# Regressão de Poisson e parentes próximos

Filipa Januário

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Janeiro 2012

# Família Exponencial

Seja Y uma variável aleatória. A distribuição de probabilidade de Y pertence à família exponencial se a sua função densidade de probabilidade é da forma

$$f(y|\theta,\varphi) = \exp\left(\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi)\right) \tag{1}$$

onde

- $\theta$ é o parâmetro de localização, designado por parâmetro canónico
- $\varphi$ é o parâmetro de dispersão
- $b(.)$ e $c(.)$ são funções reais conhecidas.

Se $f$ verificar (1), diz-se que $Y$ está na forma canónica. Nessa situação, pode-se mostrar que

$$E(Y) = \mu = b'(\theta)$$

$$Var(Y) = \sigma^2 = \varphi b''(\theta)$$

# Exemplos

- Distribuição de Poisson: Y $\sim P(\lambda)$

$$f(y|\lambda) = \lambda^y \frac{e^{-y}}{y!} = \exp\left[\log(\lambda)y - \lambda - \log(y!)\right]$$

- Distribuição Binomial: Y $\sim B(n, p)$

$$f(y|n,p) = \binom{n}{p} p^y (1-p)^{n-y} = \exp\left[y \log\left(\frac{p}{1-p}\right) + n\log(1-p) + log\binom{n}{p}\right]$$

- Distribuição Binomial Negativa: Y $\sim BN(r, p)$

$$f(y|p,r) = \binom{y+r-1}{r-1} p^r (1-p)^y = \exp\left[y\log(1-p) + r\log(p) + \log\left(\frac{y+r-1}{r-1}\right)\right]$$

- Distribuição Normal, Distribuição Gamma,...

# Modelos Lineares Generalizados

No modelo de regressão linear usual:

- A variável resposta $Y$, condicionada pelas variáveis explicativas, segue uma distribuição Normal.
- A combinação linear das variáveis explicativas prevê diretamente a média da resposta.

No modelo linear generalizado:

- A variável resposta $Y$, condicionada pelas variáveis explicativas, segue uma distribuição pertencente à família exponencial.
- A combinação linear das variáveis explicativas prevê uma função (previamente definida) da média da resposta.

# Componentes de um Modelo Linear Generalizado

Considerando $n$ observações independentes $y_1, \ldots, y_n$ com $p$ covariáveis, $X_1, \ldots, X_p$, os MLG são caracterizados pela seguinte estrutura:

1. **Componente aleatória**: As observações da variável resposta $Y$, que se quer modelar, provêm de distribuições independentes que pertencem à família exponencial.

2. **Componente sistemática**: Consiste numa combinação linear de variáveis explicativas

$$\beta_0 + \beta_1 X_{1(i)} + \beta_2 X_{2(i)} + \ldots + \beta_p X_{p(i)}, \qquad i = 1, \ldots, n.$$

3. **Função de ligação**: função diferenciável e monótona g(.) que associa as componentes aleatória e sistemática, através duma relação da forma

$$g(\mu_i) = \beta_0 + \beta_1 X_{1(i)} + \beta_2 X_{2(i)} + \ldots + \beta_p X_{p(i)}, \qquad i = 1, \ldots, n.$$

Os MLG usam o método da Máxima Verosimilhança para estimar os coeficientes.

# Regressão de Poisson

- Utilizada para modelar **dados de contagem** num período de tempo ou espaço específico.

Assume-se que $\boxed{Y|X = x_i \sim P(\mu(x_i))}$. **Pretendemos modelar a média de $Y|X = x_i$ como combinação linear das variáveis explicativas.**
Uma modelação do tipo:

$$\mu(x_i) = \beta_0 + \beta_1 X_{1(i)} + \beta_2 X_{2(i)} + \ldots + \beta_p X_{p(i)}, \qquad i = 1, \ldots, n,$$

é **desajustada** dado que o termo do lado direito pode tomar qualquer valor real, enquanto que o termo de lado esquerdo só pode tomar valores não negativos.

