boldsymbol{\beta}), \phi),$$

- Most used link *functions* :
  connection between Y_i and $\mathbf{x}_i^T\boldsymbol{\beta}$
  - to restrict $f(\mathbf{x}_i^T\boldsymbol{\beta})$ to belong to $[0, \infty[$:
    - ▷ log link: $f(z) = e^z$
  - to restrict $f(\mathbf{x}_i^T\boldsymbol{\beta})$ to belong to $[0, 1]$:
    - ▷ logit link: $f(z) = e^z/(1 + e^z) = 1/(1 + e^{-z})$ where z is positive
    - ▷ probit link: $f(z) = \Phi(z)$, where Φ denotes the $N(0, 1)$.



Distribution for dichotomous variates: Bernoulli

Example:
in *Jones (Unpublished BSc dissertation, University of Southampton, 1975)*, the main outcome is the presence/absence of bronchitis:

Sample of 212 men in Cardiff:	$i = 1$	$i = 2$	$i = 3$	$\cdots$	$i = 212$
	$\cancel{\text{B}}$	$\cancel{\text{B}}$	B	$\cdots$	$\cancel{\text{B}}$
y_i	0	0	1	$\cdots$	0

If

- n independent experiments,
- outcome of each experiment is dichotomous (success/failure),
- the probability of success π is the same for all experiments,

then, each dichotomous experiment, Y_i, follows a Bernoulli distribution with parameter π:

$$Y_i \sim Bernoulli(\pi)$$

$$P(Y_i = 1) = \pi$$

$$P(Y_i = 0) = 1 - \pi$$

Logistic regression: GLM for dichotomous variates

Example:
in *Jones (Unpublished BSc dissertation, University of Southampton, 1975)*, the main outcome is the presence/absence of bronchitis as a function of the daily number of smoked cigarettes (X_1) and level of pollution (X_2):

Sample of 212 men in Cardiff:	$i = 1$	$i = 2$	$i = 3$	$\cdots$	$i = 212$
	~~B~~	~~B~~	B	$\cdots$	~~B~~
y_i	0	0	1	$\cdots$	0
x_{1i}	5.15	0	2.5		0.9
x_{2i}	67.1	66.9	66.7		55.4

If

- n independent experiments,
- outcome of each experiment is dichotomous (success/failure),
- the probability of success π is the same for all experiments given the covariates,

then, each dichotomous experiment, Y_i, follows a Bernoulli distribution with parameter π_i:

$$Y_i \sim Bernoulli(\pi_i) \text{ where } \pi_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$$

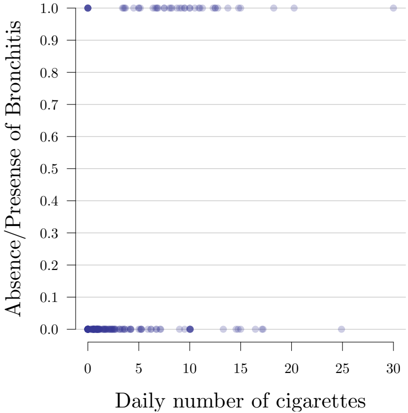
$$P(Y_i = 1) = \pi_i$$

$$P(Y_i = 0) = 1 - \pi_i$$

Logistic regression: predictions and interpretation of β

Example:
Model the probability of presence of bronchitis as a function of the daily number of smoked cigarettes (X_1) :

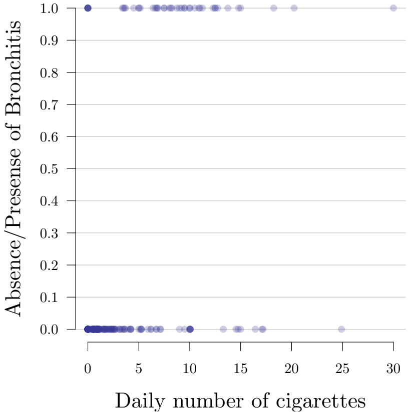
$$P(Y_i = 1) = \pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1}}}{1 + e^{\beta_0 + \beta_1 x_{i1}}}$$



