

MAE526 - Tópicos de Regressão

Fernando Henrique Ferraz Pereira da Rosa
Matheus Moreira Costa
Vagner Aparecido Pedro Júnior

16 de junho de 2005

Lista 3

1. Na tabela abaixo são apresentados os resultados de um experimento em que a resistência (em horas) de um determinado tipo de vidro foi avaliada segundo quatro níveis de voltagem (em kilovolts) e duas temperaturas (em graus Celsius). Esses dados estão também disponíveis no arquivo `vidros.dat`. Na primeira coluna do arquivo tem-se o tempo de resistência, na segunda coluna a voltagem (1: 200kV, 2: 250kV, 3: 300kV e 4: 350kV) e na terceira coluna a temperatura (1: 170 °C e 2: 180 °C). Seja Y_{ijk} o tempo de resistência da k -ésima amostra de vidro submetida à i -ésima temperatura e à j -ésima voltagem. Supor que $Y_{ijk} \sim G(\mu_{ij}, \phi)$. O interesse é comparar as médias μ_{ij} , $i = 1, 2$ e $j = 2, 3, 4$. Propor uma reparametrização tipo casela de referência em que $\mu_{11} = \alpha$, $\mu_{1j} = \alpha + \beta_j$, $\mu_{21} = \alpha + \gamma$ e $\mu_{2j} = \alpha + \gamma + \beta_j$, $j = 2, 3, 4$.

Temperatura (°C)	Voltagem (kV)			
	200	250	300	350
170	439	572	315	258
	904	690	315	258
	1092	904	439	347
	1105	1090	528	588
180	959	216	241	241
	1065	315	315	241
	1065	455	332	435
	1087	473	380	455

Procure responder de que forma os níveis de voltagem e temperatura afetam o tempo médio de resistência dos vidros. Faça também uma análise de diagnóstico.

Lemos o conjunto de dados e estabelecemos a parametrização casela de referência através dos comandos:

```

> vidros <- read.table("dados/vidros.dat", col.names = c("resistencia",
+ "voltagem", "temperatura"))
> vidros$voltagem <- factor(vidros$voltagem)
> vidros$temperatura <- factor(vidros$temperatura)
> vidros$voltagem <- C(vidros$voltagem, treatment)
> vidros$temperatura <- C(vidros$temperatura, treatment)

```

Ajustamos o modelo através do comando:

```

> mod1 <- glm(resistencia ~ temperatura + voltagem,
+ family = Gamma(link = identity), data = vidros)
> summary(mod1)

```

Call:

```

glm(formula = resistencia ~ temperatura + voltagem, family = Gamma(link = identity),
    data = vidros)

```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.7544	-0.2425	0.0117	0.1529	0.6356

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1039.9	122.5	8.49	4.2e-09 ***
temperatura2	-117.8	56.4	-2.09	0.046 *
voltagem2	-426.5	135.6	-3.14	0.004 **
voltagem3	-608.8	126.2	-4.82	4.9e-05 ***
voltagem4	-612.9	126.0	-4.86	4.4e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.118)

Null deviance: 9.4487 on 31 degrees of freedom
Residual deviance: 3.4306 on 27 degrees of freedom
AIC: 428.5

Number of Fisher Scoring iterations: 9

Notemos que, de acordo com a parametrização do enunciado, o intercepto é α , $\text{temperatura2} = \gamma$ e $\text{voltagem}_j = \beta_j$, $j = 2, 3, 4$. Observando os coeficientes ajustados podemos ver que tanto temperatura quanto voltagem têm efeito na resposta, e que quanto maior a temperatura e a voltagem, menor é o tempo médio de resistência dos vidros. Em particular, temos

que o tempo médio de resistência para um tipo de vidro submetido à temperatura de 170 °C e à voltagem de 200kV, é de 1039.9. A mudança na temperatura de 170 °C para 180 °C implica em um decréscimo de 117.8 no tempo médio, fixada a voltagem. Temos também que a mudança de voltagem de 200kV para 250kV, implica em um decréscimo na média de 462.5. Analogamente o uso de voltagens de 300kV e 350kV implica em decréscimos na média de 608.8 e 612.9, respectivamente.