**Uma Solução:** considerar a **função logarítmica** como função de ligação

$$\boxed{\log(\mu(x_i)) = \beta_0 + \beta_1 X_{1(i)} + \beta_2 X_{2(i)} + \ldots + \beta_p X_{p(i)}, \qquad i = 1, \ldots, n.}$$

# Regressão Binomial Negativa

- Utilizada para modelar **dados de contagem** quando a variância das observações excede a média. [1]
- $E(Y) = \mu$ e $Var(Y) = \mu + \alpha\mu^2$
- Assume-se que $\boxed{Y|X = x_i \sim BN(r, p(x_i))}$.
- Tal como na regressão de Poisson, pode considerar-se a função logaritmo como função de ligação:

$$\boxed{\log(\mu(x_i)) = \beta_0 + \beta_1 X_{1(i)} + \beta_2 X_{2(i)} + \ldots + \beta_p X_{p(i)}, \qquad i = 1, \ldots, n.}$$

[1] $E(Y) = Var(Y) = \mu$ na distribuição de Poisson

# Regressão Logística

- Utiliza-se quando a variável resposta é binária.

Considere-se que a variável resposta $Y$ é uma v.a binária, codificada por 0 e 1.

- Assume-se que $\boxed{Y|X = x_i \sim B(1, \pi(x_i))}$.

- A função de ligação mais usual neste tipo de regressão é o $logit(\pi_i) = \log\left(\dfrac{\pi_i}{1-\pi_i}\right)$:

$$\boxed{logit(\pi_i) = \beta_0 + \beta_1 X_{1(i)} + \beta_2 X_{2(i)} + \ldots + \beta_p X_{p(i)} \Leftrightarrow}$$

$$\Leftrightarrow \pi(x_i) = \frac{1}{1 + \exp^{-\left(\beta_0 + \beta_1 X_{1(i)} + \beta_2 X_{2(i)} + \ldots + \beta_p X_{p(i)}\right)}}$$

# Exemplos

família exponencial: $f(y,\theta,\varphi) = \exp\left[\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi)\right]$

| Distribuição $f(y,\theta,\varphi)$ | Função de ligação $\eta = g(\mu) = x.\beta$ | Modelo de regressão | $b(\theta)$ | $\varphi$ | $E(y) = b'(\theta)$ | $b''(\theta)$ | $\mathrm{Var}(y) = \varphi b''(\theta)$ |
|---|---|---|---|---|---|---|---|
| Normal$(\mu,\sigma^2)$ | $g(\mu) = \mu$ | linear | $\frac{\theta^2}{2}$ | $\sigma^2$ | $\mu = \theta$ | 1 | $\sigma^2$ |
| Binomial$(n,p)$ | $g(\mu) = \mathrm{logit}(p) = \mathrm{logit}(\frac{\mu}{n})$ | logística | $\log(1+\exp(\theta))$ | 1 | $\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$ | $\pi(1-\pi)$ | $\pi(1-\pi)$ |
| Poisson$(\lambda)$ | $g(\mu) = \log(\mu)$ | de Poisson | $\exp(\theta)$ | 1 | $\lambda = \exp(\theta)$ | $\lambda$ | $\lambda$ |
| Gamma$(\mu,\nu)$ | $g(\mu) = -\frac{1}{\mu}$ | Gama | $-\log(-\theta)$ | $\frac{1}{\nu}$ | $\mu = -\frac{1}{\theta}$ | $\mu^2$ | $\frac{\mu^2}{\nu}$ |
| IGauss$(\mu,\sigma^2)$ | $g(\mu) = -\frac{1}{\mu^2}$ | Gaussiana inversa | $-(-2\theta)^{1/2}$ | $\sigma^2$ | $\mu = (-2\theta)^{-1/2}$ | $\mu^3$ | $\mu^3\sigma^2$ |

# Modelos de Zeros Inflacionados

- Surgem com o intuito de modelar o excesso de zeros, e consequentemente sobredispersão, existente numa grande parte de dados de contagens reais.
- São aplicados em áreas como:
  - Medicina: ausência de dentes (dentes todos bons ou dentes implantados)
  - Ecologia: presença ou não de determinada espécie num determinado habitat
  - Produção: Defeitos de fabricação (contar o número de parafusos imperfeitos)
  - Alimentação: frequência alimentar de determinados alimentos

# Modelo de regressão de Poisson com zeros inflacionados

- Este modelo surge com o intuito de modelar o excesso de zeros existente nos dados.

