Modèle de Poisson et surdispersion

Master 2 Modélisation en pharmacologie clinique et épidémiologie

Jean-François Dupuy

Jean-Francois.Dupuy@insa-rennes.fr

Linear regression model

The linear regression model is used to model a quantitative outcome Y as a function of one or several predictor variables X. Predictors can be quantitative and/or qualitative. E.g. :

- **outcome** : diastolic blood pressure
- **predictors** : systolic blood pressure (num.), cholesterol level (num.), body mass index (num.), age (num. or cat.), coronary artery disease status (cat.)

Remark 1

- when there is a single predictor, the model is called a simple linear model, otherwise it is called a multiple linear model,
- when the single predictor is a categorical variable, the model is called one-way ANOVA.

Linear regression model

The simple linear regression model of Y on X is :

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

with $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

The multiple linear model of Y_i over $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ with $X_{i1} = 1$ is :

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \\ &= \beta^\top \mathbf{X}_i + \varepsilon_i \end{aligned}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top$. The β_j are interpreted as slopes.

Linear regression model

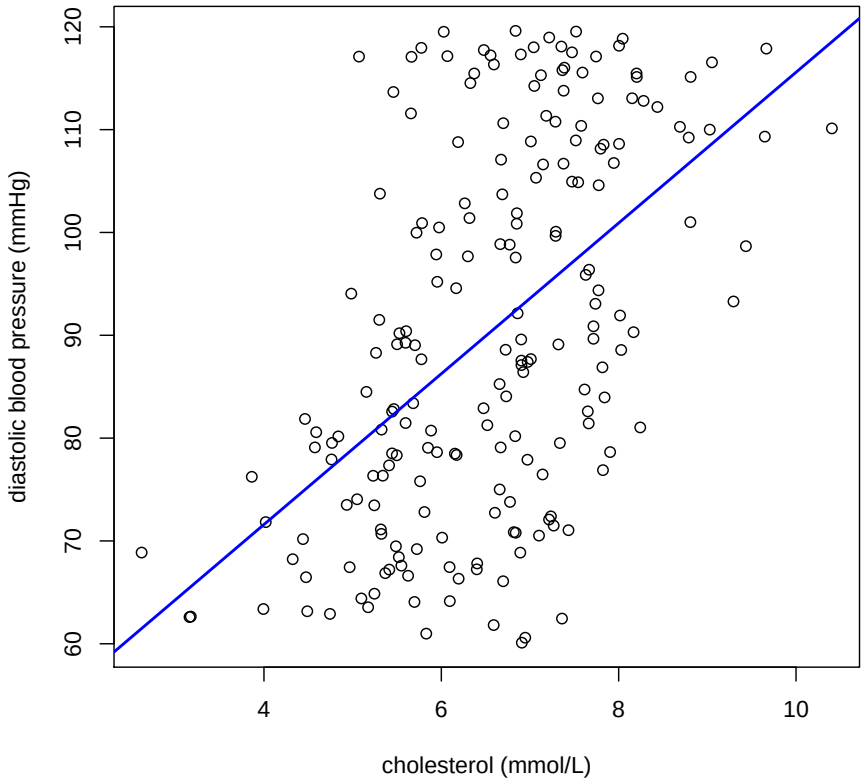

FIGURE 1 – Linear regression of diastolic blood pressure (dbp) vs cholesterol level

Linear regression model

Objectives of linear regression :

- evaluate and interpret the influence of cholesterol on dbp (e.g., "*1 mmol/L increase in cholesterol causes 7 mmHg increase in dbp*")

  $\hookrightarrow$ maximum likelihood (or least squares) estimation, confidence intervals
- test the significance of this effect (Student test, Fisher test)
- predict the dbp value for a given cholesterol level
- in a multiple regression model, select relevant predictors (AIC, stepwise procedure)

Linear regression model

FIGURE 2 – A brief research in PubMed [1]

1. PubMed comprises more than 36 million citations for biomedical literature from MEDLINE, life science journals, and online books

Logistic regression : a model for binary response

Linear regression assumes a normally distributed outcome and is not adapted for modeling a binary response.

FIGURE 3 – A brief research in PubMed (ctd)

Logistic regression

Logistic regression for a binary response $Y_i \in \{0, 1\}$ (e.g., surgical site infection, PTSD [2] symptom remission of sexual assault victims, in-hospital death in patients with cancer...) :

$$Y_i | \mathbf{X}_i \sim \mathcal{B}(\pi(\mathbf{X}_i))$$

with probability of "success" $\pi(\mathbf{X}_i) := \mathbb{P}(Y_i = 1 | \mathbf{X}_i)$ modeled as :

$$\pi(\mathbf{X}_i) = \frac{e^{\beta^\top \mathbf{X}_i}}{1 + e^{\beta^\top \mathbf{X}_i}}$$

or equivalently,

$$\text{logit}\left(\pi(\mathbf{X}_i)\right) := \log\left(\frac{\pi(\mathbf{X}_i)}{1 - \pi(\mathbf{X}_i)}\right) = \beta^\top \mathbf{X}_i.$$

If $\beta_j > 0$, then $\pi(\mathbf{X}_i)$ increases as X_{ij} increases.

2. posttraumatic stress disorder

Logistic regression

Objectives of logistic regression :

- evaluate and interpret the influence of X_{ij} on the "risk" of "success" (odds-ratio)

  $\hookrightarrow$ ML estimation, confidence intervals
- test the significance of this effect (Wald test, LR test)
- predict the probability of success (or the binary outcome Y)
- in a multiple logistic regression model, select relevant predictors (AIC, stepwise procedure)