Na Figura 1 temos os gráficos de diagnóstico do modelo ajustado. Não há indícios de fuga das suposições do modelo. Em particular temos que no gráfico do envelope simulado todos os pontos ficam na banda de confiança e nenhum ponto tem influência desproporcional no modelo.

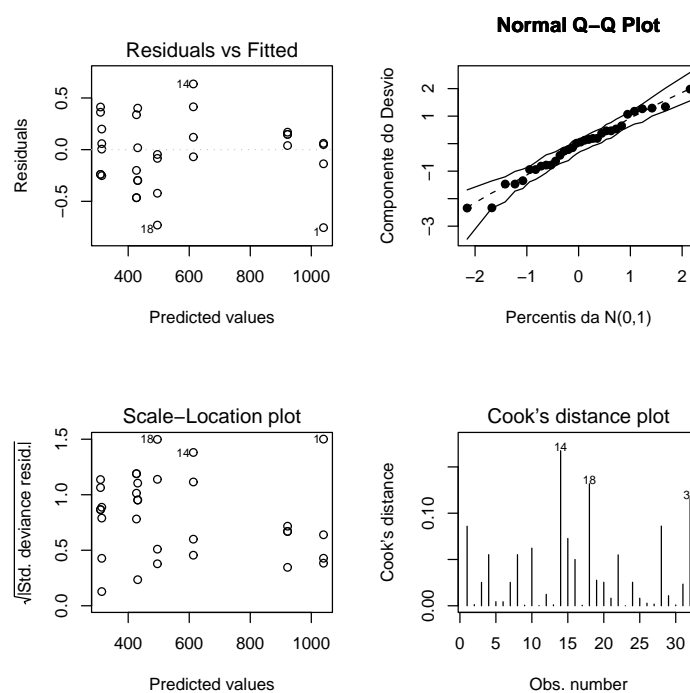


Figura 1: Gráficos de diagnóstico para o modelo ajustado

2. A tabela abaixo apresenta o resultado de uma pesquisa em que 1008 pessoas receberam duas marcas de detergente, X e M, e posteriormente responderam às seguintes perguntas: maciez da água (leve, média ou forte); uso anterior do detergente M (sim ou não); temperatura da água (alta ou baixa); preferência (marca X ou marca M).

Temperatura	Uso de M	Preferência	Maciez		
			Leve	Médio	Forte
Alta	Sim	X	19	23	24
		M	29	47	43
	Não	X	29	33	42
		M	27	23	30
Baixa	Sim	X	57	47	37
		M	49	55	52
	Não	X	63	66	68
		M	53	50	42

Ajustar um modelo log-linear de Poisson para explicar π_{ijkl} , a proporção de indivíduos que responderam, respectivamente, nível de temperatura (i=1 alta, i=2 baixa), uso prévio de M (j=1 sim, j=2 não), preferência (k=1 X, k=2 M) e nível de maciez (l = 1 leve, l = 2 médio, l = 3 forte). Selecionar através do método AIC os efeitos principais significativos. Depois incluir apenas as interações significativas de primeira ordem. Interpretar os resultados e fazer uma análise de diagnóstico.