Logistic regression: predictions and interpretation of β

Example:
Model the probability of presence of bronchitis as a function of the daily number of smoked cigarettes (X_1) :

$$P(Y_i = 1) = \pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1}}}{1 + e^{\beta_0 + \beta_1 x_{i1}}}$$

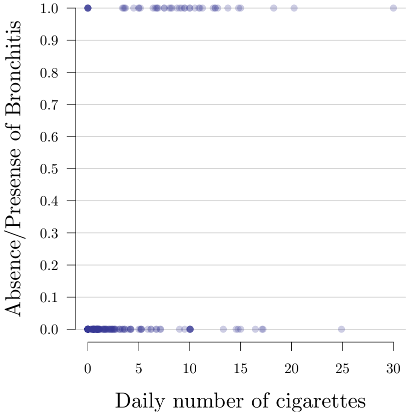


	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.2839832	0.27305505	-8.364552	6.034375e-17
cigs	0.2093618	0.03760466	5.567442	2.585062e-08

Logistic regression: predictions and interpretation of β

Example:
Model the probability of presence of bronchitis as a function of the daily number of smoked cigarettes (X_1) :

$$P(Y_i = 1) = \pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1}}}{1 + e^{\beta_0 + \beta_1 x_{i1}}}$$



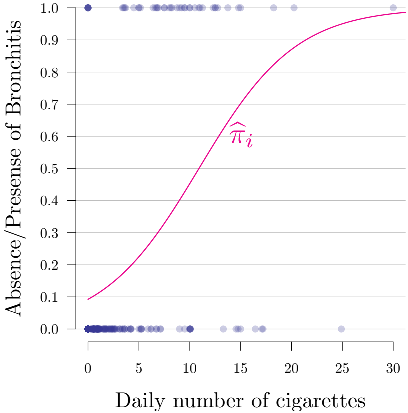
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.2839832	0.27305505	-8.364552	6.034375e-17
cigs	0.2093618	0.03760466	5.567442	2.585062e-08

- $\widehat{\pi}_{\text{0cig}} = \frac{e^{\widehat{\beta}_0}}{1+e^{\widehat{\beta}_0}}$

Logistic regression: predictions and interpretation of β

Example:
Model the probability of presence of bronchitis as a function of the daily number of smoked cigarettes (X_1) :

$$P(Y_i = 1) = \pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1}}}{1 + e^{\beta_0 + \beta_1 x_{i1}}}$$



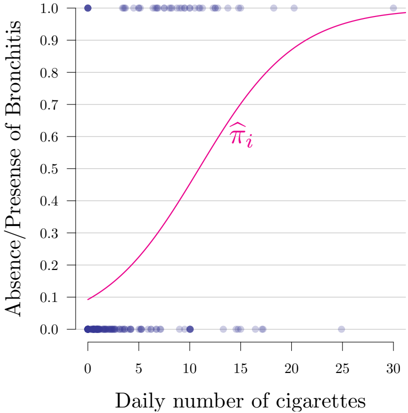
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.2839832	0.27305505	-8.364552	6.034375e-17
cigs	0.2093618	0.03760466	5.567442	2.585062e-08

- $\widehat{\pi}_{0\text{cig}} = \frac{e^{\widehat{\beta}_0}}{1+e^{\widehat{\beta}_0}} \Leftrightarrow \frac{\widehat{\pi}_{0\text{cig}}}{1-\widehat{\pi}_{0\text{cig}}} = e^{\widehat{\beta}_0}$
- $\widehat{\pi}_{1\text{cig}} = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1}} \Leftrightarrow \frac{\widehat{\pi}_{1\text{cig}}}{1-\widehat{\pi}_{1\text{cig}}} = e^{\widehat{\beta}_0+\widehat{\beta}_1}$
- $\dfrac{\frac{\widehat{\pi}_{1\text{cig}}}{1-\widehat{\pi}_{1\text{cig}}}}{\frac{\widehat{\pi}_{0\text{cig}}}{1-\widehat{\pi}_{0\text{cig}}}} = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1}}{e^{\widehat{\beta}_0}} = e^{\widehat{\beta}_1}$ (Odd ratio)