Poisson regression : a model for count data

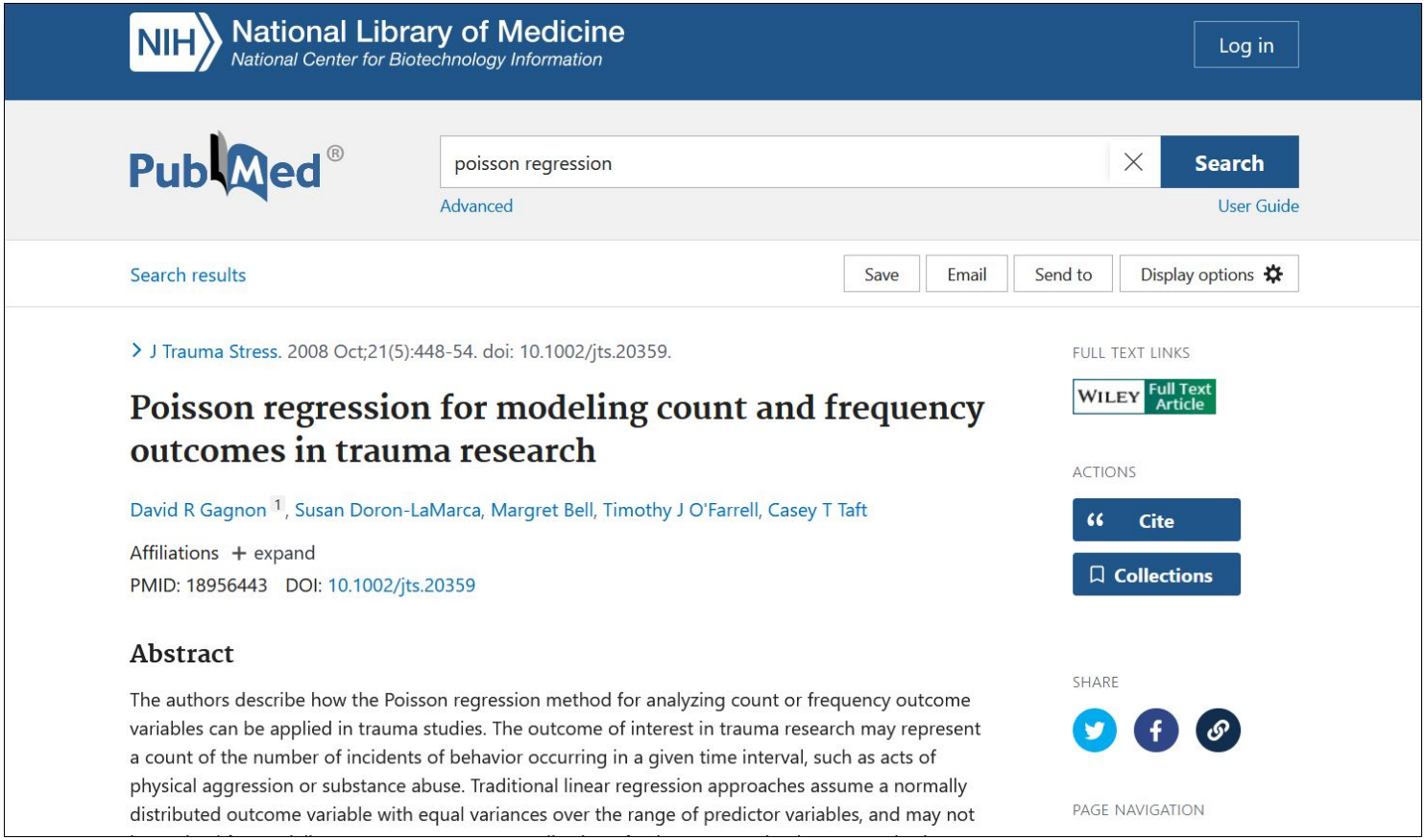

FIGURE 4 – A brief research in PubMed (ctd)

Again, the linear model based on the normal law is no longer adapted.

Poisson regression : a model for count data

FIGURE 5 – A brief research in PubMed (ctd)

Poisson regression : a model for count data

FIGURE 6 – A brief research in PubMed (ctd)

Poisson model : a model for count data

The Poisson law is a discrete law which describes the number of events occurring in a period of time.

Some examples : number of asthma episodes in early childhood, number of doctor visits over a given time period, number of Deaths of Despair (DoD)[3] in England...

We have $Y \sim \mathcal{P}(\lambda)$, with $\lambda > 0$ iff

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

A remarkable (and often unreasonable !) property :

$$\mathbb{E}(Y) = \lambda \quad \text{and} \quad \text{var}(Y) = \mathbb{E}(Y) = \lambda \quad \hookrightarrow \text{equidispersion}$$

3. DoD are socially patterned causes of death encompassing drug and alcohol misuse and suicide, see "Morts de désespoir", A. Case and A. Deaton, 2021.

Poisson model : a model for count data

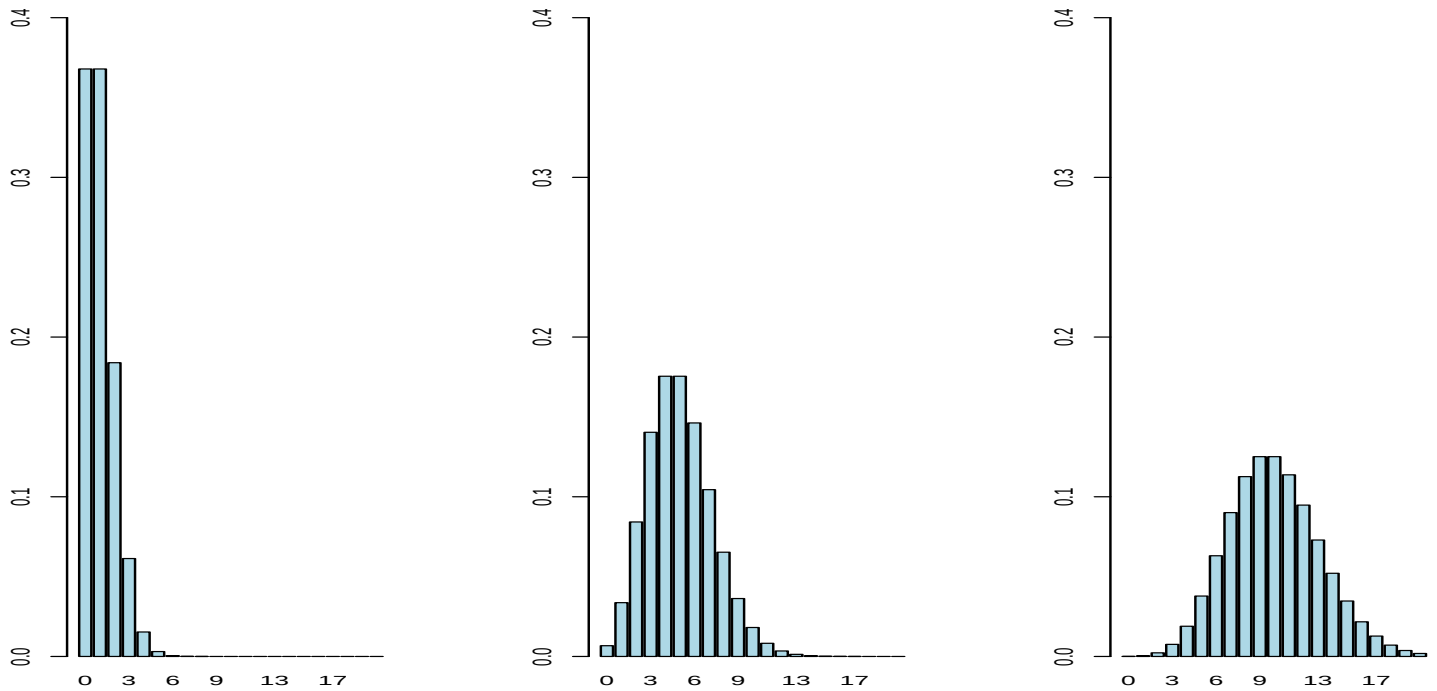

FIGURE 7 – Poisson laws $\mathcal{P}(1), \mathcal{P}(5)$ and $\mathcal{P}(10)$ (left to right). Observe how the mean and variance both increase as λ increases.

Estimation in the Poisson model

Maximum likelihood :

$$\begin{aligned} L_n(\lambda) &= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) \\ &= \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \end{aligned}$$

Take the logarithm, differentiate wrt λ and solve the estimating equation $\Rightarrow$ maximum likelihood estimate (MLE) :

$$\hat{\lambda}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}_n$$

Poisson regression model

Now, suppose that the count response Y may be influenced by some predictors. E.g. the number of doctor visits may be influenced by :

- age (num. or cat.)
- health satisfaction : 0 (low) - 10 (high) (num.)
- years of schooling (num.)
- household income (num.)
- handicap : yes/no (cat.)
- degree of handicap in percentage points (num.)
- health insurance : yes/no (cat.)
- ...