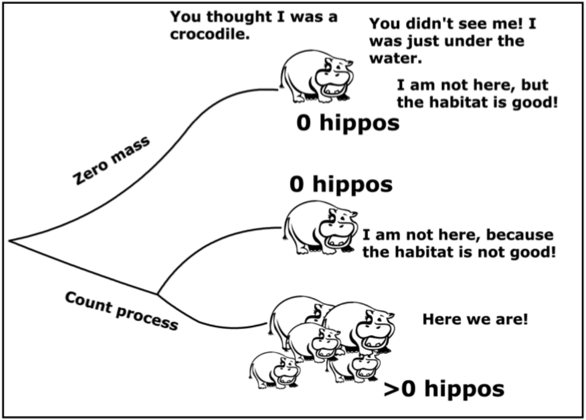
Considere-se uma amostra aleatória $Y_1, ..., Y_n$ da variável resposta $Y$. Para cada variável $Y_i$, existem **dois processos**:

$$Y_i \sim \begin{cases} 0 & \text{com probabilidade } \pi_i \\ Poisson(\mu_i) & \text{com probabilidade } 1 - \pi_i, \end{cases}$$

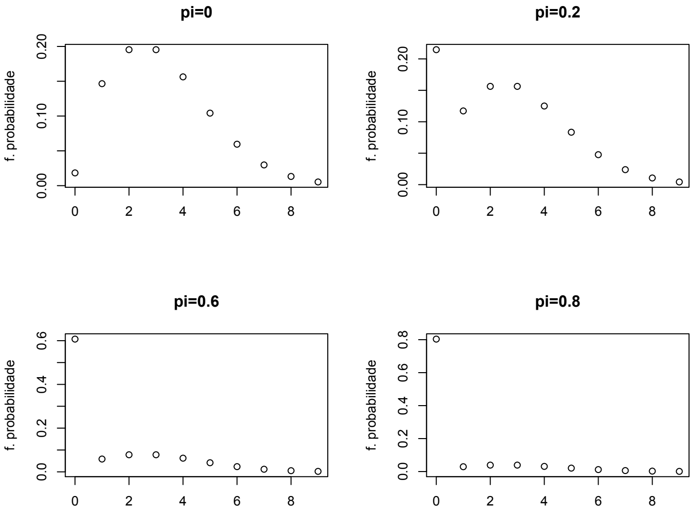
onde $\pi_i$ é a probabilidade de existirem zeros falsos. O modelo de zeros inflacionados de Poisson assume que:

$$P_r(Y_i = y_i|x_i) = \begin{cases} \pi_i + (1 - \pi_i)\exp(-\mu_i) & \text{PARA } y_i = 0 \\ (1 - \pi_i)\dfrac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} & \text{PARA } y_i \geq 1. \end{cases}$$

# Zeros Falsos vs Zeros Verdadeiros



# Comportamento da função de probabilidade do modelo ZIP



# Modelo de Regressão de Poisson com zeros inflacionados

Este modelo modela:

- **a média $\mu$ de uma variável de Poisson através de uma regressão de Poisson.**
- **a probabilidade $\pi$ de existirem zeros falsos através de uma regressão Logística.**

  Assim, os parâmetros $\mu = (\mu_1, ..., \mu_n)'$ e $\pi_i = (\pi_1, ..., \pi_n)'$ satisfazem:

$$\log(\mu) = X\beta$$
$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = Z\gamma,$$

onde $X$ e $Z$ são matrizes de observações das variáveis explicativas de interesse e $\beta$ e $\gamma$ (vetores) são os parâmetros da regressão.

