MLG para respostas positivas (assimétricas): parte 2

Prof. Caio Azevedo

Exemplo 6: pagamentos de seguros na Austrália

- Vamos considerar parte dos dados descritos em de Jong e Heller (2008, pgs. 14-15) referentes aos valores pagos de seguros individuais (em dólares australianos) por danos com acidentes pessoais no período de julho de 1989 a junho de 1999 (veja também Paula 2024).
- As análises serão restritas ao período de janeiro de 1998 a junho de 1999, correspondendo a um total de 769 seguros pagos.
- Variável reposta: valor pago ao segurado - *vpago*.

Exemplo 6: pagamentos de seguros na Austrália

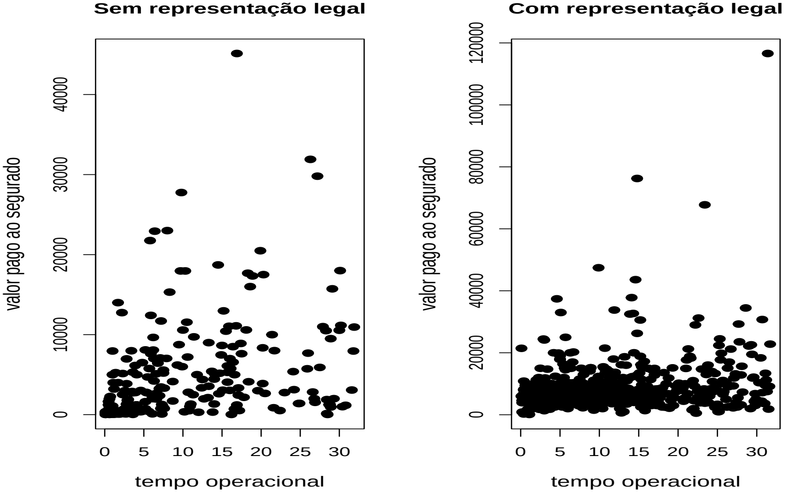
- Variáveis explicativas:
  - *legrep*, representação legal - (0: não, 1: sim) e *optime* - tempo operacional para pagamento do seguro.
  - Essa última variável assume valores no intervalo (0, 100) e por exemplo um valor 23 significa que 23% dos seguros foram pagos antes do seguro em análise.
  - Como estamos considerando apenas parte dos dados (referentes aos últimos 18 meses), os valores de *optime* variam de 0,1 a 31,9.
- Veja o arquivo insurance.prn.

Banco de dados

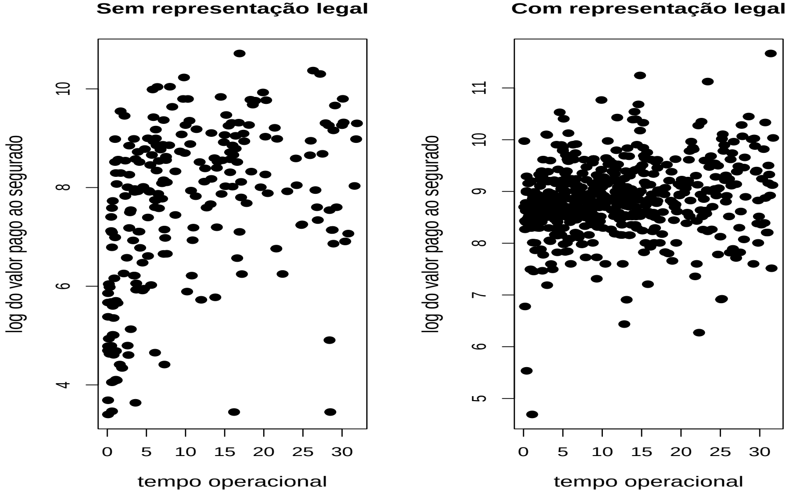
Observação	vpago	optime	legrep
1	119,60	0,10	0
2	290,00	0,10	0
3	30,00	0,10	0
4	40,00	0,10	0
5	21450,00	0,10	1
⋮	⋮	⋮	⋮
769	10947,00	31,90	0

Vamos considerar que cada observação corresponde à uma ocorrência (observações independentes).