Poisson regression model

Poisson regression model specifies the (conditional) law of Y_i as :

$$Y_i|\mathbf{X}_i \sim \mathcal{P}(\lambda(\mathbf{X}_i)),$$

and so :

$$\mathbb{P}(Y_i = y_i|\mathbf{X}_i) = \frac{e^{-\lambda(\mathbf{X}_i)}\lambda(\mathbf{X}_i)^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots.$$

We have : $\mathbb{E}[Y_i|\mathbf{X}_i] = \lambda(\mathbf{X}_i)$ and $\text{var}(Y_i|\mathbf{X}_i) = \lambda(\mathbf{X}_i)$.

A model for $\lambda(\mathbf{X}_i)$ is :

$$\begin{aligned} \lambda(\mathbf{X}_i) &= e^{\beta^\top \mathbf{X}_i} \\ &= e^{\beta_1+\beta_2 X_{i2}+\ldots+\beta_p X_{ip}} \end{aligned}$$

which is always positive.

Interpreting the coefficients

Note that :

$$\frac{\mathbb{E}[Y_i|X_i = x+1]}{\mathbb{E}[Y_i|X_i = x]} = \frac{e^{\beta_1+\beta_2(x+1)}}{e^{\beta_1+\beta_2 x}} = e^{\beta_2}.$$

- if $e^{\beta_2} = 2$, increasing X by 1 unit multiplies the expected count by 2
- if $e^{\beta_2} = 0.5$, increasing X by 1 unit divides the expected count by 2
- if $e^{\beta_2} = 1$ (i.e. $\beta_2 = 0$), then

$$\mathbb{E}[Y_i|X_i = x+1] = \mathbb{E}[Y_i|X_i = x]$$

and X has no influence of the expected count $\Rightarrow$ it will be of interest to test $\mathcal{H}_0 : \beta_2 = 0$ vs $\mathcal{H}_1 : \beta_2 \neq 0$.

- similar interpretation with more than one regressor

Poisson regression model as a GLM

The Poisson regression model is a special case of generalized linear models [4] (GLM). Other examples of GLM :

- logistic regression
- linear regression
- gamma regression (e.g., intensive care unit length of stay, mechanical ventilation duration in ventilated infants with Bronchiolitis, cost of extra working hours due to AHTO [5])
- inverse Gaussian regression (e.g., quantitative blood loss from uterine atony, hospitalization times of COVID-19 patients)
- ...

4. McCullagh P., Nelder J.A., Generalized Linear Models, Wiley, 1989
5. AHTO ("alcohol's harm to others") includes the adverse effects imposed on health, safety and QoL of other people due to an individual's alcohol consumption

Poisson regression in practice

- in R, Poisson models (or more generally, GLM) are fitted with the function glm :

```
model = glm(outcome ~ regressors, family=poisson
(link="log"))
```

- in SAS, GLM are available in the GENMOD procedure[6] :

```
proc genmod;
model outcome = regressors / dist=poi link=log ;
run;
```

6. **gen**eralized **mod**eling

Poisson regression model : link function

We assumed

$$\lambda(\mathbf{X}_i) = e^{\beta_1+\beta_2 X_{i2}+\ldots+\beta_p X_{ip}}$$

which is equivalent to

$$\ln(\lambda(\mathbf{X}_i)) = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}.$$

We say that $\ln$ is the link function. Another choice is the identity link ($\mathrm{id}(x) = x$) :

$$\lambda(\mathbf{X}_i) = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}.$$

Although less natural, from a statistical viewpoint, the id link is used, e.g., in the context of dosimetry, where it is motivated by the shape of the dose-response curve.

Poisson regression model : link function

The main purpose of biological dosimetry is to translate an observed amount of damage in cells into a radiation dose estimate. This is useful to distinguish those deemed to be critically exposed, who should be prioritised, from people who have comparatively been less exposed and are unlikely to need urgent treatment.

When fitting dose-response curves to calibration data, it is standard to relate the mean yield of aberrations, $\lambda(X_i)$, to the dose X_i via the quadratic model

$$\lambda(X_i) = \beta_1 + \beta_2 X_i + \beta_2 X_i^2.$$

Poisson regression model : link function

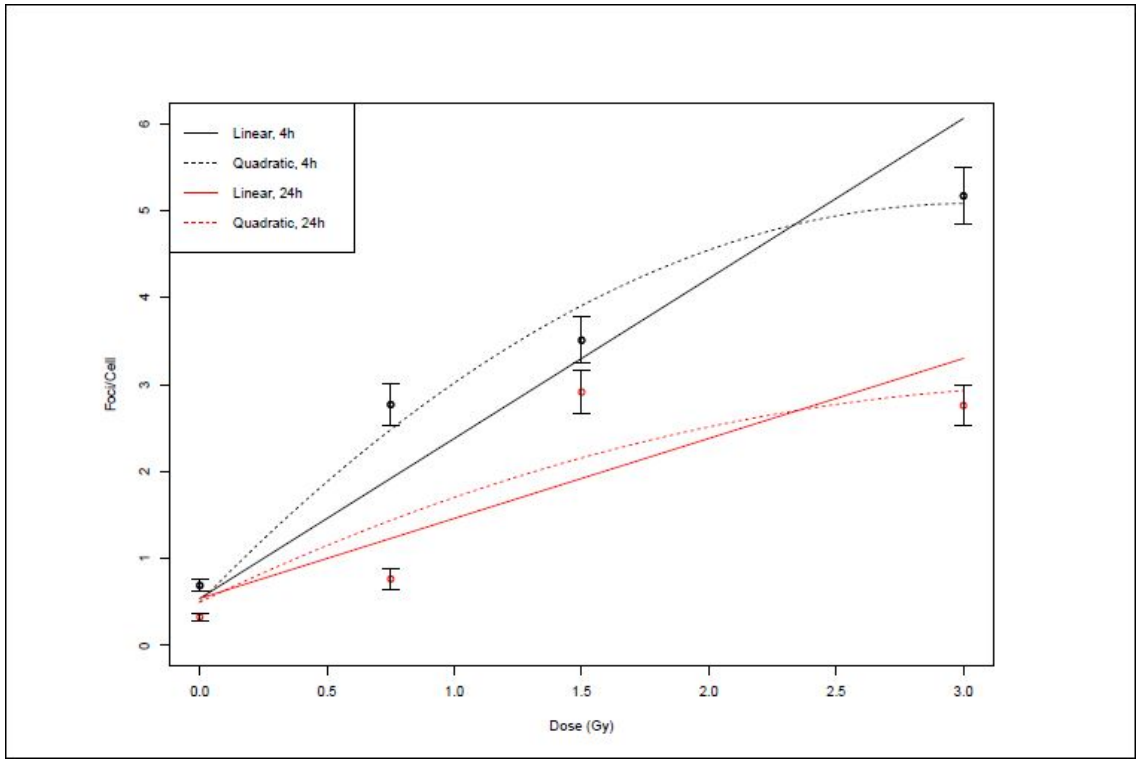

FIGURE 8 – Linear and quadratic calibration curves fitted to 4h (black) and 24h (red) PHE-Foci1 data.

Source : Errington A. *Estimating dose and exposure fraction from radiation biomarkers in the presence of overdispersion*. PhD of Durham Univ., UK, 2023.