# Como avaliar a qualidade do ajustamento de um modelo?

## Desviância:

$$D(y, \widehat{\mu}) = -2\varphi \log\left(\frac{\text{função de verosimilhança de um modelo } \omega}{\text{função de verosimilhança do modelo saturado }^{2}}\right)$$

- A desviância de um modelo avalia a discrepância entre os valores observados (modelos saturado) e os valores estimados pelo modelo $\omega$, ou seja, **avalia a qualidade do ajustamento do modelo**.
- O **valor de D será tanto maior** (em módulo) **quanto maior for a discrepância** entre o modelo estimado e os dados.

[2] melhor modelo em termos de ajustamento; tem um parâmetro por observação

# Teste de Hipóteses - Modelo $\omega$ *vs* Modelo Saturado

Pretende-se avaliar se um determinado modelo $\omega$ faz um bom ajustamento aos dados. Para isso, consideremos o seguinte teste:

$\mathbf{H_0}$ : o ajustamento do modelo $\omega$ é igual ao ajustamento do modelo saturado
$\mathbf{H_1}$ : os ajustamentos não são iguais ( e pto o ajustamento do modelo saturado é melhor)

**Estatística de teste:** Sob determinadas condições, $D \sim \mathrm{X}^2(\mathrm{df})$

onde df é o número de graus de liberdade que é igual a $p_s - p_\omega$, onde $p_s$ é o número de parâmetros do modelo saturado e $p_\omega$ é o número de parâmetros do modelo $\omega$.

**Decisão:** Rejeitar $H_0$ com um nível de significância $\alpha$ se $\boxed{\mathbf{D} > \mathbf{X}^2_{\mathbf{1}-\alpha}(\mathbf{df})}$

# Exemplo - PhD

Pretende-se saber qual o número médio de artigos publicados por um bioquímico nos últimos 3 anos de doutoramento.
As variáveis dos dados são:

- **art** - artigos publicados nos últimos 3 anos de doutoramento
- **fem** - sexo feminino (codificada como 1)
- **mar** - casado (codificada como 1)
- **kid5** - número de filhos com idade inferior a 6 anos
- **phd** - prestígio do programa de doutoramento, varia entre 0.75 e 5.
- **ment** - artigos publicados pelo orientador nos últimos três anos

# Exemplo - PhD

```
. summarize art

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         art |       915    1.692896    1.926069          0         19

. di r(Var)
3.7097416
```

- existem 915 observações (915 bioquímicos)
- a média do número de artigos publicados é de 1.69
- a variância do número de artigos publicados é de 3.71

# PhD - Regressão de Poisson

**Modelo:**

$$log(\mu(x)) = \beta_0 + \beta_1 * fem + \beta_2 * mar + \beta_3 * kid5 + \beta_4 * phd + \beta_5 * ment$$

```
. glm art fem mar kid5 phd ment, family(poisson) nolog

Generalized linear models                          No. of obs      =       915
Optimization     : ML                              Residual df     =       909
                                                   Scale parameter =         1
Deviance         =  1634.370984                    (1/df) Deviance =  1.797988
Pearson          =   1662.54655                    (1/df) Pearson  =  1.828984

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  3.621981
Log likelihood   = -1651.056316                    BIC             = -4564.031

------------------------------------------------------------------------------
             |                 OIM
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fem |  -.2245942   .0546138    -4.11   0.000    -.3316352   -.1175532
         mar |   .1552434   .0613747     2.53   0.011     .0349512    .2755356
        kid5 |  -.1848827   .0401272    -4.61   0.000    -.2635305   -.1062349
         phd |   .0128226   .0263972     0.49   0.627     -.038915    .0645601
        ment |   .0255427   .0020061    12.73   0.000     .0216109    .0294746
       _cons |   .3046168   .1029822     2.96   0.003     .1027755    .5064581
------------------------------------------------------------------------------
```

# PhD - Modelo de Regressão de Poisson

- Qual o número médio de artigos publicados por um bioquímico do sexo **masculino**,**casado**, com **1 filho com idade inferior a 6**, com um programa de doutoramento de **prestígio 5** e com **4 artigos** publicados pelo orientador nos últimos 3 anos?