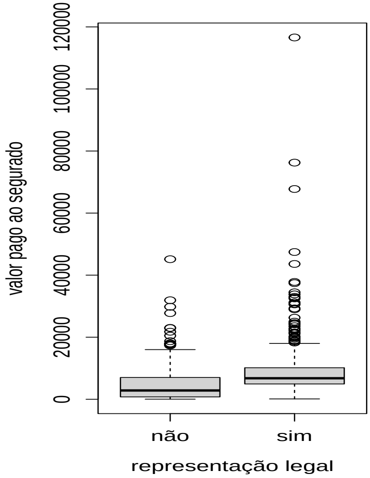
Gráficos de dispersão



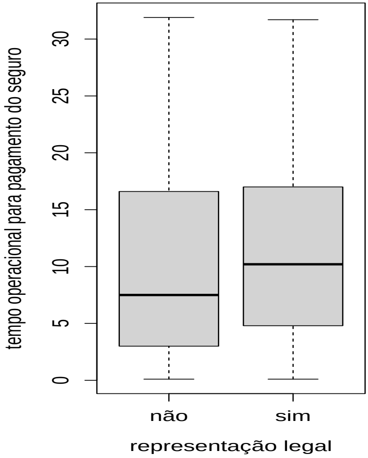
Gráficos de dispersão (log da resposta)



Boxplots



valor pago ao segurado
0
20000
40000
60000
80000
100000
120000
não
sim
representação legal



tempo operacional para pagamento do seguro
0
5
10
15
20
25
30
não
sim
representação legal

Medidas resumo

	vpago		optime	
Medida-resumo	nao	sim	nao	sim
média	5069,27	8996,41	10,74	11,86
DP	6338,01	8790,71	9,18	8,53
Var.	40170367,59	77276620,67	84,26	72,69
CV(%)	125,02	97,71	85,45	71,90
CA	2,51	5,58	0,73	0,63
curt	12,05	54,31	12,05	54,31
Min.	30,0	109,00	0,10	0,10
Max.	45132,03	116586,72	31,90	31,70

Modelo

$$
\begin{aligned}
Y_{ij} &\overset{ind.}{\sim} \text{gama}(\mu_{ij}, \phi), \\
i &= 1, 2\,(\text{legpre}, 1 - \text{não}, 2 - \text{sim});\,; j = 1, 2, ..., n_i \\
& \quad (\text{observação}, n_1 = 227, n_2 = 542) \\
\ln(\mu_{ij}) &= \alpha + \alpha_i + (\beta + \beta_i)(x_{ij} - 0,1) + (\gamma + \gamma_i)(x_{ij} - 0,1)^2, \\
\alpha_1 &= \beta_1 = \gamma_1 = 0 \\
\mathcal{E}(Y_{ij}) = \mu_{ij} &\quad ; \quad \mathcal{V}(Y_{ij}) = \frac{\mu_{ij}^2}{\phi}
\end{aligned}
$$

Modelo

- Y_{ij}: valor do seguro pago na j-ésima ocorrência sob a i-ésima representação legal.
- x_{ij}: tempo operacional para pagamento do seguro na j-ésima ocorrência sob a i-ésima representação legal.
- e^{α} : valor médio do seguro pago sem representação legal com o tempo operacional igual à 0,1.
- $e^{\alpha+\alpha_2}$: valor médio do seguro pago com representação legal com o tempo operacional igual à 0,1.