Poisson regression model

Remark 2 (offset term)

- $\lambda(\mathbf{X}_i)$ is the expected count in a "unit" time length.
- If t_i is the time length in which events occur (t_i known), then the count distribution is modeled as

$$Y_i|\mathbf{X}_i \sim \mathcal{P}(t_i \cdot \lambda(\mathbf{X}_i))$$

- Assuming that $\lambda(\mathbf{X}_i) = e^{\beta^\top \mathbf{X}_i}$, then $t_i \cdot \lambda(\mathbf{X}_i)$ can be rewritten as

$$\begin{aligned} t_i \cdot \lambda(\mathbf{X}_i) &= e^{\ln t_i} \cdot e^{\beta^\top \mathbf{X}_i} \\ &= e^{\beta^\top \mathbf{X}_i + \mathbf{1} \ln t_i} \end{aligned}$$

The term $\ln t_i$ is called *offset* and can be treated as a covariate with coefficient forced to 1.

Estimation in the Poisson regression model

The parameter $\beta = (\beta_1, \ldots, \beta_p)^\top$ is estimated by MLE :

$$\begin{aligned} L_n(\beta) &= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | \mathbf{X}_i) \\ &= \prod_{i=1}^{n} \frac{e^{-\lambda(\mathbf{X}_i)} \lambda(\mathbf{X}_i)^{Y_i}}{Y_i!} \\ &= \prod_{i=1}^{n} \frac{e^{-e^{\beta^\top \mathbf{X}_i}} e^{Y_i \beta^\top \mathbf{X}_i}}{Y_i!} \end{aligned}$$

Take $\ln$, differentiate wrt β_j's and solve the system of p equations :

$$\sum_{i=1}^{n} X_{ij}(Y_i - e^{\beta^\top \mathbf{X}_i}) = 0, \quad j = 1, \ldots, p. \tag{1}$$

$\hookrightarrow$ no explicit formula for the MLE $\Rightarrow$ numerical algorithms (e.g. Newton-Raphson, Fisher-scoring) are used

Matrix formulation

- in the linear model $Y_i = \beta^\top \mathbf{X}_i + \varepsilon_i$, the MLE (OLS) of β is[7] :

$$\hat{\beta}_n = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{Y}$$

This comes from the estimating equation

$$\mathbb{X}^\top (\mathbf{Y} - \mu) = 0$$

with $\mu = (\mu_1, \ldots, \mu_n)^\top$ and $\mu_i = \mathbb{E}(Y_i | \mathbf{X}_i) = \beta^\top \mathbf{X}_i$.

- in Poisson regression, the system (1) can also be expressed as :

$$\mathbb{X}^\top (\mathbf{Y} - \mu) = 0$$

where $\mu = (\mu_1, \ldots, \mu_n)^\top$ and $\mu_i = \mathbb{E}(Y_i | \mathbf{X}_i) = e^{\beta^\top \mathbf{X}_i}$.

- this comes from the fact that both linear and Poisson models are special cases of GLM

7. if $\mathbb{X}^\top \mathbb{X}$ is invertible, where $\mathbb{X}$ is the $(n \times p)$ design matrix

Inference in Poisson regression

For "n large", we have[8], for every coefficient β_j :

$$\hat{\beta}_{n,j} \approx \mathcal{N}(\beta_j, (\mathsf{se}(\hat{\beta}_{n,j}))^2)$$

or equivalently

$$\frac{\hat{\beta}_{n,j} - \beta_j}{\mathsf{se}(\hat{\beta}_{n,j})} \approx \mathcal{N}(0,1).$$

$\hookrightarrow$ "usual" properties of the MLE

From this, we can calculate confidence intervals and tests of hypothesis.

8. the estimated standard deviation of $\hat{\beta}_{n,j}$ is called the *standard error* of $\hat{\beta}_{n,j}$ and we have : $\mathsf{var}(\hat{\beta}_{n,j}) = (\mathsf{se}(\hat{\beta}_{n,j}))^2$

Confidence intervals for β_j

For example, a 95%-confidence interval for β_j is :

$$[\hat{\beta}_{n,j} - 1.96\,\mathsf{se}(\hat{\beta}_{n,j}), \hat{\beta}_{n,j} + 1.96\,\mathsf{se}(\hat{\beta}_{n,j})].$$

This comes from the fact that :

$$\mathbb{P}\left(-1.96 \leq \frac{\hat{\beta}_{n,j} - \beta_j}{\mathsf{se}(\hat{\beta}_{n,j})} \leq 1.96\right) \approx \underbrace{\mathbb{P}(-1.96 \leq \mathcal{N}(0,1) \leq 1.96)}_{0.95}.$$

Test for a single component of β (Wald test)

We wish to test that the j-th covariate in $\beta^\top \mathbf{X}_i$ is non-significant :

$$\mathcal{H}_0 : \beta_j = 0 \text{ against } \mathcal{H}_1 : \beta_j \neq 0$$

Under $\mathcal{H}_0$, the Wald statistic

$$z_j := \frac{\hat{\beta}_{n,j}}{\mathsf{se}(\hat{\beta}_{n,j})} \approx \mathcal{N}(0,1).$$

The following decision rule has level α :

$$\text{reject } \mathcal{H}_0 \text{ if } |z_j| \geq u_{1-\alpha/2}$$

Remark 3 (p-value)

The p-value is $\mathbb{P}(|\mathcal{N}(0,1)| > |z_j|)$. Reject $\mathcal{H}_0$ if p-value $\leq 5\%$.

Likelihood-ratio test (LRT)

We wish to test [9] :

$$\mathcal{H}_0 : \beta_1 = \ldots = \beta_q = 0$$

against

$$\mathcal{H}_1 : \text{ there exists } i \in \{1, \ldots, q\} \text{ such that } \beta_i \neq 0.$$

Remark 4

Useful for testing significance of a qualitative variable with more than 3 categories, e.g., stage of disease, age, smoking status with categories of : current smoker, former smoker, never smoked, and smoking status unknown.

9. without loss of generality, up to a re-ordering of the coefficients

Likelihood-ratio test (LRT)

The LRT compares the likelihoods under $\mathcal{H}_0$ and $\mathcal{H}_1$ and accepts $\mathcal{H}_0$ if they are "close." Define :

$$D_n = 2 \ln \left(\frac{L_n(\hat{\beta}_n)}{L_n(\hat{\beta}_{n,\mathcal{H}_0})} \right) = 2(\ln L_n(\hat{\beta}_n) - \ln L_n(\hat{\beta}_{n,\mathcal{H}_0}))$$

Under H_0, if n is "large",

$$D_n \approx \chi^2_q.$$

Therefore, we reject $\mathcal{H}_0$ at the level α if :

$$D_n \geq c_q(1-\alpha)$$

where $c_q(1-\alpha)$ is the $(1-\alpha)$-quantile of χ^2_q.

Remark 5 (p-value)

The p-value is $\mathbb{P}(\chi^2_q > D_n)$. Reject $\mathcal{H}_0$ if p-value $\leq 5\%$.

Model selection : information criteria

Several information criteria available :

- Akaike information criterion (AIC) is :

$$AIC = -2 \ln L_n(\hat{\beta}_n) + 2p$$

- Bayesian information criterion (BIC) is :

$$BIC = -2 \ln L_n(\hat{\beta}_n) + p \ln n$$

(with n the sample size and p the number of regressors)

$\hookrightarrow$ keep the model that minimizes the chosen criterion

Remark 6 (automatic selection procedures)

AIC and BIC can be used in automatic stepwise procedures for model selection (such as backward and forward procedures).

