MPS814 - Tópicos Especiais em Epidemiologia

R para epidemiologia

Rodrigo Citton Padilha dos Reis
rodrigocpdosreis@gmail.com

Universidade Federal de Minas Gerais
Faculdade de Medicina
Programa de Pós-Graduação em Saúde Pública

Belo Horizonte
Julho de 2017

Modelos de regressão no R

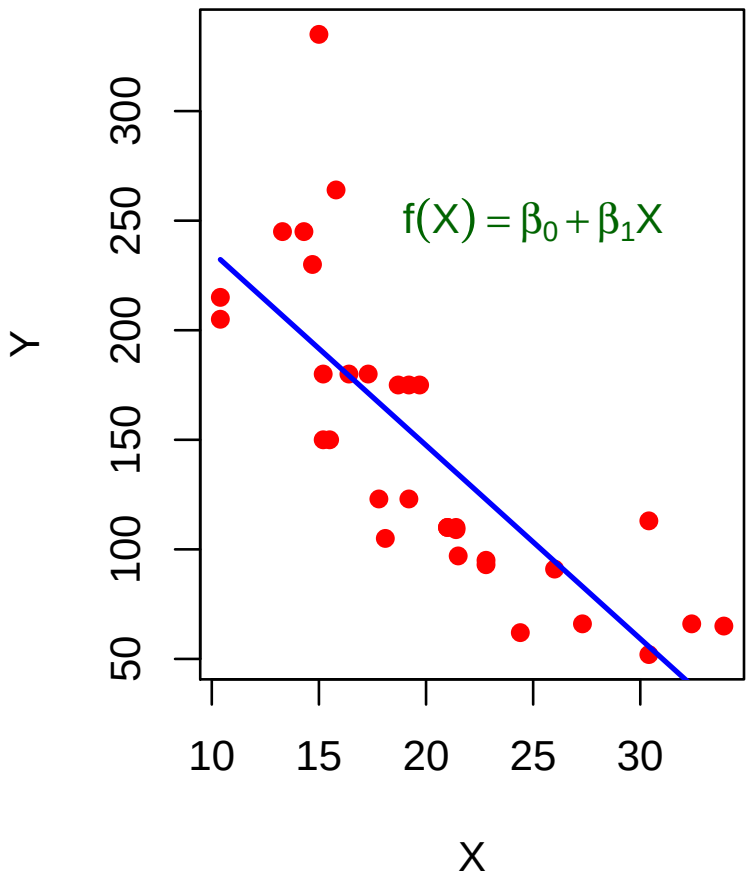
Modelos estatísticos

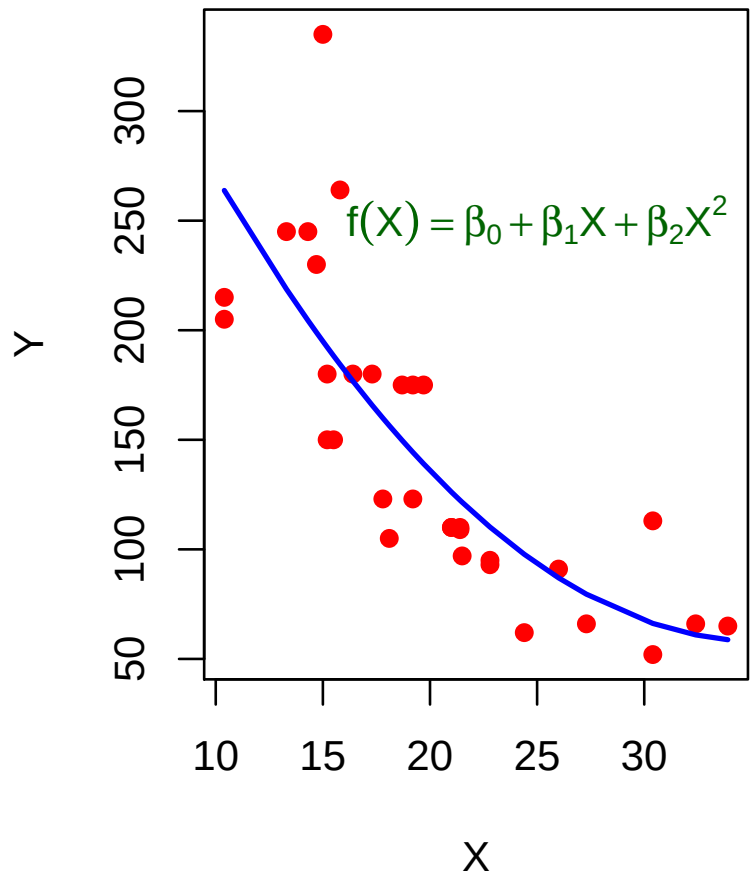
Suponha que observamos uma variável resposta Y e p diferentes preditores, $X_1, X_2, \ldots, X_p$. Vamos assumir que existe alguma relação entre Y e $X = (X_1, X_2, \ldots, X_p)$ que pode ser escrita em uma forma bastante geral

$$Y = f(X) + \epsilon,$$

em que f é uma **função fixa**, porém **desconhecida** de X e ϵ é um termo de **erro aleatório**.