Começamos ajustando um modelo log-linear de Poisson, com o uso de uma parametrização de efeitos de tratamento.

```
> options(contrasts = c("contr.sum", "contr.poly"))
> det.glm0 <- glm(freq ~ temperatura + uso + preferencia +
+   maciez, family = poisson, data = detergente)
> det.glm0
```

```
Call: glm(formula = freq ~ temperatura + uso + preferencia + maciez, family =
```

```
Coefficients:
(Intercept)  temperatur1      uso1  preferencia1
 3.69921    -0.27455    -0.04368    0.00794
 maciez1      maciez2
-0.02996    0.02378
```

```
Degrees of Freedom: 23 Total (i.e. Null); 18 Residual
Null Deviance:      119
Residual Deviance: 42.9      AIC: 187
```

Fazemos então a seleção dos efeitos principais por AIC, através do comando:

```
> det.glm1 <- stepAIC(det.glm0, trace = 0)
```

Ficando com o modelo:

```
> summary(det.glm1)
```

Call:

```
glm(formula = freq ~ temperatura, family = poisson, data = detergente)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.358	-0.972	-0.245	0.706	2.717

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.7004	0.0327	113.2	<2e-16 ***
temperatura1	-0.2746	0.0327	-8.4	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 118.627 on 23 degrees of freedom
Residual deviance: 45.415 on 22 degrees of freedom
AIC: 181.8

Number of Fisher Scoring iterations: 4

Donde vemos que o único fator principal significativo foi a temperatura. A partir do desvio residual obtemos o nível descritivo:

```
> 1 - pchisq(deviance(det.glm1), det.glm1$df.resid)
```

```
[1] 0.00235
```

o que indica que o modelo não tem um ajuste satisfatório. Consideramos então a possibilidade dos outros efeitos junto com os efeitos de interação. Começamos a partir do modelo completo com interações de primeira ordem, e selecionamos o melhor modelo pelo AIC:

```
> det.glm2 <- update(det.glm0, . ~ . + temperatura:uso +  
+ temperatura:preferencia + temperatura:maciez +  
+ preferencia:uso + preferencia:maciez + uso:maciez)  
> det.glm3 <- stepAIC(det.glm2, trace = 0)
```

E ficamos com o modelo:

```
> summary(det.glm3)
```

```
Call:
glm(formula = freq ~ temperatura + uso + preferencia + maciez +
     temperatura:preferencia + temperatura:maciez + uso:preferencia,
     family = poisson, data = detergente)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4338	-0.3560	-0.0736	0.3975	1.5135

Coefficients:

	Estimate	Std. Error	z value
(Intercept)	3.68270	0.03328	110.66
temperatura1	-0.27826	0.03294	-8.45
uso1	-0.04343	0.03185	-1.36
preferencia1	-0.01659	0.03312	-0.50
maciez1	-0.05907	0.04750	-1.24
maciez2	0.02778	0.04608	0.60
temperatura1:preferencia1	-0.06836	0.03278	-2.09
temperatura1:maciez1	-0.10159	0.04750	-2.14
temperatura1:maciez2	0.00345	0.04608	0.07
uso1:preferencia1	-0.14377	0.03185	-4.51

Pr(>|z|)

(Intercept)	< 2e-16 ***
temperatura1	< 2e-16 ***
uso1	0.173
preferencia1	0.616
maciez1	0.214
maciez2	0.547
temperatura1:preferencia1	0.037 *
temperatura1:maciez1	0.032 *
temperatura1:maciez2	0.940
uso1:preferencia1	6.4e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 118.627 on 23 degrees of freedom
Residual deviance: 11.886 on 14 degrees of freedom
AIC: 164.3

Number of Fisher Scoring iterations: 4

O nível descritivo para o dado desvio residual é dado por:

```
> 1 - pchisq(deviance(det.glm3), det.glm3$df.resid)
```

```
[1] 0.615
```

o que nos leva a considerá-lo satisfatório. Na Figura 2 temos os gráficos de resíduos para esse modelo. Vemos que não há fuga das suposições nem observações desproporcionalmente influentes.