Poisson regression : an example in R

Dataset : `docvisits` in the `R` package `zic`.

- this data set gives the number of doctor visits in the last three months for a sample of German male individuals in 1994
- the data set is a subsample of the German Socioeconomic Panel (SOEP)
- the SOEP is a longitudinal survey of approximately 15,000 private households in Germany (starting from 1984 ; the SOEP Wave 38 was released in July 2023)

The equidispersion property

Poisson distribution is equidispersed : if $Y \sim \mathcal{P}(\lambda)$, then

$$\mathbb{E}(Y) = \text{var}(Y) = \lambda.$$

Aim of this part :

1. how can we check that equidispersion holds on a given data set ?
2. why should we check that equidispersion holds ? if answer is no, is this really an issue ?
3. what can we do if equidispersion does not hold ?

Checking equidispersion

Empirically, one can check equidispersion by calculating the sample mean and variance. E.g., in R :

```
> Y=rpois(100,2)
> mean(Y)
2.1
> var(Y)
2.171717
> var(Y)/mean(Y)
1.034151
```

Remark 7

The ratio $\phi = \frac{\text{var}(Y)}{\mathbb{E}(Y)}$ is called the dispersion index. Under equidispersion, ϕ should be "close" to 1.

Checking equidispersion

Checking equidispersion is more tricky in a Poisson regression model !

In the model $Y|\mathbf{X} \sim \mathcal{P}(\lambda(\mathbf{X}))$, equidispersion can be stated as :

$$\mathbb{E}(Y|\mathbf{X}) = \text{var}(Y|\mathbf{X}) = \lambda(\mathbf{X}).$$

With only one categorical covariate (e.g., smoking status), one could estimate $\mathbb{E}(Y|X = x)$ and $\text{var}(Y|X = x)$ for each x, and compare.

We will see a more practical tool with general $\mathbf{X} \hookrightarrow$ dispersion index

Equidispersion vs overdispersion

- if $\text{var}(Y_i|\mathbf{X}_i) > \mathbb{E}[Y_i|\mathbf{X}_i]$, data are said to be overdispersed
- intuitively, overdispersion arises when the variance of the data is underestimated $\Rightarrow$ Poisson standard errors are too small
- why is overdispersion a problem ? because both confidence intervals and tests of hypothesis are affected :
  - confidence intervals $[\hat{\beta}_{n,j} \pm 1.96\, \text{se}(\hat{\beta}_{n,j})]$ are too narrow $\Rightarrow$ incorrect coverage probabilities
  - Wald tests $z_j := \frac{\hat{\beta}_{n,j}}{\text{se}(\hat{\beta}_{n,j})}$ are too large $\Rightarrow$ non-significant explanatory variables may appear to be significant

Overdispersion : a simulation illustration

Overdispersion can arise for various reasons, e.g. :

- presence of unobserved heterogeneity in the data
- zero-inflation (not seen in this course)

For example, let Y_i be the number of episodes of chronic bronchitis and suppose :

$$Y_i \sim \begin{cases} \mathcal{P}(1) \text{ if } X_i = \text{non smoker (NS)} \\ \mathcal{P}(4) \text{ if } X_i = \text{smoker (S)} \end{cases}$$

Equivalently, $Y_i|X_i = \text{NS} \sim \mathcal{P}(1)$ and $Y_i|X_i = \text{S} \sim \mathcal{P}(4)$.

Overdispersion : a simulation illustration

Check equidispersion within each sub-group :

```
> Y.NS=rpois(100,1)
> Y.S=rpois(100,4)
> c(mean(Y.NS),mean(Y.S))
0.72 4.10
> c(var(Y.NS),var(Y.S))
0.7692929 3.9898990
> c(var(Y.NS)/mean(Y.NS),var(Y.S)/mean(Y.S))
1.0684624 0.9731461
```

Now, merge both sub-groups and do again :

```
> mean(c(Y.NS,Y.S))
2.41
> var(c(Y.NS,Y.S))
5.23809
> var(c(Y.NS,Y.S))/mean(c(Y.NS,Y.S))
2.173482
```

Overdispersion : a mathematical explanation

Let $Y_i|X_i = 0 \sim \mathcal{P}(\lambda_0)$ and $Y_i|X_i = 1 \sim \mathcal{P}(\lambda_1)$ and $X_i \sim \mathcal{B}(\pi_i)$. Suppose that we only observe Y_i (or that X_i is observed but its effect is not modeled).Then :

$$\begin{aligned}\mathbb{E}(Y_i) &= \mathbb{E}[\mathbb{E}[Y_i|X_i]] \\ &= \mathbb{E}(\lambda_0(1-X_i)+\lambda_1 X_i) \\ &= \pi_i\lambda_1 + (1-\pi_i)\lambda_0\end{aligned}$$

and

$$\begin{aligned}\text{var}(Y_i) &= \mathbb{E}[\text{var}(Y_i|X_i)] + \text{var}(\mathbb{E}[Y_i|X_i]) \\ &= \mathbb{E}(Y_i) + \text{var}(\lambda_0 + (\lambda_1-\lambda_0)X_i) \\ &= \mathbb{E}(Y_i) + (\lambda_1-\lambda_0)^2\pi_i(1-\pi_i).\end{aligned}$$

If $\pi_i \neq 0, 1$, then :

$$\text{var}(Y_i) > \mathbb{E}(Y_i)$$

and the distribution of Y_i is overdispersed.

Overdispersion : omitting covariates

Omission of one or several key explanatory variables from the linear predictor can yield overdispersion.

A numerical experiment :

1. simulate n indep. values of $X_1, X_2 \sim \mathcal{U}[0,1]$ and $X_3 \sim \mathcal{N}(0,1)$
2. simulate n indep. counts $Y_i \sim \mathcal{P}(e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}})$
3. discretize each X_k into 8 categories, yielding $N = 8^3 = 512$ classes $\mathbf{x}_1, \ldots, \mathbf{x}_N$
4. estimate $\mathbb{E}(Y|\mathbf{X} = \mathbf{x}_j)$ and $\text{var}(Y|\mathbf{X} = \mathbf{x}_j)$, $j = 1, \ldots, N$ and plot the N estimated pairs (blue circles)
5. regress the N empirical variances on the N empirical means (dashed blue line)
6. redo steps 4-5 while omitting X_3 (red dotted line)