Modelos estatísticos





Modelagem: conjunto de técnicas para especificar e estimar f.

Modelos estatísticos

Breiman, 2001[1] sugere que existem dois objetivos na análise de dados:

- **Predição**: ser capaz de prever o que as respostas vão ser para futuras variáveis de entrada.
- **Informação**: extrair algumas informações sobre como a natureza está associando as variáveis de resposta às variáveis de entrada.

[1] Breiman, L. Statistical modeling: the two cultures. *Statistical Science*, 16: 199-231, 2001.

Modelos estatísticos

To Explain or to Predict?[2]

- **Modelagem explicativa:** aplicação de modelos estatísticos aos dados para testar hipóteses causais.
- **Modelagem preditiva:** aplicação de modelos estatísticos **(mineração de dados, aprendizado estatístico/máquina)** para predição/classificação de novas ou futuras observações.
- **Modelagem descritiva:** aplicação de modelos estatísticos para representar de maneira compacta a estrutura dos dados; captura a associação entre as variáveis dependente e independentes **(ausência de hipóteses causais)**.

As abordagens explicativa e preditiva são diferentes

- No entanto, não são necessariamente inconsistentes ou incompatíveis.

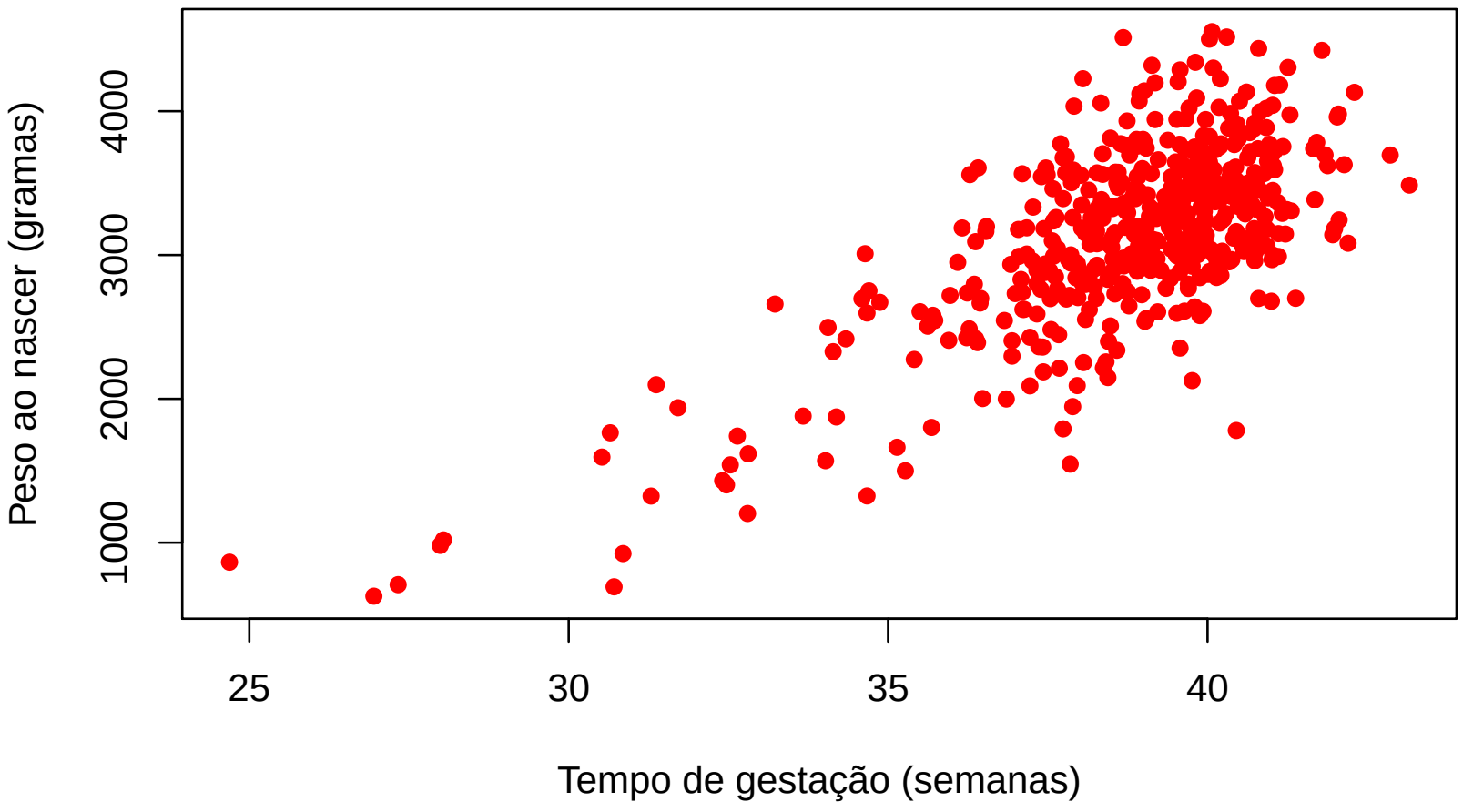
[2] Shmueli, G. To explain or to predict. *Statistical Science*, 25: 289-310, 2010.