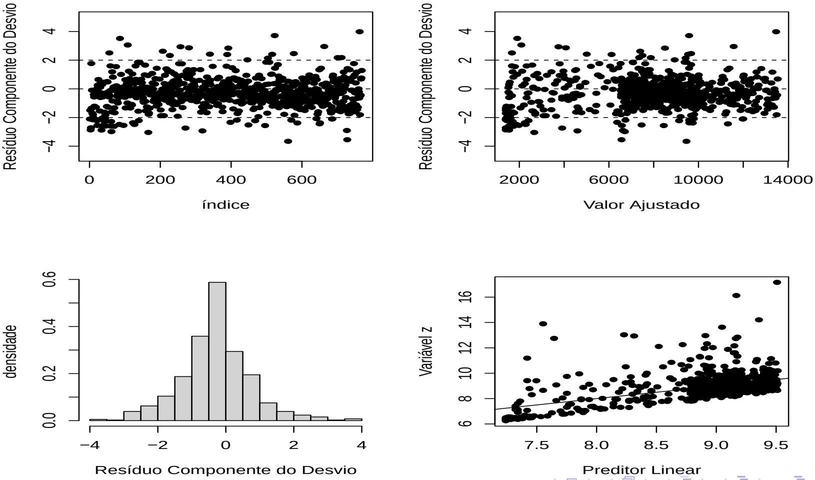
Modelo

- $-\dfrac{\beta}{2\gamma} + 0,1$: valor do tempo operacional para o qual o log do valor do esperado (consequentemente o próprio valor esperado) esperado do seguro pago é máximo/mínimo, sem representação legal.
- $-\dfrac{\beta + \beta_2}{2(\gamma + \gamma_2)} + 0,1$: valor do tempo operacional para o qual o log do valor esperado (consequentemente o próprio valor esperado) esperado do seguro pago é máximo/mínimo, com representação legal.
- Exercício: suponha que $\gamma = \gamma_2 = 0$. Interprete e^{β} e $e^{\beta+\beta_2}$.

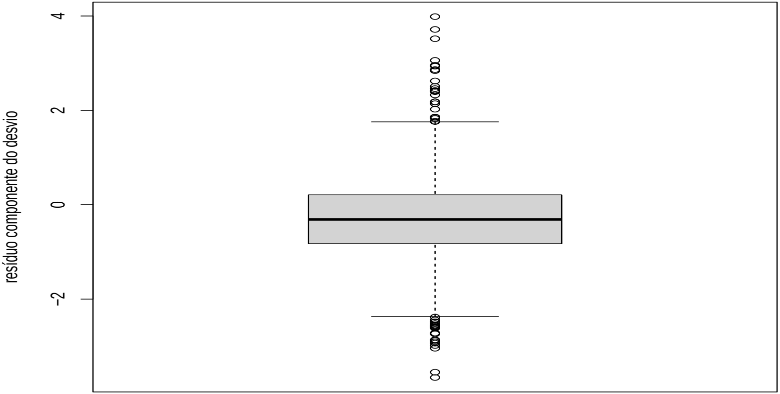
Modelo

- $D(\boldsymbol{y};\widetilde{\boldsymbol{\mu}}) = 856,25 (p = 0,0104)$ (considerando-se a aproximação pela $\chi^2_{(763)}$ adequada), o que indica que o modelo não se ajustou de modo satisfatório aos dados.
- Neste caso, a estimativa de ϕ (veja adiante) é pequena e a aproximação pela χ^2 pode não ser razoável.