Overdispersion : omitting covariates

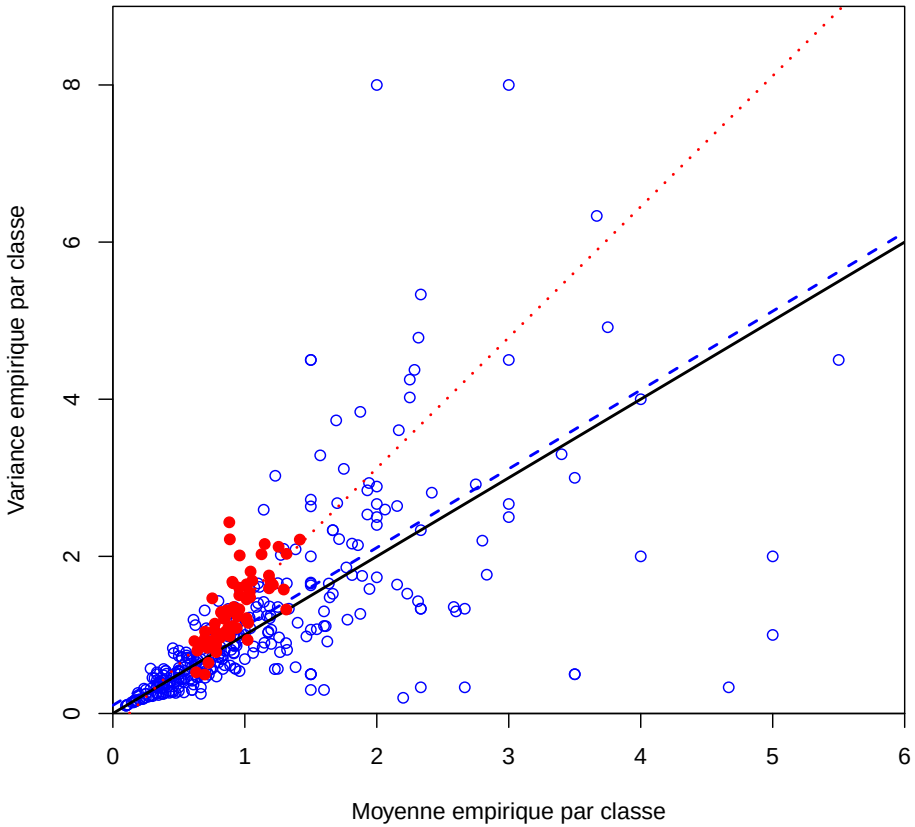

FIGURE 9 – Estimated relation between $\mathbb{E}(Y|\mathbf{X}=\mathbf{x})$ and $\text{var}(Y|\mathbf{X}=\mathbf{x})$ (solid line is $y=x$, i.e. equidispersed case).

Dispersion index

Let

$$\begin{aligned}\phi &= \frac{\text{var}(Y|\mathbf{X})}{\mathbb{E}(Y|\mathbf{X})} \quad \text{with } \mathbb{E}(Y|\mathbf{X}) = e^{\beta^\top \mathbf{X}} \\ &= \frac{\mathbb{E}((Y - e^{\beta^\top \mathbf{X}})^2|\mathbf{X})}{e^{\beta^\top \mathbf{X}}}\end{aligned}$$

To estimate ϕ, estimate β by its MLE $\hat{\beta}_n$ in a Poisson model, then $e^{\beta^\top \mathbf{X}}$ by $e^{\hat{\beta}_n^\top \mathbf{X}}$ and $\mathbb{E}((Y - e^{\beta^\top \mathbf{X}})^2|\mathbf{X})$ by the empirical variance

$$\frac{1}{n-p}\sum_{i=1}^{n}(Y_i - e^{\hat{\beta}_n^\top \mathbf{X}_i})^2.$$

Finally, estimate ϕ by :

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\left(\frac{Y_i - e^{\hat{\beta}_n^\top \mathbf{X}_i}}{\sqrt{e^{\hat{\beta}_n^\top \mathbf{X}_i}}}\right)^2 = \frac{1}{n-p}\sum_{i=1}^{n} r^2_{\text{Pearson},i}$$

Overdispersion : some solutions

Non-exhaustive list :

1. quasi-Poisson model
2. negative-binomial distribution

Basic idea

Under overdispersion, $\text{var}(Y|\mathbf{X}) \approx \hat{\phi}\, \mathbb{E}(Y|\mathbf{X})$. Poisson regression will underestimate the variability of the observations, and thus, the variance estimates, by a factor $\hat{\phi} \Rightarrow$ the idea is to correct the variance estimates by a factor $\hat{\phi}$.

- still assume that $\mathbb{E}(Y|\mathbf{X}) = e^{\beta^\top \mathbf{X}}$
- estimate β using the same equation as in Poisson model (but we **do not suppose** that the observations are Poisson !) :

$$\sum_{i=1}^{n} X_{ij}(Y_i - e^{\beta^\top \mathbf{X}_i}) = 0, \quad j = 1, \ldots, p$$

- calculate Poisson variance estimates $\text{var}_P(\hat{\beta}_{n,j}) = (\text{se}_P(\hat{\beta}_{n,j}))^2$
- correct the variance estimates by calculating

$$\text{var}_Q(\hat{\beta}_{n,j}) = \hat{\phi} \times \text{var}_P(\hat{\beta}_{n,j})$$

Remarks

Remark 8

- if $\text{var}_Q(\hat{\beta}_{n,j}) = \hat{\phi} \times \text{var}_P(\hat{\beta}_{n,j})$ then

$$\text{se}_Q(\hat{\beta}_{n,j}) = \sqrt{\hat{\phi}} \times \text{se}_P(\hat{\beta}_{n,j})$$

$\Rightarrow$ a correction factor of $\sqrt{\hat{\phi}}$ is introduced in the usual Poisson confidence intervals and tests of hypothesis

Remarks

Remark 9

- β is estimated by solving the same equation as in Poisson regression.
  However, we **do not** assume equidispersion, thus we **do not** assume that the sample comes from Poisson. Hence the name quasi-Poisson [a].
- in fact, we only assume :

$$\mathbb{E}(Y|\mathbf{X}) = e^{\beta^\top \mathbf{X}} \quad \text{and} \quad \text{var}(Y|\mathbf{X}) = \phi\, \mathbb{E}(Y|\mathbf{X})$$

  Practical implication : LRT and AIC/BIC no longer available !
- in `R`, quasi-Poisson is available in `glm` with the option `family=quasipoisson()`

a. Wedderburn, 1974

Quasi-Poisson in public health

FIGURE 10 – Quasi-Poisson : a brief research in PubMed

Quasi-Poisson in public health

FIGURE 11 – Quasi-Poisson : a brief research in PubMed (ctd)

Negative binomial model

An alternative solution to overdispersion : consider a more flexible distribution (than Poisson) that does not impose equidispersion.

The most widely used distribution in this context is the negative binomial (NB).

Negative binomial model

Consider a series of independent binary experiments, with success probability π. We repeat the experiment until the k-th success occurs (k fixed in $\{1, 2, \ldots\}$).
Let Y be the number of failures before the k-th success. The law of Y is the NB law. We have :

$$\mathbb{P}(Y = y) = C_{y+k-1}^{y} \pi^k (1-\pi)^y, \quad y = 0, 1, 2, \ldots$$

Remark 10

The case $k = 1$ is the geometrical law.

 A more tricky but more relevant interpretation : the NB law can be viewed as a Poisson-gamma mixture.

Negative binomial model

Let $U_i \sim G(1, \nu)$ and suppose

$$Y_i|U_i \sim \mathcal{P}(\lambda_i U_i) \text{ with } \lambda_i > 0.$$

Remark 11

In practice, U_i is unobserved. It allows to introduce heterogeneity between individuals by adding variability to their mean response.

Let $\kappa = 1/\nu$. Then Y_i has a (unconditional) NB law, and :

$$\mathbb{E}(Y_i) = \lambda_i$$

and $\text{var}(Y_i) = \lambda_i + \kappa\lambda_i^2 > \mathbb{E}(Y_i) \Rightarrow$ overdispersed law

Remark 12

$\text{var}(Y_i)$ is quadratic in $\mathbb{E}(Y_i)$, hence the name "NB2" model.