$$log(\mu(x)) = \beta_0 + \beta_1 * fem + \beta_2 * mar + \beta_3 * kid5 + \beta_4 * phd + \beta_5 * ment$$

$$\Leftrightarrow \mu(x) = \exp(0.305 - 0.2245 * 0 + 0.1552 * 1 - 0.1849 * 1 + 0.0128 * 5 + 0.3046 * 4) \Leftrightarrow$$

$$\Leftrightarrow \mu(x) = 4.7$$

Tendo em conta o teste de hipóteses: [3]

- D=1634, df=909, $\alpha$ =0.05, $X^2_{0.95}(909)$=980.3

1634 > 980.3 logo **rejeitamos** $H_0$.

[3] $\mathbf{H_0}$ : o ajustamento do modelo $\omega$ é igual ao ajustamento do modelo saturado $\mathbf{H_1}$ : os ajustamentos não são iguais ( e pto o ajustamento do modelo saturado é melhor); **Estatística de teste:** Sob determinadas condições, $D_\omega \sim \mathrm{X}^2(909)$; **Decisão:** Rejeitar $H_0$ com um nível de significância $\alpha$ se $D > X^2_{1-\alpha}(909)$

# PhD

A tabela seguinte contém os dados sobre o número de artigos publicados tendo em conta o sexo, casamento, nº de filhos com idade inferior a 6, prestígio do programa doutoral e nº de artigos publicados pelo orientador.

| phd | $fem = 1 \| mar = 1 \| kid5 = 1 \| ment = [0-6]$ |
|---|---|
| [0-4.00] | 1.52 ← **média** |
| | 2.03 ←**variância** |
| | 27 ← **nº observações** |

| phd | $fem = 0 \| mar = 0 \| kid5 = 1 \| ment = [23-30]$ |
|---|---|
| [0-4.00] | 4.29 ← **média** |
| | 11.24 ←**variância** |
| | 7 ← **nº observações** |

- Existe sobredispersão $\Rightarrow$ Modelo de Regressão Binomial Negativa

# PhD - Modelo de Regressão Binomial Negativa

**Modelo:**

$$log(\mu(x)) = \beta_0 + \beta_1 * fem + \beta_2 * mar + \beta_3 * kid5 + \beta_4 * phd + \beta_5 * ment$$

```
. glm art fem mar kid5 phd ment, family(nb `v') nolog

Generalized linear models                   No. of obs      =        915
Optimization     : ML                       Residual df     =        909
                                            Scale parameter =          1
Deviance         =     1004.2815            (1/df) Deviance =    1.10482
Pearson          =   944.5494622            (1/df) Pearson  =   1.039108

Variance function: V(u) = u+(.4416000000000001)u^2 [Neg. Binomial]
Link function    : g(u) = ln(u)                    [Log]

                                            AIC             =   3.425046
Log likelihood   = -1560.958338             BIC             =   -5194.12

------------------------------------------------------------------------------
             |                 OIM
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fem |  -.2164184   .0726706    -2.98   0.003    -.3588501   -.0739867
         mar |   .1504895   .0821062     1.83   0.067    -.0104358    .3114147
        kid5 |  -.1764152   .0530587    -3.32   0.001    -.2804084   -.0724221
         phd |   .0152712   .0360382     0.42   0.672    -.0553624    .0859047
        ment |   .0290823   .0034657     8.39   0.000     .0222896    .0358751
       _cons |    .256144   .1385256     1.85   0.064    -.0153613    .5276493
------------------------------------------------------------------------------
```

# PhD - Modelo de Regressão Binomial Negativa

- Qual o número médio de artigos publicados por um bioquímico do sexo **masculino**,**casado**, com **1 filho com idade inferior a 6**, com um programa de doutoramento de **prestígio 5** e com **4 artigos** publicados pelo orientador nos últimos 3 anos?