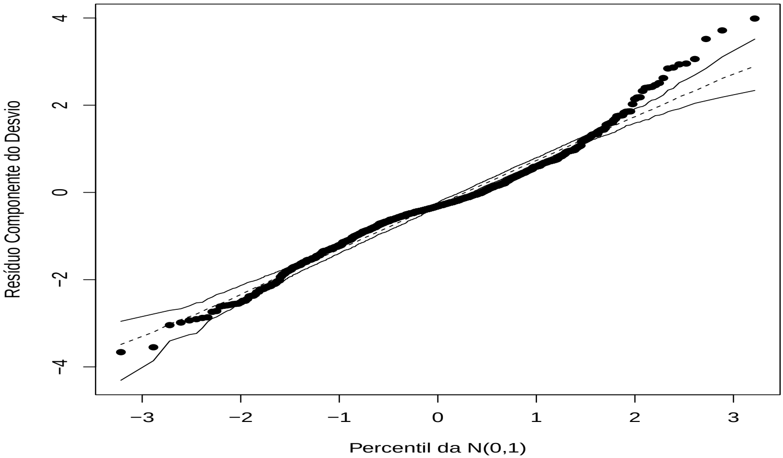
Gráficos de diagnóstico



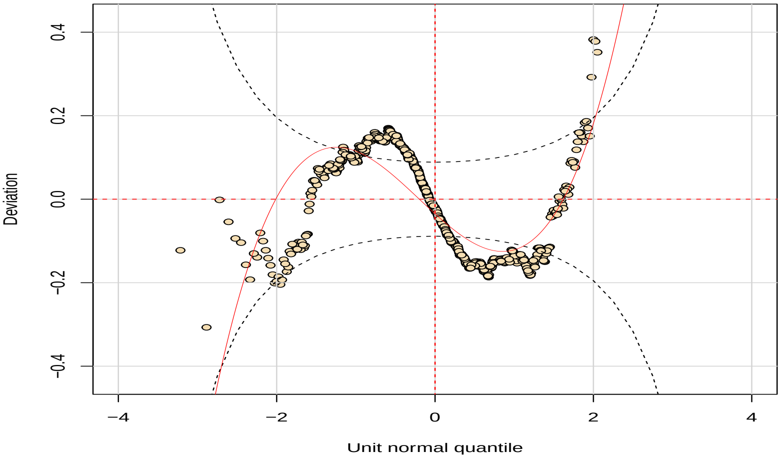
Boxplot para os resíduos



Envelope para os resíduos



Worm plot para os resíduos



Comentários

- O modelo não se ajustou bem aos dados.
- O RCD apresenta caudas pesadas (com vários candidatos à outliers) e uma aparente heterocedasticidade.
- Além disso, o gráfico da variável z $\times$ o preditor linear apresenta muitos valores "discrepantes" (o que pode ser devido à necessidade de utilização de um predito não paramétrico).
- O worm plot endossa o mal ajuste do modelo.

Estimativas dos parâmetros

Par.	Estimativa	EP	IC(95%)	Estat. Z	p-valor
α	7,241	0,112	[7,021 ; 7,462]	64,440	$<$0,0001
β	0,203	0,021	[0,162 ; 0,244]	9,679	$<$0,0001
γ	-0,005	0,001	[-0,007 ; -0,004]	-7,313	$<$0,0001
α_2	1,542	0,144	[1,260 ; 1,823]	10,726	$<$0,0001
β_2	-0,174	0,026	[-0,225 ; -0,124]	-6,732	$<$0,0001
γ_2	0,005	0,001	[0,003 ; 0,007]	5,774	$<$0,0001
ϕ	1,407	0,065	[1,280 ; 1,535]	-	-

Estimativas dos parâmetros

- O modelo gama é preferível ao modelo exponencial.
- Todos os parâmetros de regressão significativos.
- Curvas:

  (sem representação legal)

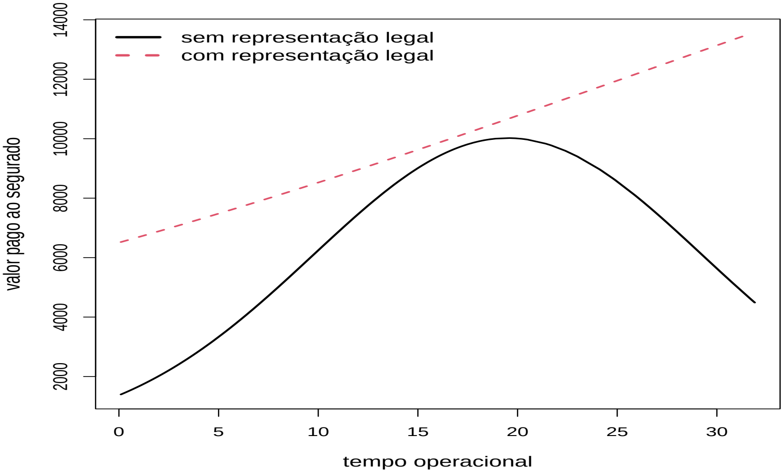
$$\ln(\widetilde{\mu}_{1j}) = 7,241 + 0,203(x_{1j} - 0,1) - 0,005(x_{1j} - 0,1)^2.$$