Negative binomial model

Proof : 

$$\begin{aligned}
\mathbb{P}(Y_i = y) &= \int_0^\infty e^{-\lambda_i u} \frac{(\lambda_i u)^y}{y!} \frac{\nu^\nu}{\Gamma(\nu)} u^{\nu-1} e^{-u\nu}\, du \\
&= \frac{\nu^\nu \lambda_i^y}{\Gamma(\nu) y!} \int_0^\infty e^{-(\lambda_i+\nu)u} u^{y+\nu-1}\, du \\
&= \frac{\nu^\nu \lambda_i^y}{\Gamma(\nu) y!} \frac{\Gamma(y+\nu)}{(\lambda_i+\nu)^{y+\nu}} \\
&= \frac{\Gamma(y+\nu)}{\Gamma(\nu) y!} \left(\frac{\lambda_i}{\nu+\lambda_i}\right)^y \left(\frac{\nu}{\nu+\lambda_i}\right)^\nu
\end{aligned}$$

Set $\kappa = 1/\nu$ and $\lambda = \lambda_i$. We recognize the density of the NB.

Negative binomial model

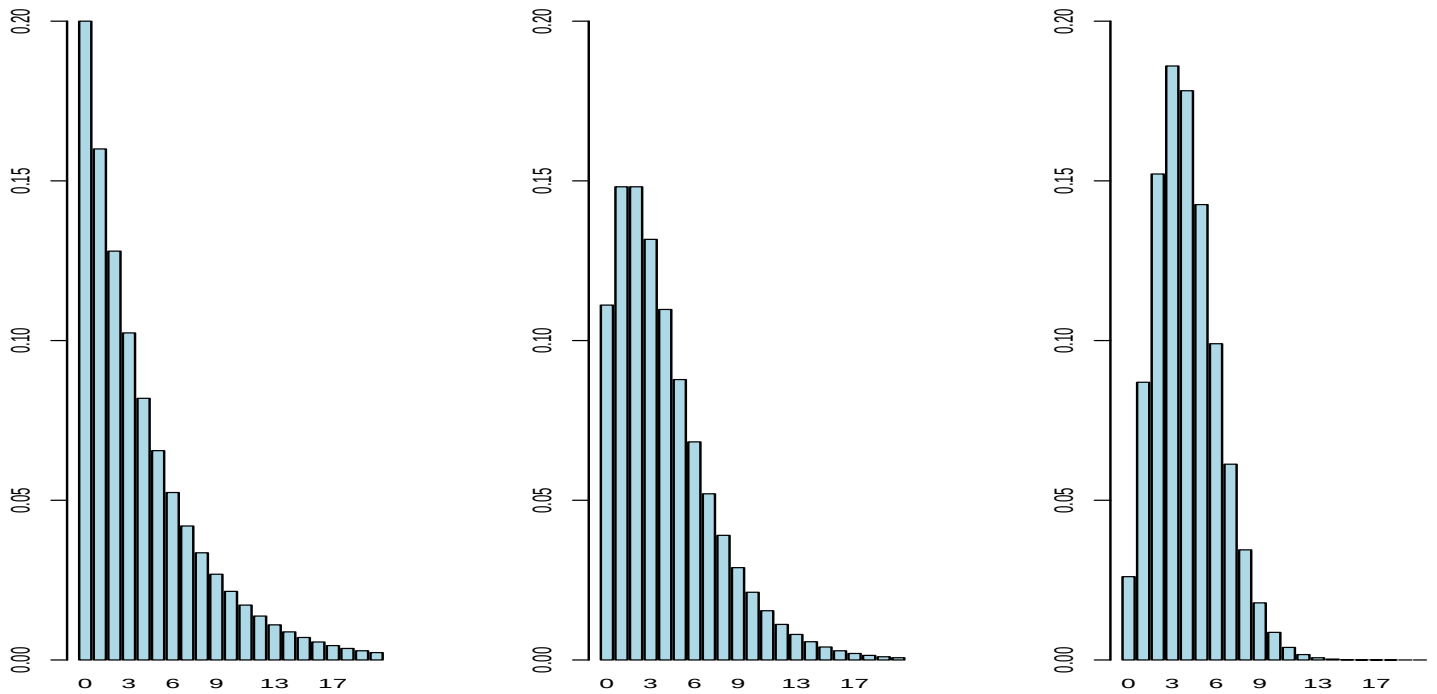

FIGURE 12 – NB laws with mean 3 and decreasing variance (left to right).

Negative binomial model

In the NB regression model,

$$\mathbb{E}(Y_i|\mathbf{X}_i) = \lambda(\mathbf{X}_i)$$

and $\text{var}(Y_i|\mathbf{X}_i) = \lambda(\mathbf{X}_i) + \kappa\lambda(\mathbf{X}_i)^2 > \mathbb{E}(Y_i|\mathbf{X}_i)$ and we set, as in a Poisson model :

$$\lambda(\mathbf{X}_i) = e^{\beta^\top \mathbf{X}_i}$$

β and κ are estimated by MLE (Fisher-scoring).