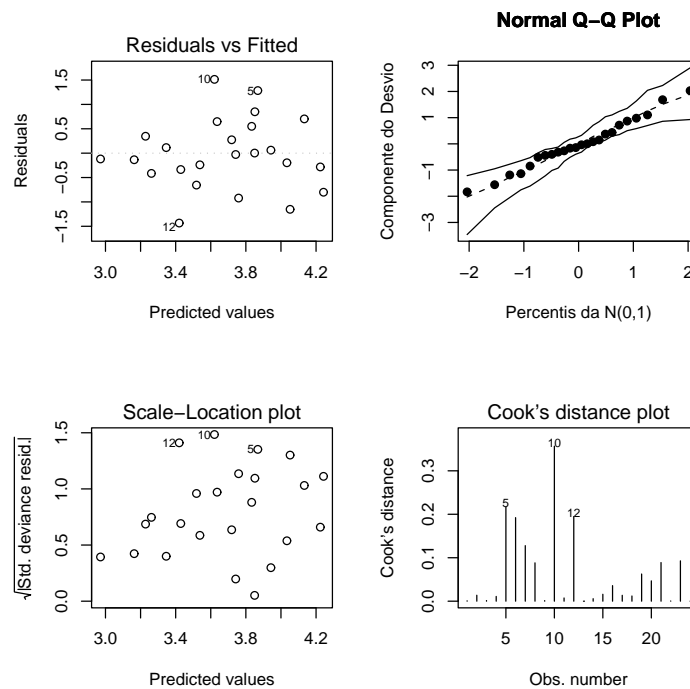


Figura 2: Gráficos de diagnóstico para o modelo ajustado

Ficamos então com o modelo log-linear de Poisson, cuja taxa de ocorrência fica dada por:

$$\log \lambda_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \tau_l + (\alpha\gamma)_{ij} + (\alpha\tau)_{ik} + (\beta\gamma)_{jk},$$

onde μ é um intercepto, α é o efeito da temperatura, β é o efeito do uso anterior de M, γ é a preferência e τ é a maciez. Os outros termos são as respectivas interações entre os efeitos principais. Observando as estimativas dos coeficientes desse modelo vemos que os efeitos significativos

são o efeito principal de temperatura, e as interações entre temperatura e maciez, temperatura e preferência e uso e preferência.

O coeficiente estimado para temperatura (-0.278) indica que fixados os outros fatores, a proporção de pessoas que usam temperatura alta é menor do que aquelas que usam temperatura baixa. O coeficiente de interação entre temperatura e preferência (-0.0684) indica que o comportamento da preferência pela marca X ou M varia de acordo com o nível de temperatura (alta ou baixa). Em particular, sob a temperatura baixa, há uma proporção de preferência maior pela marca X. A interação entre temperatura e maciez (coeficiente -0.102) indica que a percepção de maciez pode variar de acordo com a temperatura. Em particular, fixados os outros fatores, para a temperatura baixa é maior a proporção de maciez leve. Por fim a interação entre uso e preferência (coeficiente -0.144) está indicando que as proporções da preferência pela marca X ou M variam de acordo com o uso anterior de M. Em particular, ter usado M anteriormente implica em uma maior chance de se ter preferência por ele.

3. No arquivo `olhos.dat` são apresentados dados referentes a 78 famílias com pelo menos seis filhos cada uma. Na primeira coluna tem-se a classificação dos olhos dos pais segundo a cor (1: ambos claros, 2: ambos castanhos, 3: ambos escuros, 4: claro e castanho, 5: claro e escuro e 6: castanho e escuro), na segunda coluna a classificação dos olhos dos avós segundo a cor (1: todos claros, 2: todos castanhos, 3: todos escuros, 4: três claros e um castanho, 5: três claros e um escuro, 6: um claro e três castanhos, 7: um escuro e três castanhos, 8: um claro e três escuros, 9: um castanho e três escuros, 10: dois claros e dois castanhos, 11: dois claros e dois escuros, 12: dois castanhos e dois escuros, 13: dois claros, um castanho e um escuro, 14: um claro, dois castanhos e um escuro e 15: um claro, um castanho e dois escuros), na terceira coluna tem-se o número de filhos na família e na última coluna o número de filhos com olhos claros. Seja Y_i o número de filhos com olhos claros pertencentes a i -ésima família. Assuma inicialmente que $Y_i \sim B(n_i, \pi_i), i = 1, \dots, 78$. Ajustar um modelo logístico linear apenas com o fator 'cor dos olhos dos pais'. Construir gráficos de resíduos. Identificar os pontos aberrantes. Quais as mudanças nos resultados com a eliminação desses pontos. Há indícios de superdispersão? Comente.