  (com representação legal)

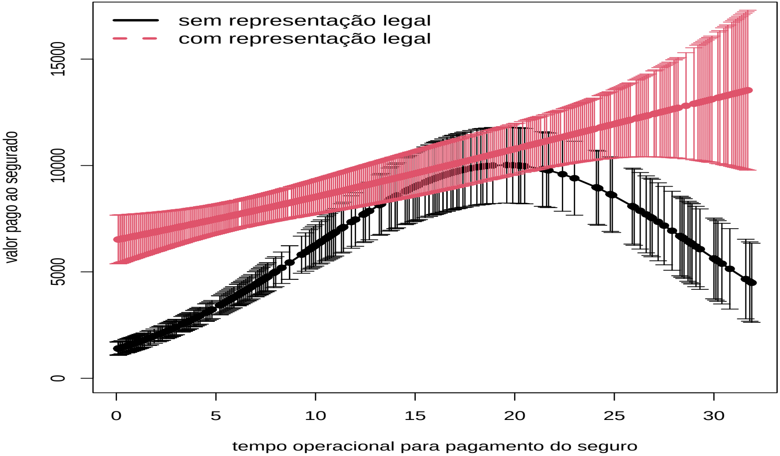
$$\ln(\widetilde{\mu}_{2j}) = 8,783 + 0,029(x_{2j} - 0,1) - 0,00018(x_{2j} - 0,1)^2.$$

- A curva do primeiro grupo (sem representação legal) apresenta um ponto de máximo, enquanto que a do segundo é, essencialmente, uma reta.