$$log(\mu(x)) = \beta_0 + \beta_1 * fem + \beta_2 * mar + \beta_3 * kid5 + \beta_4 * phd + \beta_5 * ment$$

$$\Leftrightarrow \mu(x) = \exp(0.2561 - 0.02164*0 + 0.1505*1 - 0.1764*1 + 0.0153*5 + 0.0291*4) \Leftrightarrow$$

$$\mu(x) = 1.5267$$

Tendo em conta o teste de hipóteses: [4]

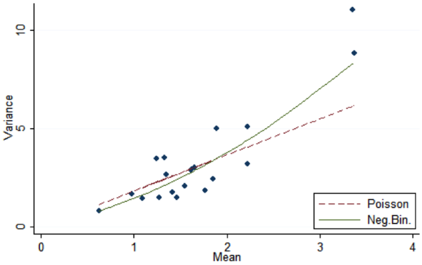
- D=1004, df=909, $\alpha$ =0.05, $X^2_{0.95}(909)$=980.3

$$1004 > 980.3 \text{ logo } \textbf{rejeitamos } H_0.$$

- O modelo de regressão binomial negativa faz um ajustamento **muito melhor** do que o modelo de regressão de poisson.

[4] $\mathbf{H_0}$ : o ajustamento do modelo $\omega$ é igual ao ajustamento do modelo saturado $\mathbf{H_1}$ : os ajustamentos não são iguais ( e pto o ajustamento do modelo saturado é melhor); **Estatística de teste:** Sob determinadas condições, $D_\omega \sim X^2(909)$; **Decisão:** Rejeitar $H_0$ com um nível de significância $\alpha$ se $D > X^2_{1-\alpha}(909)$

# PhD - Modelo de Regressão de Poisson *vs* Modelo de Regressão Binomial Negativa




- A função variância no caso da distribuição de Poisson faz um bom trabalho na maior parte dos dados, mas não consegue captar as variações mais elevadas dos bioquímicos mais produtivos.
- A função variância no caso da distribuição Binomial Negativa não é muito diferente mas sendo uma função quadrática, faz um melhor trabalho nas variações mais elevadas.

# PhD - Modelo de Regressão de Poisson com zeros inflacionados

```
. sum zobs zfitp

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        zobs |       915    .3005464    .4587464          0          1
       zfitp |       915    .2092071    .0794247   .0000659   .4113403
```

- 30% dos bioquímicos não publicaram artigos nos últimos 3 anos de doutoramento
- O modelo de Regressão de Poisson prevê que apenas 20.9% dos bioquímicos não publicaram artigos nos últimos 3 anos de doutoramento -**existe inflação de zeros**.

**Modelo ZIP:**

$$log(\mu(x)) = \beta_0 + \beta_1 * fem + \beta_2 * mar + \beta_3 * kid5 + \beta_4 * phd + \beta_5 * ment$$

$$logit(\pi(x)) = \gamma_0 + \gamma_1 * fem + \gamma_2 * mar + \gamma_3 * kid5 + \gamma_4 * phd + \gamma_5 * ment$$

# PhD - Modelo de Regressão de Poisson com zeros inflacionados

- **Zeros falsos** - aqueles estudantes para quem publicar não é relevante; querem entrar, por exemplo, no mundo do trabalho e não estão interessados em publicar.

- **Zeros verdadeiros** - aqueles estudantes cujas publicações são importantes mas não conseguem publicar por determinado motivo.