Logistic regression: predictions and interpretation of β

Example:
Model the probability of presence of bronchitis as a function of the daily number of smoked cigarettes (X_1) :

$$P(Y_i = 1) = \pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1}}}{1 + e^{\beta_0 + \beta_1 x_{i1}}}$$

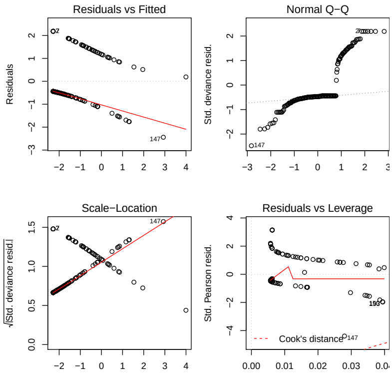


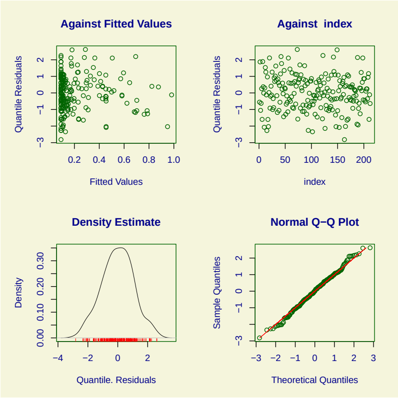

```
            Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -2.2839832 0.27305505 -8.364552 6.034375e-17
cigs         0.2093618 0.03760466  5.567442 2.585062e-08
```

- $\widehat{\pi}_{\text{0cig}} = \frac{e^{\widehat{\beta}_0}}{1+e^{\widehat{\beta}_0}} \Leftrightarrow \frac{\widehat{\pi}_{\text{0cig}}}{1-\widehat{\pi}_{\text{0cig}}} = e^{\widehat{\beta}_0}$
- $\widehat{\pi}_{\text{1cig}} = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1}} \Leftrightarrow \frac{\widehat{\pi}_{\text{1cig}}}{1-\widehat{\pi}_{\text{1cig}}} = e^{\widehat{\beta}_0+\widehat{\beta}_1}$
- $\dfrac{\frac{\widehat{\pi}_{\text{1cig}}}{1-\widehat{\pi}_{\text{1cig}}}}{\frac{\widehat{\pi}_{\text{0cig}}}{1-\widehat{\pi}_{\text{0cig}}}} = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1}}{e^{\widehat{\beta}_0}} = e^{\widehat{\beta}_1}$ (Odd ratio)
- **H0$_1$:** $\beta_0 = 0$ versus **H1$_1$:** $\beta_0 \neq 0$
  H0$_2$: $\beta_1 = 0$ versus **H1$_2$:** $\beta_1 \neq 0$

Logistic regression: model check

- pearson residuals $(y_i - \widehat{\pi})/\sqrt{\mathrm{Var}(\widehat{\pi})}$,
- deviance residuals [Default in **R**],
- randomised normalised quantile residuals [Default in package **gamlss()**]





Distribution for count data: Poisson

Example:
Interest for the number of high school students diagnosed with an infectious disease

Sample of 115 days:	$t = 1$	$t = 2$	$t = 3$	$\cdots$	$t = 115$
y_i	6	8	12	$\cdots$	0