Na Figura 3 temos um diagrama de pontos da proporção de filhos com olhos claros pelo tipo dos olhos dos pais. Podemos observar como a variância é diferente para os vários grupos e como a proporção de filhos com olhos claros é maior para pais com ambos os olhos claros.

Começamos ajustando um modelo logístico linear, da forma

$$\log \left\{ \frac{\pi(x)}{1 - \pi(x)} \right\} = X\beta,$$

em que $\pi(x)$ é a probabilidade de ter filhos com olhos claros e β um vetor de parâmetros.

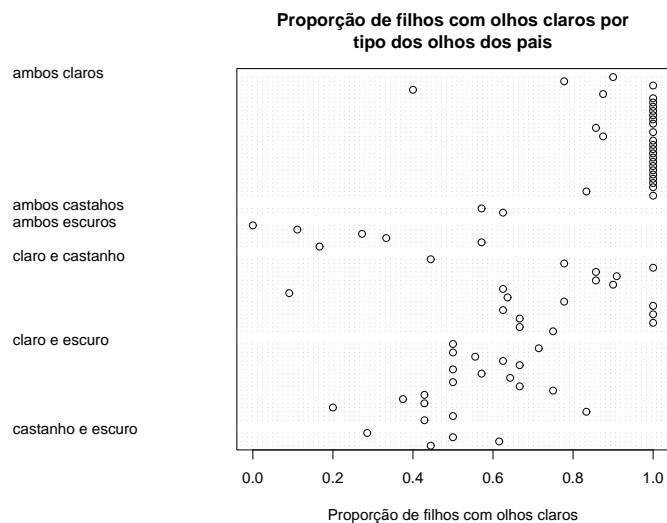


Figura 3: Diagrama de pontos para os dados do exercício 3

Ajustamos esse modelo no R da seguinte maneira:

```
> resp <- cbind(olhos$olhosclaros, olhos$filhos -
+   olhos$olhosclaros)
> logis1 <- glm(resp ~ pais, family = binomial,
+   data = olhos)
> summary(logis1)
```