Negative binomial model

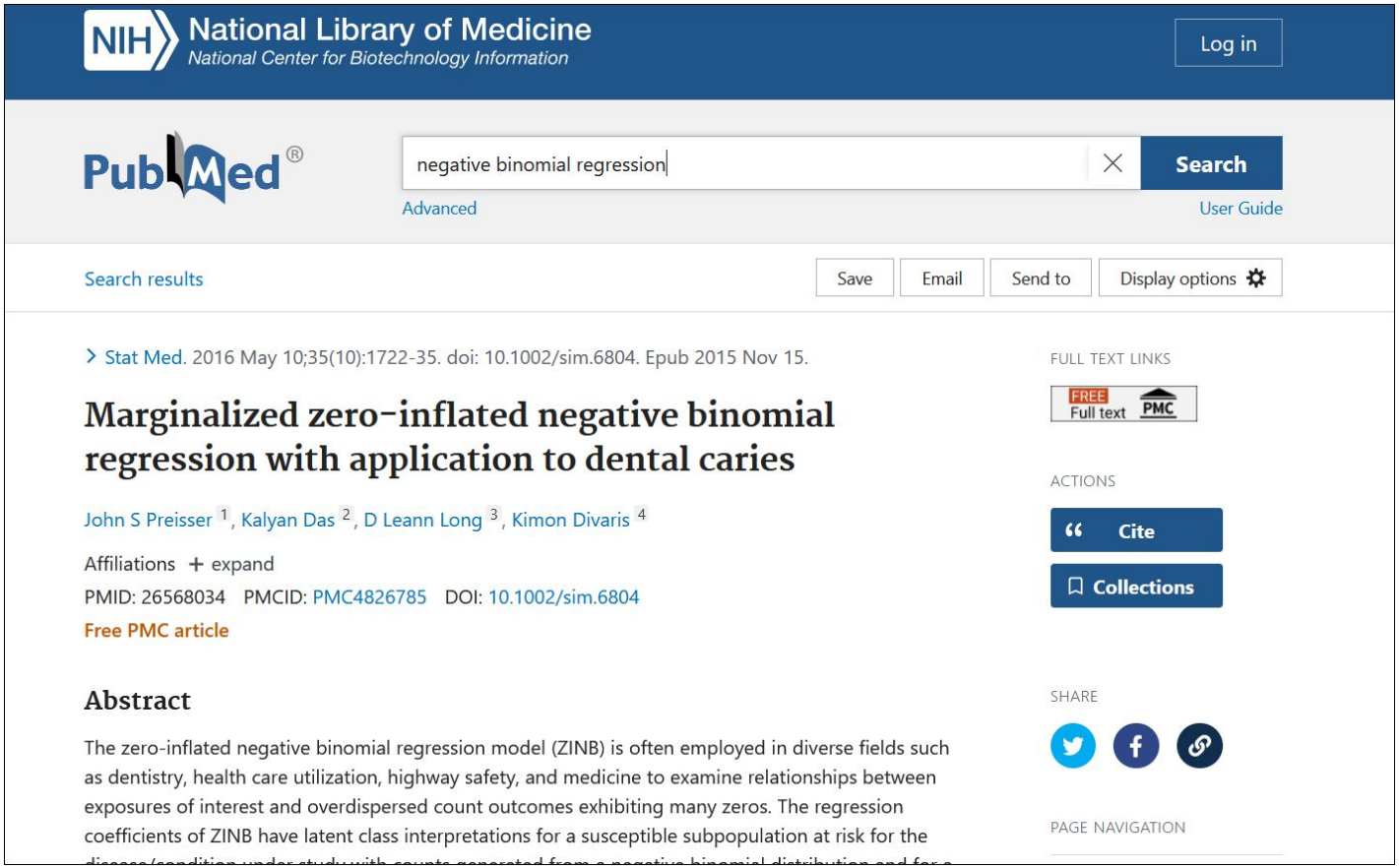

FIGURE 13 – A brief research in PubMed

Negative binomial model

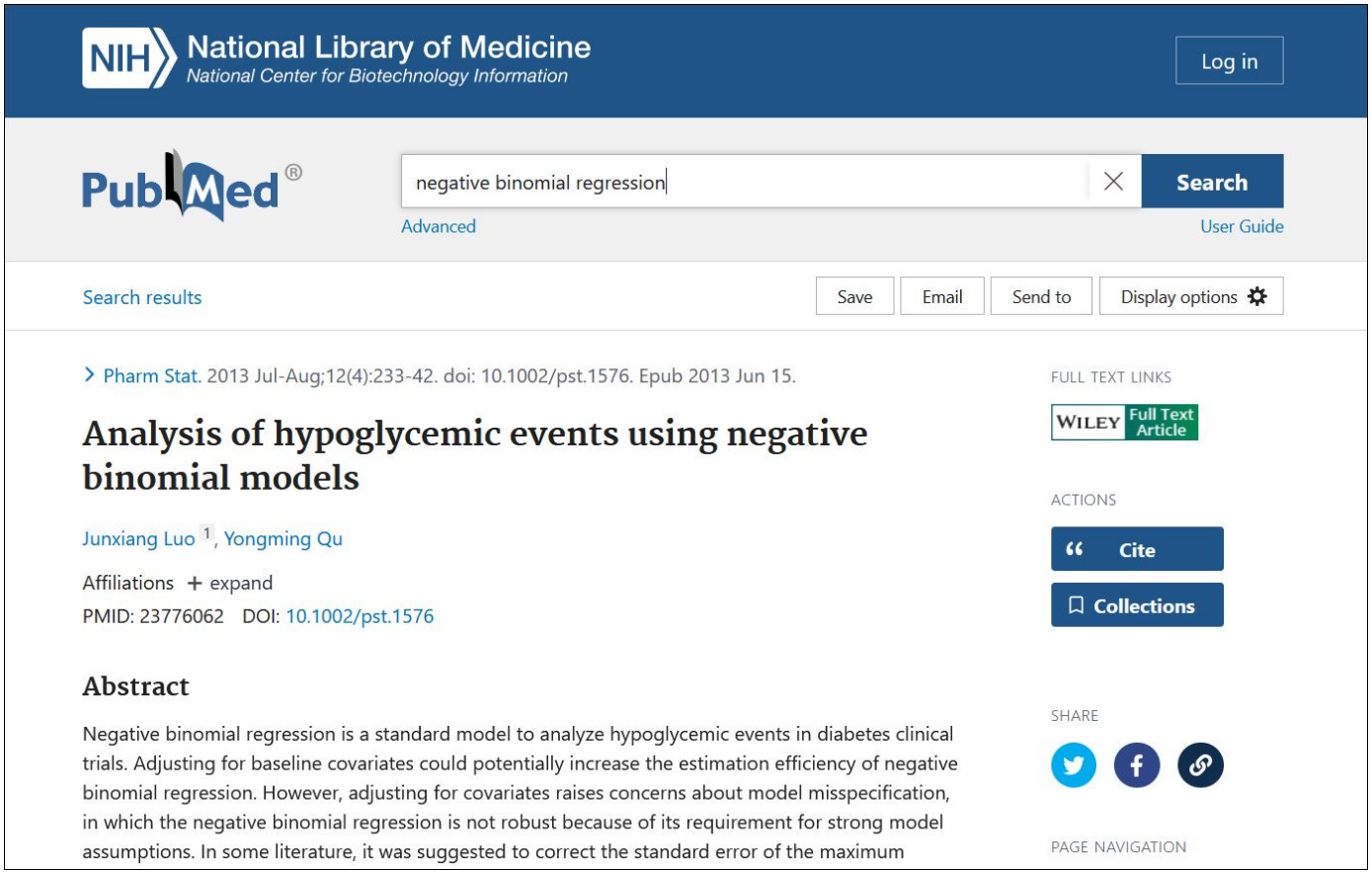

FIGURE 14 – A brief research in PubMed (ctd)

NB1 and NB2 variance functions

Another common parametrization is obtained with $U_i \sim G(\lambda_i, \phi\lambda_i)$ and $Y_i|U_i \sim \mathcal{P}(U_i)$, with $\phi > 0 \Rightarrow$ NB distribution with

$$\mathbb{E}(Y_i) = \lambda_i \text{ and } \text{var}(Y_i) = \lambda_i \left(1 + \frac{1}{\phi}\right).$$

The variance function is linear in μ_i (hence the name "NB1").

Overdispersion tests

Several tests are available to test $\mathcal{H}_0$: "the data are equidispersed" against $\mathcal{H}_1$: "the data are overdispersed".

$\hookrightarrow$ a likelihood ratio test (view Poisson as a special case of NB2 when $\kappa = 0$). Let :

$$LRT = 2\left(\ln L_n^{NB}(\hat{\beta}_n, \hat{\kappa}_n) - \ln L_n^{P}(\hat{\beta}_n)\right)$$

Under $\mathcal{H}_0$, if n is large,

$$LRT \sim \frac{1}{2}\delta_0 + \frac{1}{2}\chi_1^2.$$

Remark 13

This non-standard distribution arises because $\kappa \geq 0$ ($\mathcal{H}_0$ lies on the boundary of the parameter space). To test $\mathcal{H}_0$ at the level α, use $\chi_1^2(1-2\alpha)$ as critical value.
LRT available in the `odTest` function (`pscl` package).

NB vs quasi-Poisson

Remark 14

One advantage of NB model relative to quasi-Poisson is that it is associated with a formal likelihood so that information criteria (such as AIC) are readily available.

Let's apply quasi-Poisson and NB on the `docvisits` dataset (`R` package `zic`).