If, during a time interval or in a given area,

- events occur independently,
- at the same rate,
- and the probability of an event to occur in a small interval (area) is proportional to the length of the interval (size of the area),

then,

- a count occurring in a fixed time interval or in a given area, Y, may be modelled by means of a Poisson distribution with parameter μ:

$$Y \sim Poisson(\mu) \text{ where } \mu = \mathsf{E}[Y] = \mathsf{Var}[Y],$$

- the probability of observing x events during a fixed time interval or in a given area is given by

$$P(Y = y|\mu) = \frac{\mu^y e^{-\mu}}{y!}.$$

Poisson regression: GLM for count data

Example:
Interest for the number of high school students diagnosed with an infectious disease as a function of the number of days from the initial outbreak

Sample of 115 days: $t=1$ $t=2$ $t=3$ $\cdots$ $t=115$

y_t	6	8	12	$\cdots$	0
t	1	2	3	$\cdots$	115

If, during a time interval or in a given area,

- events occur independently given the covariates,
- at the same rate given the covariates,
- and the probability of an event to occur in a small interval (area) is proportional to the length of the interval (size of the area) given the covariates,

then,

- each count occurring in a fixed time interval or in a given area, Y_t, may be modelled by means of a Poisson distribution with parameter μ_t:

$$Y_t \sim Poisson(\mu_t) \text{ where } \mu_t = \mathsf{E}[Y] = \mathsf{Var}[Y] = e^{\mathbf{x}_i^T \boldsymbol{\beta}},$$

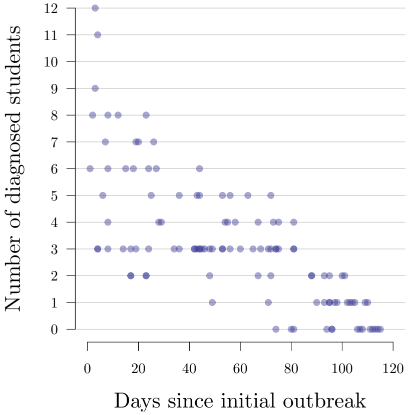
- the probability of observing y during the fixed time interval or in the given area is given by

$$P(Y_t = y_t | \mu_t) = \frac{\mu_t^{y_t} e^{-\mu_t}}{y_t!}.$$

Poisson regression: predictions and interpretation of β

Example:
Model the mean count of diagnosed students, μ_t, as a function of the number of days from the outbreak (T) :

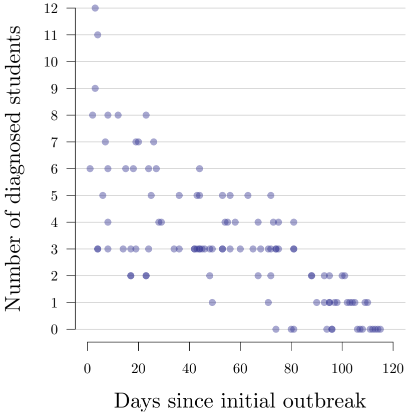
$$\mu_t = e^{\beta_0 + \beta_1 t}$$



Poisson regression: predictions and interpretation of β

Example:
Model the mean count of diagnosed students, μ_t, as a function of the number of days from the outbreak (T) :

$$\mu_t = e^{\beta_0 + \beta_1 t}$$

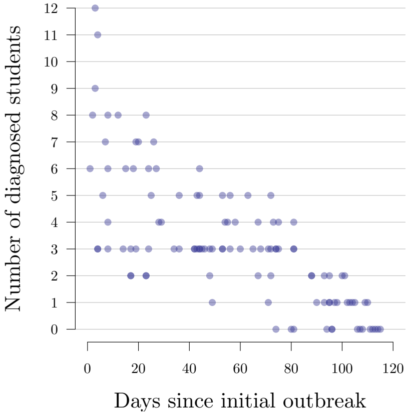



```
              Estimate  Std. Error   z value       Pr(>|z|)
(Intercept)  1.99023497 0.083935207  23.71156 2.739875e-124
day         -0.01746317 0.001726709 -10.11356  4.810392e-24
```