Call:

```
glm(formula = resp ~ pais, family = binomial, data = olhos)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.651	-0.529	0.477	0.953	2.363

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.539	0.135	4.00	6.4e-05	***
pais1	2.290	0.269	8.50	< 2e-16	***
pais2	-0.134	0.451	-0.30	0.767	
pais3	-1.668	0.314	-5.32	1.1e-07	***
pais4	0.472	0.202	2.34	0.020	*
pais5	-0.364	0.190	-1.91	0.056	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 270.64 on 77 degrees of freedom
Residual deviance: 119.10 on 72 degrees of freedom
AIC: 250.5

Number of Fisher Scoring iterations: 5

O alto valor do desvio residual em relação aos graus de liberdade é indicativo de um ajuste inadequado (nível descritivo: 0.000405). Na Figura 4 temos os gráficos de resíduos do modelo ajustado. Vemos que vários pontos ficaram fora da banda de confiança e que 3 observações (13, 42 e 58) aparentam ter uma influência desproporcional no ajuste.

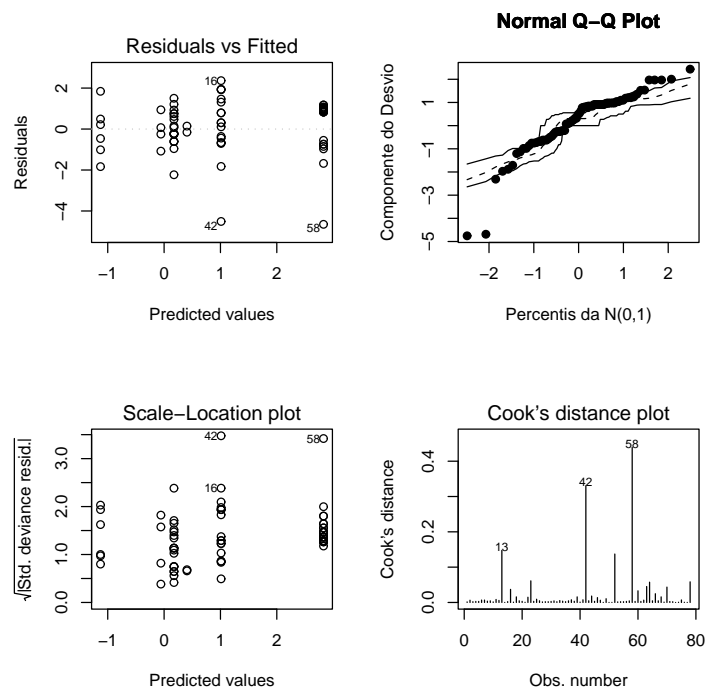


Figura 4: Gráficos de diagnóstico para o modelo ajustado

Para verificar a influência das observações 13, 42 e 58 realizamos o ajuste do modelo sem cada uma delas e depois sem as três. Na Tabela 1 temos

as variações dos parâmetros e do desvio com cada retirada. Vemos que todas têm uma influência desproporcional no modelo. Em particular, os efeitos pais12 e pais4 são drasticamente afetados pelas três. Também o desvio residual tem uma variação desproporcional.

Parâmetro	Observações retiradas			
	13	42	58	13,42 e 58
intercepto	-11	9	19	16
pais1	3	-2	22	22
pais2	-45	35	75	65
pais3	18	3	6	27
pais4	13	49	-21	41
pais5	-16	13	27	24
desvio residual	-4	-19	-21	-43

Tabela 1: Variação em % das estimativas para o modelo logístico linear ajustado no exercício 3, retirando as observações potencialmente influentes

Consideremos o modelo sem as 3 observações, cujo ajuste se encontra abaixo. O nível descritivo do desvio residual é de 0.51, o que é razoável. Nas Figuras 5 temos os gráficos de resíduos para o modelo dessa forma. Notamos que apesar de não haver pontos com uma influência tão grande quanto antes, ainda há alguns pontos caindo fora da banda de confiança. O fato de quase todos os coeficientes estarem significantes, assim como os desvios residuais altos que obtivemos anteriormente, são indícios de superdispersão. O uso de modelos alternativos, que acomodem essa estrutura de variação é recomendado, como o modelo binomial negativa.

```
> summary(logis1.semas3)
```

Call:

```
glm(formula = resp ~ pais, family = binomial, data = olhos, subset = -c(13,
42, 58))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.232	-0.606	0.492	0.735	2.092

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.626	0.148	4.24	2.2e-05	***
pais1	2.804	0.347	8.09	6.0e-16	***
pais2	-0.220	0.455	-0.48	0.628	
pais3	-2.114	0.372	-5.68	1.4e-08	***
pais4	0.664	0.224	2.97	0.003	**

```

pais5          -0.451      0.200   -2.26    0.024 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 245.671 on 74 degrees of freedom
Residual deviance: 68.038 on 69 degrees of freedom
AIC: 192.3