O modelo de regressão linear

- Considere o conjunto de dados `births` do pacote `Epi`.

```
library(Epi)
data(births)
plot(bweight ~ gestwks,
     data = births,
     pch = 16,
     col = "red",
     xlab = "Tempo de gestação (semanas)",
     ylab = "Peso ao nascer (gramas)")
```

O modelo de regressão linear



O modelo de regressão linear

- O modelo mais **simples** que podemos pensar é o da equação da reta:

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

em que Y representa a variável "peso ao nascer", x representa a variável "tempo de gestação", β_0 e β_1 são coeficientes de regressão e ϵ é um termo de erro aleatório.

- Chamamos este modelo de **modelo de regressão linear**.
- Assumimos que $\epsilon \overset{indep.}{\sim} N(0, \sigma^2)$.
- β_0 representa o ponto em que a reta intercepta o eixo y.
- β_1 representa o incremento esperado de Y condicional ao valor de x.
- **Inferência:** em geral, é feita por **mínimos quadrados** e **máxima verossimilhança**.

O modelo de regressão linear no R

- Para ajustarmos um modelo de regressão linear no R utilizaremos a função `lm()`.

```
mod.lm <- lm(bweight ~ gestwks,
             data = births)
```

- Note que a função `lm()` resulta em um objeto da classe `lm` (Sua vez: `class(mod.lm)`).
- As seguintes funções podem ser aplicadas a um objeto `lm`:
  - `print()`: imprime os coeficientes estimados.
  - `summary()`: resumo do ajuste do modelo.
  - `coef()`: imprime os coeficientes estimados.
  - `plot()`: gera gráficos de resíduos.
  - `resid()`, `fitted()`, `predict()`, `abline()`, `confint()`.

O modelo de regressão linear no R

```
print(mod.lm)
```

```
##
## Call:
## lm(formula = bweight ~ gestwks, data = births)
##
## Coefficients:
## (Intercept)      gestwks
##       -4489          197
```

O modelo de regressão linear no R

```
summary(mod.lm)
```

```
##
## Call:
## lm(formula = bweight ~ gestwks, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1698.40  -280.14    -3.64   287.61  1382.24
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4489.140    340.899  -13.17   <2e-16 ***
## gestwks       196.973      8.788   22.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 449.7 on 488 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.5073, Adjusted R-squared:  0.5062
## F-statistic: 502.4 on 1 and 488 DF,  p-value: < 2.2e-16
```

O modelo de regressão linear no R

```
coef(mod.lm)
```

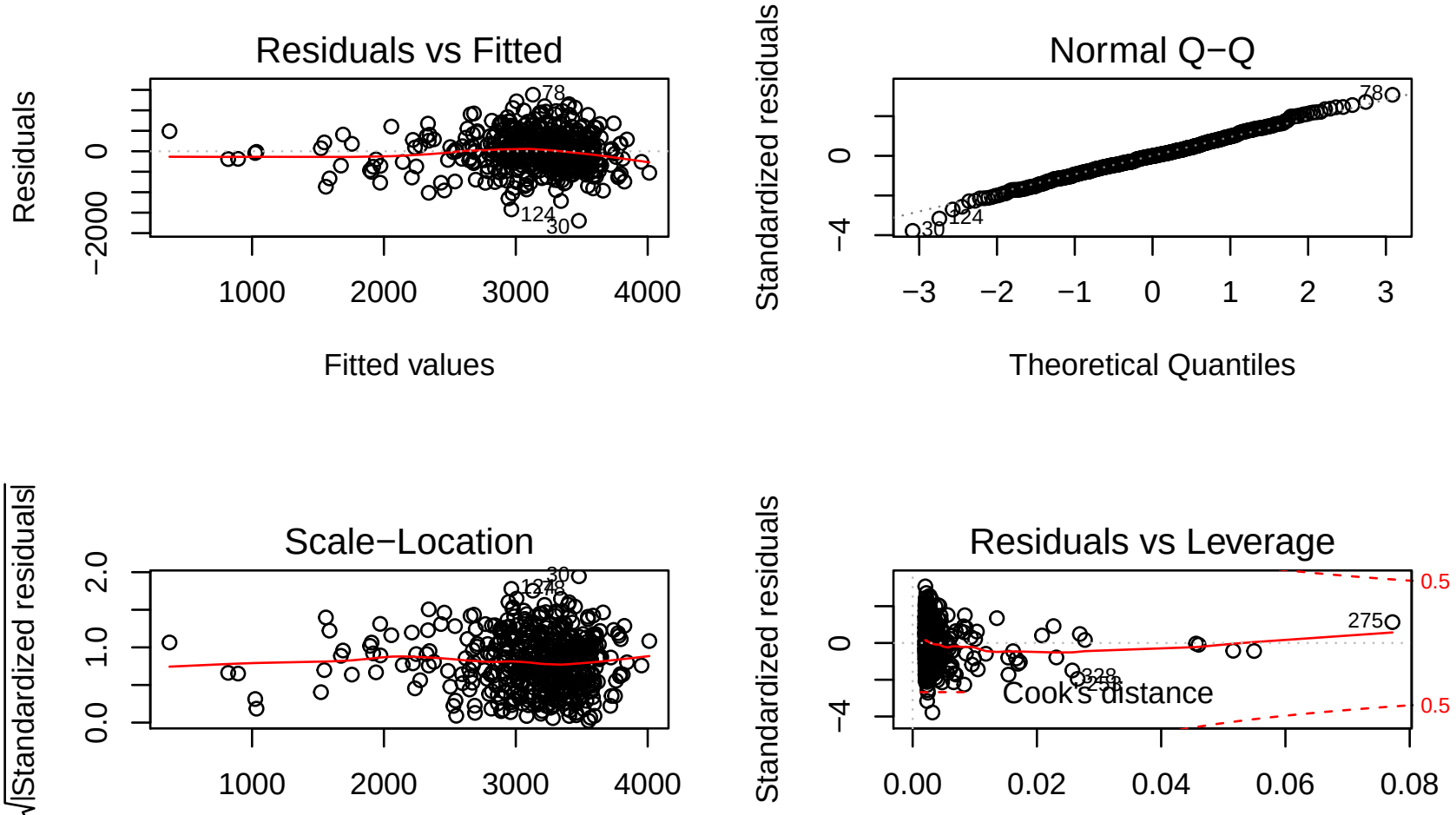
```
## (Intercept)     gestwks
##  -4489.1398    196.9726
```

```
confint(mod.lm, level = 0.95)
```

```
##                    2.5 %     97.5 %
## (Intercept) -5158.9503 -3819.3293
## gestwks       179.7054   214.2399
```