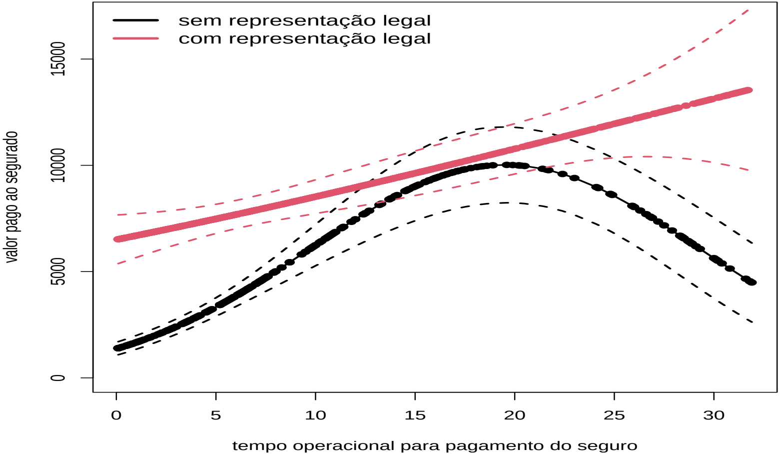
Médias preditas pelo modelo



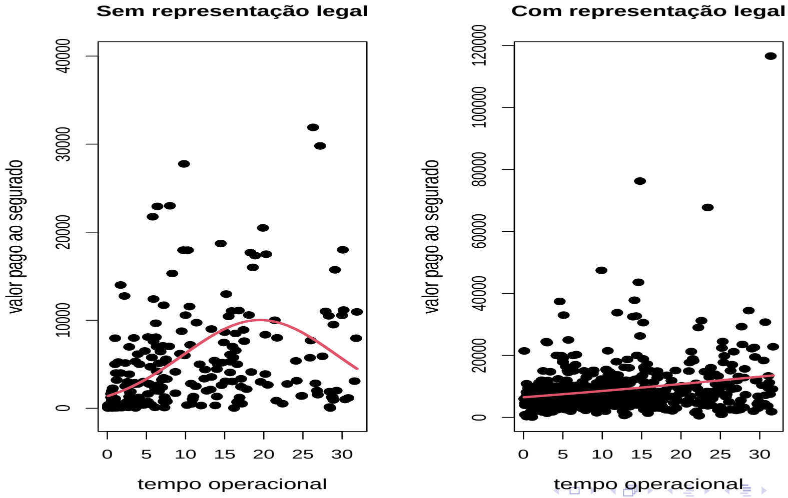
Médias preditas pelo modelo com IC's assintótico



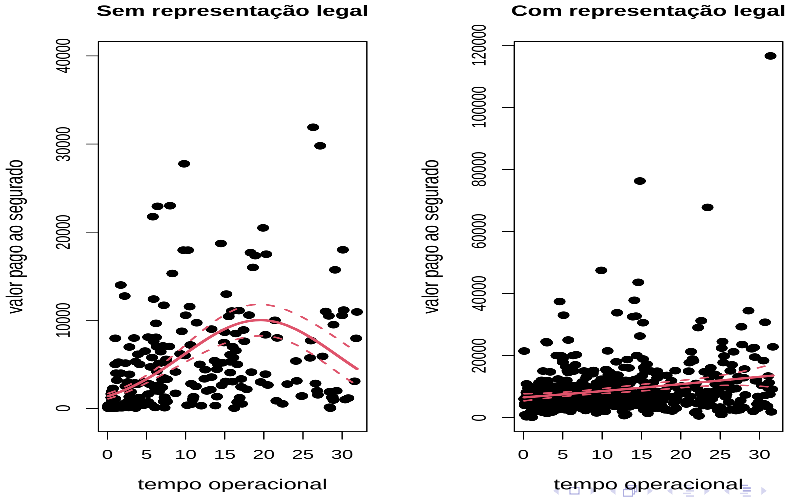
Médias preditas pelo modelo com IC's assintóticos