Poisson regression: predictions and interpretation of β

Example:
Model the mean count of diagnosed students, μ_t, as a function of the number of days from the outbreak (T) :

$$\mu_t = e^{\beta_0 + \beta_1 t}$$



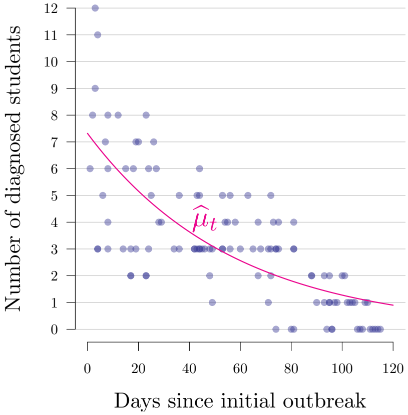
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.99023497	0.083935207	23.71156	2.739875e-124
day	-0.01746317	0.001726709	-10.11356	4.810392e-24

- $\widehat{\mu}_{\text{day0}} = e^{\widehat{\beta}_0}$

Poisson regression: predictions and interpretation of β

Example:
Model the mean count of diagnosed students, μ_t, as a function of the number of days from the outbreak (T) :

$$\mu_t = e^{\beta_0 + \beta_1 t}$$



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.99023497	0.083935207	23.71156	2.739875e-124
day	-0.01746317	0.001726709	-10.11356	4.810392e-24

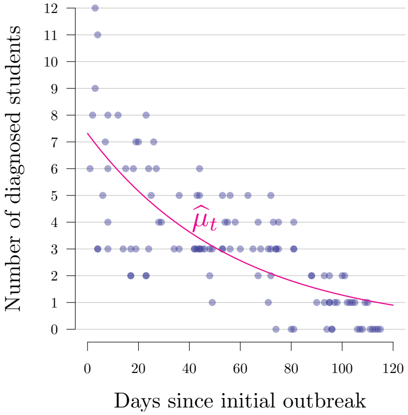
- $\widehat{\mu}_{\text{day0}} = e^{\widehat{\beta}_0}$
- $\widehat{\mu}_{\text{day1}} = e^{\widehat{\beta}_0 + \widehat{\beta}_1}$
- $\dfrac{\widehat{\mu}_{\text{day1}}}{\widehat{\mu}_{\text{day0}}} = \dfrac{e^{\widehat{\beta}_0 + \widehat{\beta}_1}}{e^{\widehat{\beta}_0}} = e^{\widehat{\beta}_1}$

UNIVERSITY OF CAMBRIDGE

Poisson regression: predictions and interpretation of β

Example:
Model the mean count of diagnosed students, μ_t, as a function of the number of days from the outbreak (T) :

$$\mu_t = e^{\beta_0 + \beta_1 t}$$




```
              Estimate  Std. Error   z value      Pr(>|z|)
(Intercept)  1.99023497 0.083935207  23.71156 2.739875e-124
day         -0.01746317 0.001726709 -10.11356  4.810392e-24
```

- $\widehat{\mu}_{\text{day0}} = e^{\widehat{\beta}_0}$
- $\widehat{\mu}_{\text{day1}} = e^{\widehat{\beta}_0 + \widehat{\beta}_1}$
- $\dfrac{\widehat{\mu}_{\text{day1}}}{\widehat{\mu}_{\text{day0}}} = \dfrac{e^{\widehat{\beta}_0 + \widehat{\beta}_1}}{e^{\widehat{\beta}_0}} = e^{\widehat{\beta}_1}$

- **H0**$_1$**:** $\beta_0 = 0$ versus **H1**$_1$**:** $\beta_0 \neq 0$
  H0$_2$**:** $\beta_1 = 0$ versus **H1**$_2$**:** $\beta_1 \neq 0$

Poisson regression: model check

- pearson residuals $(y_i - \widehat{\pi})/\sqrt{\mathsf{Var}(\widehat{\pi})}$,
- deviance residuals [Default in **R**],
- randomised normalised quantile residuals [Default in package **gamlss()**]