O modelo de regressão linear no R

```
par(mfrow = c(2,2))
plot(mod.lm)
```



O modelo de regressão linear no R

- Veja que estas funções, quando aplicadas ao objeto `lm`, resultam em novos objetos.

```
summary(mod.lm)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -4489.1398 340.898753 -13.16854 4.219248e-34
## gestwks       196.9726   8.788133  22.41348 5.064415e-77
```

O modelo de regressão linear no R

- **Observação:** o principal argumento da função `lm()` é um objeto `formula`.
  - `y ~ x`: regressão linear simples.
  - `y ~ x - 1`: reta passando pela origem ($\beta_0 = 0$).
  - `y ~ x1 + x2 + x3`: regressão múltipla.
  - `y ~ .`: regressão múltipla (com todas a variáveis do dataframe).
  - `y ~ x + I(x^2)`: regressão quadrática.
  - `log(y) ~ x1 + x2`: regressão múltipla da variável transformada.
  - `y ~ x1*x2`: regressão múltipla com termo de interação.

Sua vez!

1. Explore as funções `lm()`, `summary()`, etc. com diferentes definições do argumento `formula` para o conjunto de dados `births` do pacote `Epi`.
2. Veja quais são os outros argumentos da função `lm()`.

Apresentando resultados

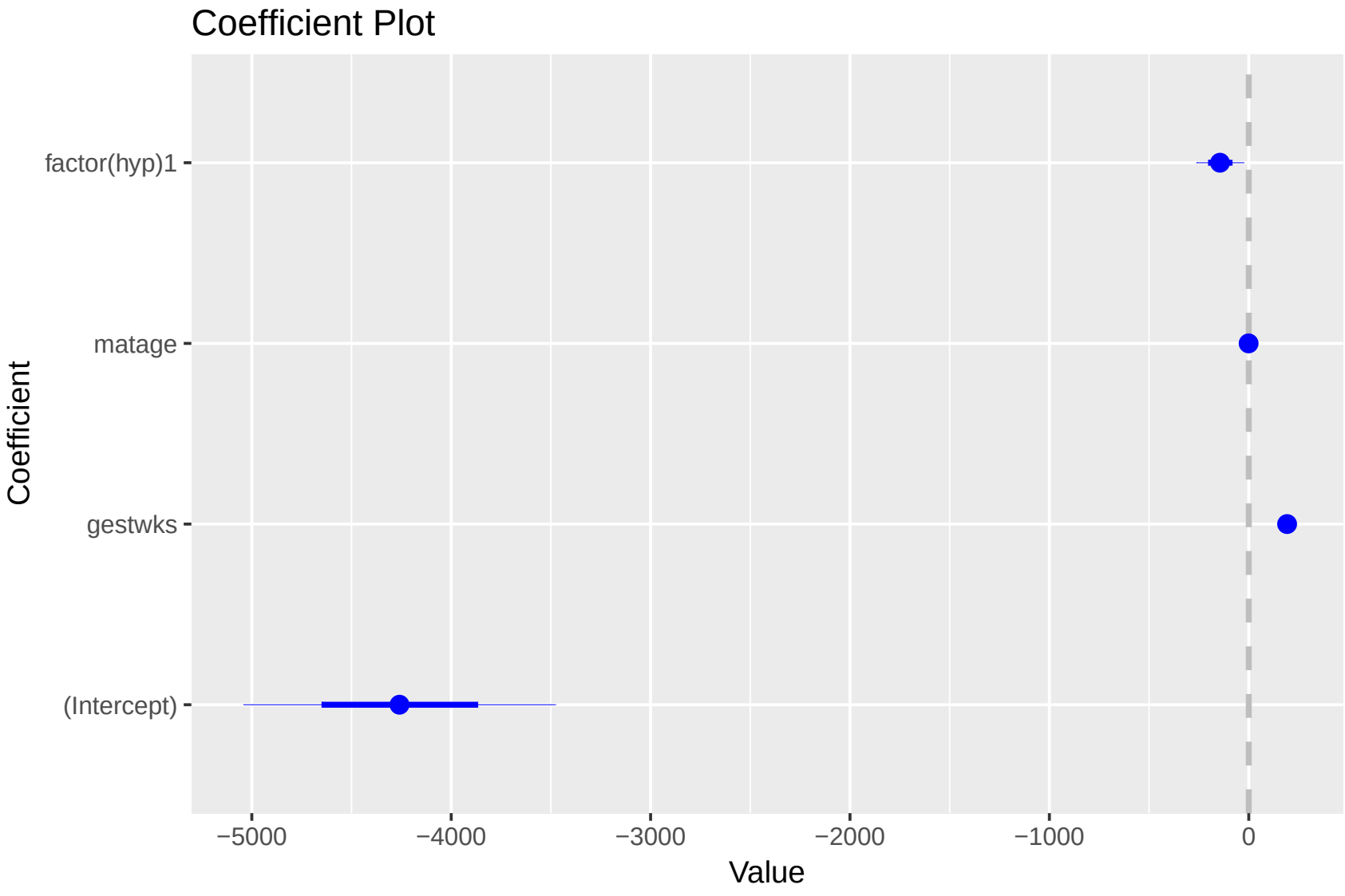
```
mod.lm2 <- lm(bweight ~ gestwks + matage + factor(hyp),
              data = births)
library(knitr)
kable(summary(mod.lm2)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4259.381056	390.643186	-10.9035079	0.0000000
gestwks	192.252362	8.965506	21.4435590	0.0000000
matage	-0.765879	5.206016	-0.1471142	0.8831029
factor(hyp)1	-144.223427	58.997058	-2.4445868	0.0148560

Apresentando resultados

```
library(coefplot)
coefplot(mod.lm2)
```

Apresentando resultados



Apresentando resultados