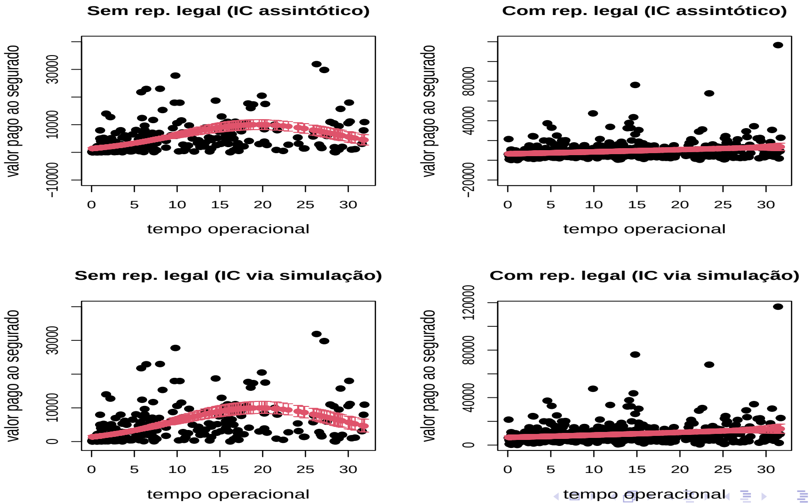
Médias preditas pelo modelo



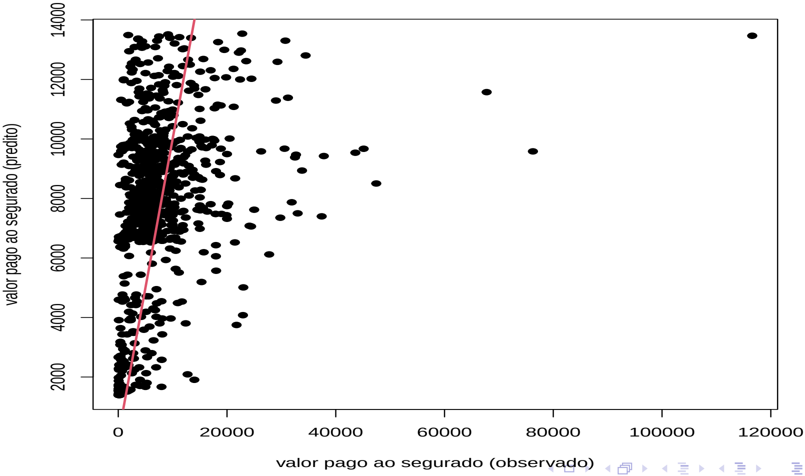
Médias preditas pelo modelo com IC's assintóticos



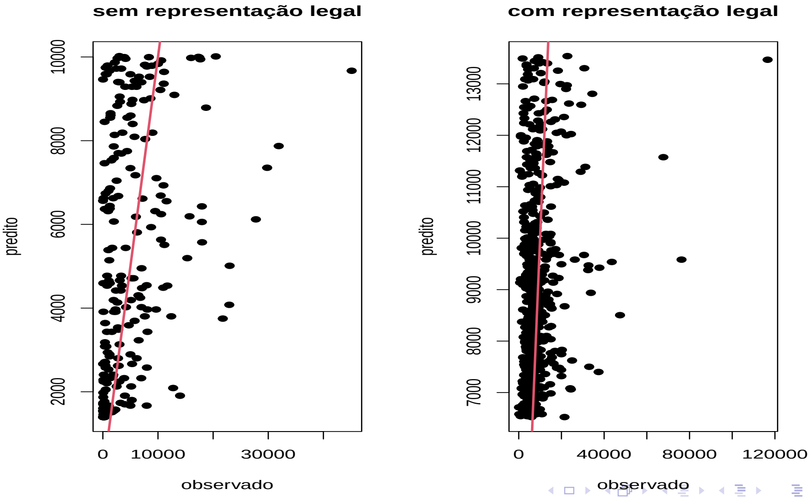
Médias preditas pelo modelo com intervalos de confiança



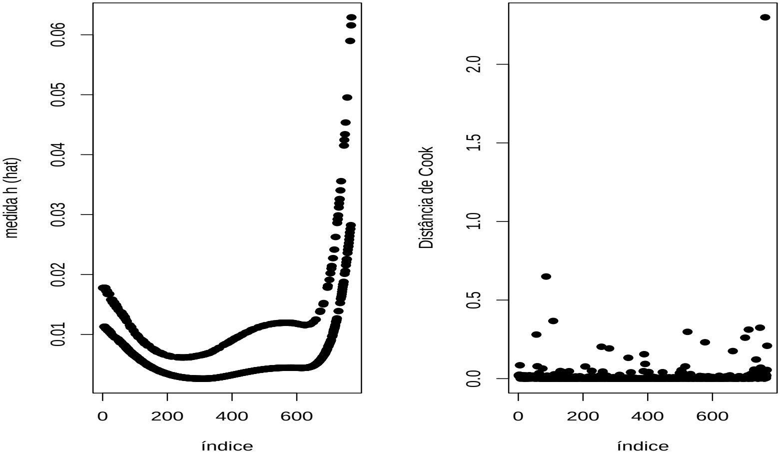
Valores pagos observados e preditos



Valores pagos observados e preditos



Pontos alavanca e Distância de Cook



Análise de sensibilidade

Parâmetro (α)

Observações	Est.	EP	IC(95%)	z value	p-valor
todos	7,241	0,112	[7,021;7,462]	64,440	$<$0,0001
-763	7,241	0,111	[7,023;7,460]	65,027	$<$0,0001
-765	7,245	0,113	[7,024;7,465]	64,312	$<$0,0001
-768	7,234	0,113	[7,013;7,454]	64,214	$<$0,0001
-769	7,226	0,113	[7,005;7,446]	64,174	$<$0,0001