# PhD - Modelo de Regressão de Poisson com zeros inflacionados

```
Zero-inflated Poisson regression                  Number of obs   =       915
                                                  Nonzero obs     =       640
                                                  Zero obs        =       275

Inflation model = logit                           LR chi2(5)      =     78.56
Log likelihood  = -1604.773                       Prob > chi2     =    0.0000

------------------------------------------------------------------------------
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
art          |
         fem |  -.2091446   .0634047    -3.30   0.001    -.3334155   -.0848737
         mar |    .103751    .071111     1.46   0.145     -.035624    .243126
        kid5 |  -.1433196   .0474293    -3.02   0.003    -.2362793   -.0503599
         phd |  -.0061662   .0310086    -0.20   0.842     -.066942    .0546096
        ment |   .0180977   .0022948     7.89   0.000     .0135999    .0225955
       _cons |    .640839   .1213072     5.28   0.000     .4030814    .8785967
-------------+----------------------------------------------------------------
inflate      |
         fem |   .1097465   .2800813     0.39   0.695    -.4392028    .6586958
         mar |  -.3540107   .3176103    -1.11   0.265    -.9765155    .2684941
        kid5 |   .2171001    .196481     1.10   0.269    -.1679956    .6021958
         phd |   .0012702   .1452639     0.01   0.993    -.2834418    .2859821
        ment |   -.134111   .0452461    -2.96   0.003    -.2227918   -.0454302
       _cons |  -.5770618   .5093853    -1.13   0.257    -1.575439     .421315
------------------------------------------------------------------------------
```

# PhD - Modelo de Regressão de Poisson com zeros inflacionados

- Tendo em conta a equação de inflação, a única variável que **influencia a probabilidade de existirem zeros falsos** é a variável **ment**.

$$OR = \frac{Odds(P|ment+1)}{Odds(P|ment)} = \exp(-0.134111) \Leftrightarrow$$

$$\Leftrightarrow \frac{Odds(P|ment+1)}{Odds(P|ment)} = 0.875 \Leftrightarrow$$
$$\Leftrightarrow Odds(P|ment+1) = (1-0.125)Odds(P|ment) \Leftrightarrow$$

$$\Leftrightarrow Odds(P|ment+1) = Odds(P|ment) - 0.125\,Odds(P|ment).$$

- Por cada artigo publicado pelo orientador, o odds para a publicação de artigos pelos alunos bioquímicos baixa de 12.5 %.

# PhD - Modelo de Regressão de Poisson com zeros inflacionados

- Analisando a equação que traduz o número médio de artigos, verifica-se que as variáveis **fem**, **kid5** e **ment são estatisticamente significativas**.
- Encontram-se desvantagens para bioquímicos do sexo feminino* ou com filhos com idade inferior a 6 anos.

  $$* \; RR = \frac{E(P|F)}{E(P|\overline{F})} = \exp(-0.209) \Leftrightarrow E(P|F) = 0.81E(P|\overline{F}) \Leftrightarrow$$
  $$E(P|F) = (1-0.19)E(P|\overline{F}) \Leftrightarrow E(P|F) = E(P|\overline{F}) - 0.19E(P|\overline{F})$$
- Um bioquímico do sexo feminino publica **em média** menos 19% do que um bioquímico do sexo masculino.

# PhD - O modelo ZIP resolveu o problema do excesso de zeros?

O modelo prevê:

- $\pi$ para saber qual a probabilidade de não existirem publicações (zeros falsos)
- $\mu$ para prever o número médio de publicações

```
. gen zfitz = pz + (1-pz)*exp(-muz)

. sum zfitz

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       zfitz |       915    .2985684    .1280144    .0007119   .5815108
```

- O modelo previu que $29,9\%$ (cerca de 273) dos bioquímicos não publicaram nenhum artigo, muito perto do valor observado, 30%.
- **O modelo ZIP resolveu o problema do excesso de zeros.**

# Referências

[1] Cheung, Y. (2002). Zero-Inflated models for regression analysis of count data: a study of growth and development. In: Statistics in Medicine, n° 21, pp. 1461-1469.

[2] Hilbe, J. M. (2011). Negative Binomial Regression. Cambridge:University Press.

[3] Ridout, M.; Demétrio, C.; Hinde, J. (1998). Models for Count data with many zeros. In: International Biometric Conference, Cape Town.

[4] Rodríguez,G. (2007). Lecture Notes on Generalized Linear Models. Disponível em `http://data.princeton.edu/wws509/`.

[5] Zuur, A.; Ieno, E.; Walker, N.; Saveliev, A.; Smith, G. (2009). Mixed Effects Models and Extensions in Ecology with R. New York: Springer Science+Bussiness Media.

# Regressão de Poisson e parentes próximos

Filipa Januário

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Janeiro 2012