```
library(stargazer)
stargazer(mod.lm, mod.lm2, header = FALSE)
```

Table 2

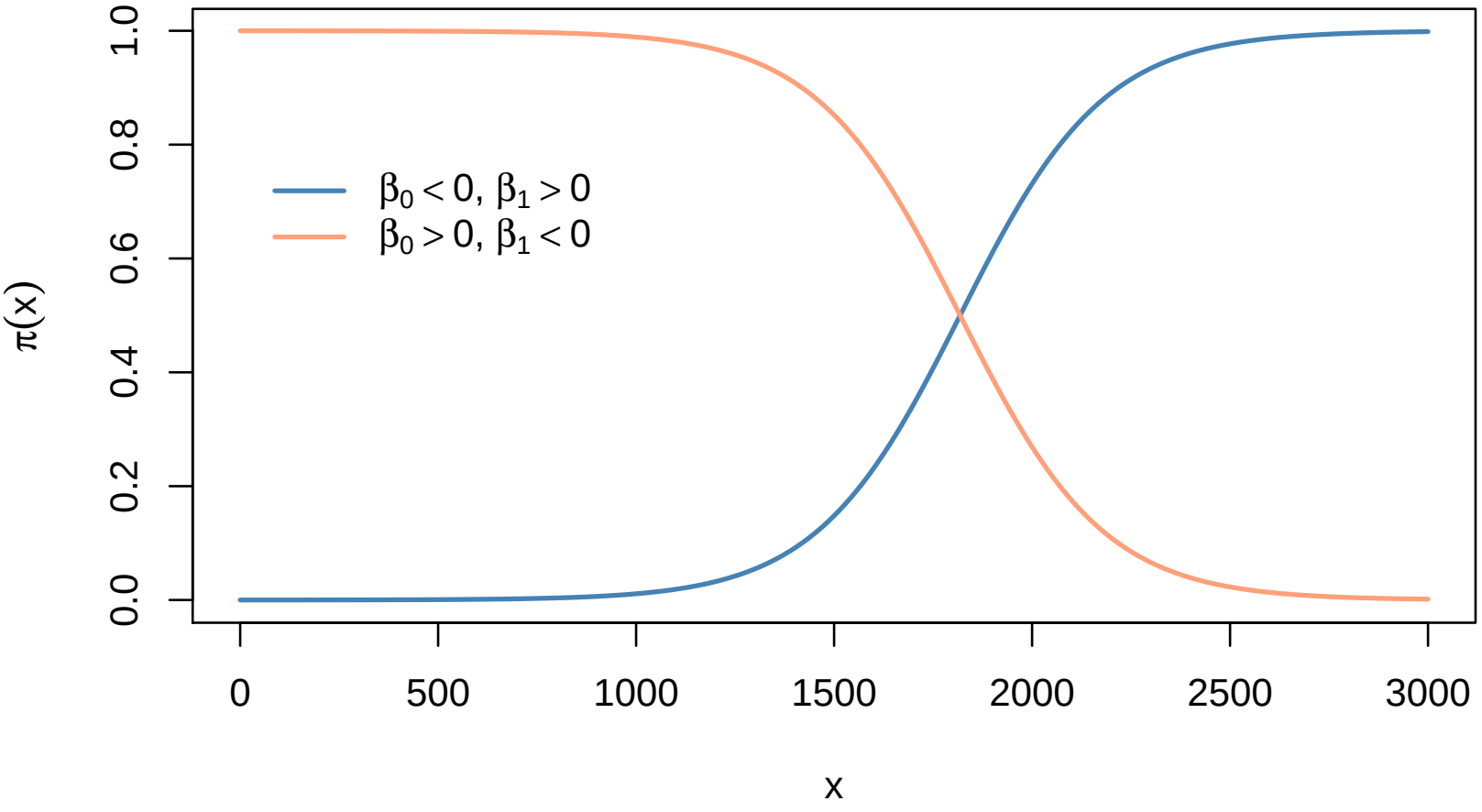
	Dependent variable:	
	bweight	
	(1)	(2)
gestwks	196.973***	192.252***
	(8.788)	(8.966)
matage		−0.766
		(5.206)
factor(hyp)1		−144.223**
		(58.997)
Constant	−4,489.140***	−4,259.381***
	(340.899)	(390.643)
Observations	490	490
R^2	0.507	0.513
Adjusted R^2	0.506	0.510
Residual Std. Error	449.724 (df = 488)	447.903 (df = 486)
F Statistic	502.364*** (df = 1; 488)	170.811*** (df = 3; 486)
Note:		*p<0.1; **p<0.05; ***p<0.01

O modelo de regressão logístico

- Seja Y a variável resposta que assume valores: 0, se o evento de interesse NÃO ocorre; 1, se o evento de interesse ocorre.
- Seja $\pi(x) = \text{Pr}(Y = 1|x)$, ou seja, o $\text{risco}(x)$.
- O modelo de regressão logística utiliza a **função logística** para modelar a associação entre x e $\pi(x)$:

$$\pi(x) = \frac{\exp\{\beta_0 + \beta_1 x\}}{1 + \exp\{\beta_0 + \beta_1 x\}}.$$

O modelo de regressão logístico



O modelo de regressão logístico

A quantidade $\pi(x)/[1-\pi(x)]$ é chamada **chance** (*odds*). Após algumas manipulações algébricas, temos:

$$\frac{\pi(x)}{1-\pi(x)} = \exp\{\beta_0 + \beta_1 x\}.$$

A função **logito** (*logit*) é o logaritmo da chance. Desta forma, temos:

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x.$$

O modelo de regressão logístico

Note que para dois valores de x (x_0 e x_1), podemos definir a **razão de chances** (*odds ratio*) $OR = \frac{\pi(x_1)/[1-\pi(x_1)]}{\pi(x_0)/[1-\pi(x_0)]}$. Pela especificação do modelo de regressão logístico, temos:

$$OR = e^{\beta_1(x_1-x_0)}$$

- A estimação dos coeficientes é feita geralmente pelo método da **máxima verossimilhança**.
- **Interpretação:** $\hat{OR} = e^{\hat{\beta}_1(x_1-x_0)}$.
- **Predição:** $\hat{\pi}(x) = \frac{\exp\{\hat{\beta}_0+\hat{\beta}_1 x\}}{1+\exp\{\hat{\beta}_0+\hat{\beta}_1 x\}}$.

O modelo de regressão logístico no R

- Para ajustarmos um modelo de regressão logístico no R utilizaremos a função `glm()` **(modelos lineares generalizados)**.