Análise de sensibilidade

Parâmetro (β)

Observações	Est.	EP	IC(95%)	z value	p-valor
todos	1,542	0,144	[1,260;1,823]	10,726	$<$0,0001
-763	1,496	0,142	[1,217;1,776]	10,500	$<$0,0001
-765	1,538	0,144	[1,256;1,821]	10,686	$<$0,0001
-768	1,549	0,144	[1,267;1,832]	10,763	$<$0,0001
-769	1,557	0,144	[1,276;1,839]	10,825	$<$0,0001

Análise de sensibilidade

Parâmetro (γ)

Observações	Est.	EP	IC(95%)	z value	p-valor
todos	0,203	0,021	[0,162;0,244]	9,679	$<$0,0001
-763	0,203	0,021	[0,162;0,244]	9,767	$<$0,0001
-765	0,202	0,021	[0,161;0,244]	9,544	$<$0,0001
-768	0,206	0,021	[0,164;0,247]	9,707	$<$0,0001
-769	0,208	0,021	[0,167;0,250]	9,833	$<$0,0001

Análise de sensibilidade

Parâmetro (α_2)

Observações	Est.	EP	IC(95%)	z value	p-valor
todos	-0,005	0,001	[-0,007;-0,004]	-7,313	$<$0,0001
-763	-0,005	0,001	[-0,007;-0,004]	-7,380	$<$0,0001
-765	-0,005	0,001	[-0,007;-0,004]	-7,123	$<$0,0001
-768	-0,005	0,001	[-0,007;-0,004]	-7,341	$<$0,0001
-769	-0,005	0,001	[-0,007;-0,004]	-7,503	$<$0,0001

Análise de sensibilidade

Parâmetro (β_2)

Observações	Est.	EP	IC(95%)	z value	p-valor
todos	-0,174	0,026	[-0,225;-0,124]	-6,732	$<$0,0001
-763	-0,161	0,026	[-0,211;-0,111]	-6,266	$<$0,0001
-765	-0,173	0,026	[-0,224;-0,122]	-6,652	$<$0,0001
-768	-0,177	0,026	[-0,228;-0,126]	-6,787	$<$0,0001
-769	-0,179	0,026	[-0,230;-0,128]	-6,890	$<$0,0001

Análise de sensibilidade

Parâmetro (γ_2)

Observações	Est.	EP	IC(95%)	z value	p-valor
todos	0,005	0,001	[0,003;0,007]	5,774	$<$0,0001
-763	0,004	0,001	[0,003;0,006]	5,118	$<$0,0001
-765	0,005	0,001	[0,003;0,007]	5,653	$<$0,0001
-768	0,005	0,001	[0,003;0,007]	5,834	$<$0,0001
-769	0,005	0,001	[0,004;0,007]	5,968	$<$0,0001

Conclusões

- Supondo que o modelo tivesse se ajustado adequadamente, poderíamos concluir que:
  - Para o grupo com representação é esperado que quanto maior o tempo operacional para pagamento do seguro, maior será o valor pago, enquanto que para o grupo sem representação há um valor do tempo operacional para o qual o valor é máximo, ou seja, 19,51 [17,67;21,35].
  - Exercício: utilizar o método Delta para encontrar o intervalo de confiança para $-\dfrac{\boldsymbol{\beta}}{2\gamma} + 0,1$.
  - Essencialmente, a única observação que influenciou os resultados inferenciais foi a de $\#$ 763, embora a influencia não aparenta ser estatisticamente significativa.

Conclusões

- O valor do seguro pago, para o tempo operacional máximo, é da ordem de: 10.021,04 [8.326,24 ; 11.805,85].
- Exercício: utilizar o método Delta para encontrar o intervalo de confiança para o valor máximo do seguro pago.
- Observação. Para modelar de forma mais apropriada possível a influência do tempo operacional no valor pago sugere-se a utilização de polinômios fracionários ou de uma estrutura não paramétrica.