```
mod.glm <- glm(lowbw ~ gestwks,
               data = births,
               family = binomial(link = "logit"))
```

- A maior parte das funções aplicadas a objetos `lm` se aplicam a objetos `glm`.

O modelo de regressão logístico no R

```
coef(mod.glm)
```

```
## (Intercept)     gestwks
##  31.8476573  -0.8964603
```

```
exp(coef(mod.glm))
```

```
##  (Intercept)      gestwks
## 6.780502e+13 4.080114e-01
```

```
library(MASS)
exp(confint(mod.glm, "gestwks", level = 0.95))
```

```
##     2.5 %    97.5 %
## 0.3242986 0.4967603
```

O modelo de regressão logístico no R

```
library(oddsratio)
or_glm(data = births,
       model = mod.glm,
       incr = list(gestwks = 1))
```

```
##   predictor oddsratio CI_low (2.5 %) CI_high (97.5 %) increment
## 1   gestwks     0.408          0.324             0.497         1
```

O modelo de regressão de Poisson

Seja Y uma variável aleatória com distribuição de Poisson com parâmetro λ. Então,

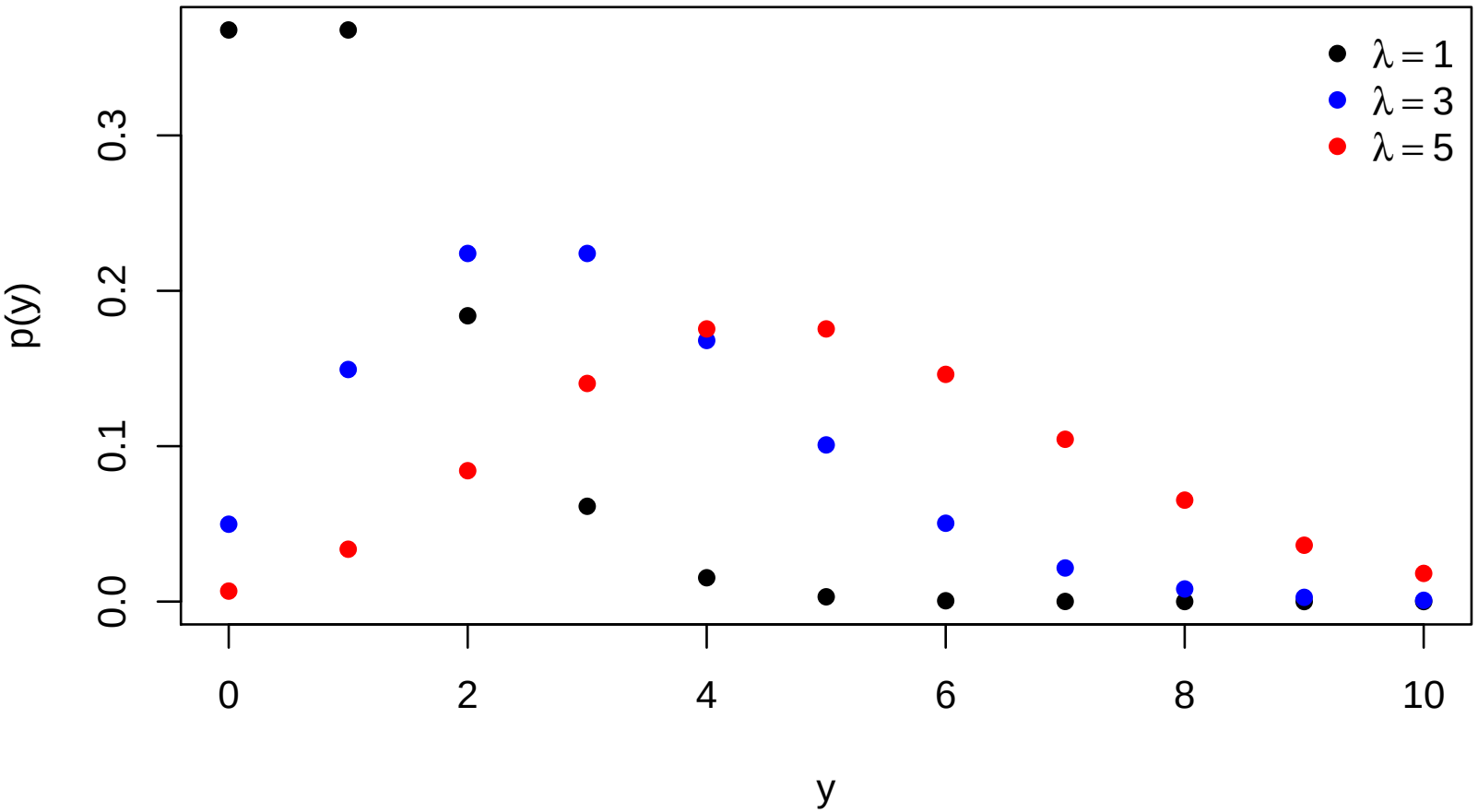
$$p(y) = \Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, y = 0, 1, 2, \ldots, \lambda > 0.$$

Note que

$$\begin{aligned} E[Y] &= \sum_{y=0}^{\infty} y \times p(y) = \sum_{y=0}^{\infty} y \times \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \lambda e^{-\lambda} \sum_{y=1}^{\infty} \frac{\lambda^{y-1}}{(y-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned}$$

De forma similar, temos que $E[Y^2] = \lambda + \lambda^2$, e portanto $Var[Y] = E[Y^2] - (E[Y])^2 = \lambda$ $(E[Y] = Var[Y])$.

O modelo de regressão de Poisson



O modelo de regressão de Poisson

Comentário: O modelo de Poisson descreve a incerteza com respeito a uma variável de contagem.

Exemplos de variáveis de contagem:

- Número de acidentes de carro;
- Número de mortes por determinada doença;
- Número de casos de dengue em uma certa região.

O modelo de regressão de Poisson

Age	Smoking	Deaths	PersonYears
35 to 44	smoker	32	52407
45 to 54	smoker	104	43248
55 to 64	smoker	206	28612
65 to 74	smoker	186	12663
75 to 84	smoker	102	5317
35 to 44	non-smoker	2	18790
45 to 54	non-smoker	12	10673
55 to 64	non-smoker	28	5710
65 to 74	non-smoker	28	2585
75 to 84	non-smoker	31	1462

O modelo de regressão de Poisson

Assumimos que as observações (dados) foram geradas de um certo modelo probabilístico, e então estimamos, através dos dados, os parâmetros desse modelo.

Exemplo: assumindo que o modelo de Poisson é um modelo adequado para o número de mortes nos dados da Tabela anterior, queremos estimar $\lambda = E[Deaths]$.

- $\hat{\lambda} = \bar{Deaths} = 73.1$.

O modelo de regressão de Poisson

Age	non-smoker	smoker
35 to 44	2	32
45 to 54	12	104
55 to 64	28	206
65 to 74	28	186
75 to 84	31	102

O modelo de regressão de Poisson

- $\hat{\lambda}_{smk} = 126$; $\hat{\lambda}_{n-smk} = 20.2$;
- $\hat{\lambda}_{35-44} = 17$; $\hat{\lambda}_{45-54} = 58$; $\hat{\lambda}_{55-64} = 117$; $\hat{\lambda}_{65-74} = 107$; $\hat{\lambda}_{75-84} = 66.5$.

O modelo de regressão de Poisson

Comentário: o número médio de mortes parece mudar de acordo com o *estrato*.

- Como considerar o efeitos simultâneos de diversas variáveis de exposição no número médio de mortes?

O modelo de regressão de Poisson

- Iremos assumir que os dados observados foram gerados de um determinado modelo; neste caso, o modelo de Poisson: $d_{jk} \overset{indep.}{\sim} Poisson(\lambda_{jk}), j = 1, \ldots, J, k = 1, \ldots, K$.
- **Comentário:** o modelo de Poisson é um modelo da classe dos *modelos lineares generalizados* (MLG).
- Iremos assumir que λ_{jk} varia conforme um componente sistemático:

Modelos aditivo: $\lambda_{jk} = \alpha_j + \beta_k$;

Modelos mutiplicativo: $\lambda_{jk} = \theta_j \psi_k$.

O modelo de regressão de Poisson

Note que tomando logaritmo dos dois lados da equação acima, temos

$$\log(\lambda_{jk}) = \log(\theta_j) + \log(\psi_k).$$

- $\alpha_j = \log(\theta_j), \beta_k = \log(\psi_k)$; ψ_k representa o risco relativo (razão de taxas).

Estes são os dois modelos mais comuns usados para descrever a relação entre efeitos de exposição e efeitos de idade e outros fatores de "perturbação".

O modelo de regressão de Poisson

No caso em que $E[d_{jk}] = \lambda_{jk} n_{jk}$, em que n_{jk} representa o número de pessoas-ano, temos

$$\begin{aligned} \log(E[d_{jk}]) &= \log(\lambda_{jk}) + \log(n_{jk}), \\ &= \alpha_j + \theta_k + \log(n_{jk}). \end{aligned}$$

- n_{jk} é dita um *offset*.
- Note que $\lambda_{jk} = \exp\{\alpha_j + \theta_k\}$.

O modelo de regressão de Poisson

Geralmente, o **método da máxima verossimilhança** é utilizado para se fazer a estimação dos parâmetros do modelo.

- **P1:** o que é a função de verossimilhança?
- **P2:** como obter o máximo da função de verossimilhança?

O modelo de regressão de Poisson

- Temos interesse em estimar as taxas λ_{jk}, porém o ajuste do modelo nos fornece estimativas dos parâmetros α_j e β_k.
- Pela **propriedade da invariância** dos estimadores de máxima verossimilhança, temos $\hat{\lambda_{jk}} = \exp\{\hat{\alpha}_j + \hat{\theta}_k\}$.
- **Razões de taxas** são uma maneira comum de descrever os coeficientes do modelo de regressão de Poisson, em uma escala mais interpretável

$$\frac{\lambda_{jk}}{\lambda_{jk'}} = \frac{\exp\{\alpha_j + \beta_k\}}{\exp\{\alpha_j + \beta'_k\}} = \exp\{\beta_k - \beta_{k'}\}.$$

O modelo de regressão de Poisson no R

- O modelo de regressão de Poisson também é um modelo da classe dos modelos lineares generalizados.
  - Logo, para ajustarmos um modelo de regressão de Poisson novamente utilizaremos a função `glm()`.

```
mod.Poi <- glm(Deaths ~ Age + Smoking,
               offset = log(PersonYears/1000),
               data = dados, family = poisson)
```

Próximos passos