```

Number of Fisher Scoring iterations: 5

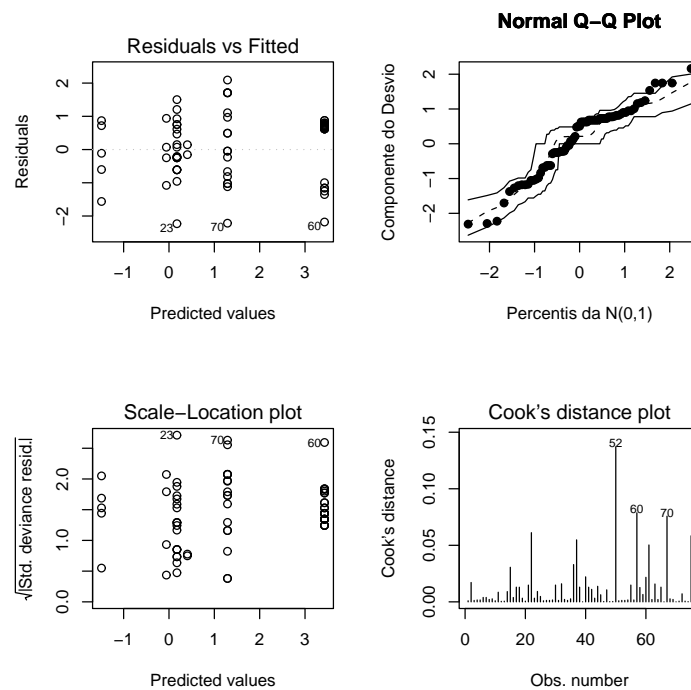


Figura 5: Gráficos de diagnóstico para o modelo ajustado

- No arquivo `recrutas.dat` são descritos os resultados de um estudo desenvolvido em 1990 com recrutas americanos referente a associação entre o número de infecções de ouvido e alguns fatores. Os dados são apresentados na seguinte ordem: hábito de nadar (ocasional ou frequente), local onde costuma nadar (piscina ou praia), faixa etária (15-19, 20-25 ou 25-29), sexo (masculino ou feminino) e número de infecções de ouvido diagnosticadas pelo próprio recruta. Verifique qual dos modelos, log-linear de Poisson ou

log-linear binomial negativa, se ajusta melhor aos dados. Selecionar para cada modelo através do método de AIC os efeitos principais e depois as interações de primeira ordem. Utilize métodos de diagnóstico como critério. Interprete os resultados do modelo escolhido através de razões de taxas.

Começamos ajustando um modelo log-linear de Poisson:

```
> rec.logis1 <- glm(infecoes ~ (.)^2, family = poisson,
+ data = recrutas)
> summary(rec.logis1)
```

Call:

```
glm(formula = infecoes ~ (.)^2, family = poisson, data = recrutas)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.267	-1.561	-1.048	0.576	5.948

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.1401	0.0699	2.00	0.04514 *
habito1	-0.3383	0.0633	-5.35	8.9e-08 ***
local1	-0.2543	0.0669	-3.80	0.00014 ***
fetaria1	0.2673	0.0810	3.30	0.00097 ***
fetaria2	-0.2309	0.1003	-2.30	0.02127 *
sexo1	0.0251	0.0634	0.40	0.69235
habito1:local1	0.0876	0.0595	1.47	0.14118
habito1:fetaria1	0.1357	0.0729	1.86	0.06274 .
habito1:fetaria2	0.0832	0.0905	0.92	0.35798
habito1:sexo1	0.0137	0.0609	0.22	0.82222
local1:fetaria1	0.1103	0.0792	1.39	0.16349
local1:fetaria2	-0.1791	0.1001	-1.79	0.07363 .
local1:sexo1	0.1616	0.0622	2.60	0.00938 **
fetaria1:sexo1	0.0907	0.0805	1.13	0.25974
fetaria2:sexo1	-0.1083	0.0927	-1.17	0.24254

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 824.51 on 286 degrees of freedom
 Residual deviance: 730.31 on 272 degrees of freedom
 AIC: 1133

Number of Fisher Scoring iterations: 6

Fazemos então a seleção de variáveis pelo procedimento AIC passo a passo:

```
> rec.logis2 <- stepAIC(rec.logis1, trace = 0)
> summary(rec.logis2)
```

Call:

```
glm(formula = infecoes ~ habito + local + fetaria + sexo + habito:local +
     habito:fetaria + local:fetaria + local:sexo, family = poisson,
     data = recrutas)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.239	-1.535	-1.030	0.575	6.125

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.1197	0.0685	1.75	0.08053	.
habito1	-0.3410	0.0616	-5.53	3.1e-08	***
local1	-0.2455	0.0662	-3.71	0.00021	***
fetaria1	0.2437	0.0777	3.14	0.00171	**
fetaria2	-0.2207	0.1000	-2.21	0.02736	*
sexo1	0.0347	0.0586	0.59	0.55386	
habito1:local1	0.0919	0.0554	1.66	0.09737	.
habito1:fetaria1	0.1330	0.0725	1.83	0.06685	.
habito1:fetaria2	0.0874	0.0894	0.98	0.32865	
local1:fetaria1	0.1495	0.0738	2.03	0.04284	*
local1:fetaria2	-0.2210	0.0965	-2.29	0.02201	*
local1:sexo1	0.1941	0.0586	3.31	0.00093	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 824.51 on 286 degrees of freedom
Residual deviance: 732.16 on 275 degrees of freedom
AIC: 1129

Number of Fisher Scoring iterations: 6

Ambos os modelos têm um desvio residual muito grande (P-valor 0+).
Na Figura 6 temos os gráficos de resíduos para o modelo selecionado pelo

procedimento do AIC. Os pontos ficam quase todos fora da banda de confiança no gráfico do envelope e há algumas observações com influência muito grande no ajuste. Isso junto com o fato de quase todas as variáveis incluídas no modelo terem sido significantes, é um indício de que um modelo de superdispersão seja mais adequado.

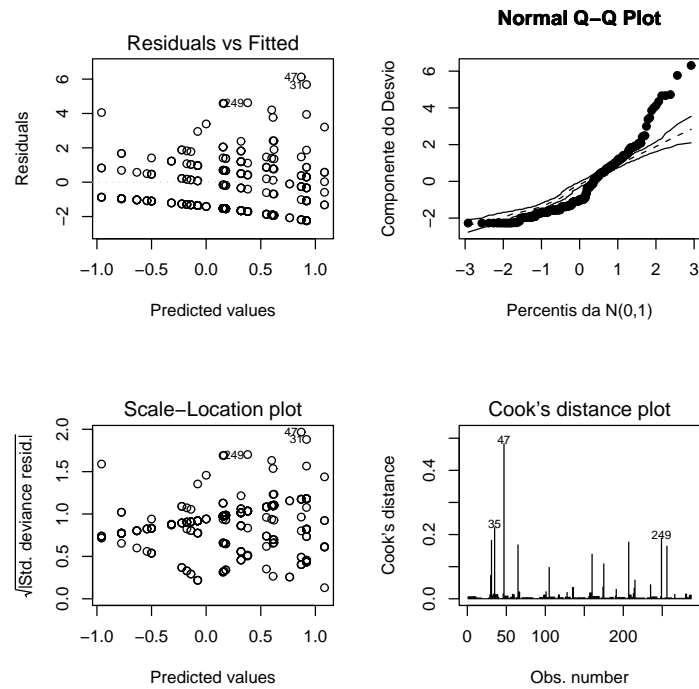


Figura 6: Gráficos de diagnóstico para o modelo Poisson

Ajustamos então um modelo log-linear binomial negativa através do comando:

```
> rec.nb1 <- glm.nb(infecoes ~ (.)^2, data = recrutas)
> summary(rec.nb1, cor = FALSE)
```

Call:

```
glm.nb(formula = infecoes ~ (.)^2, data = recrutas, init.theta = 0.613896073248922,
link = log)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.422	-1.138	-0.816	0.304	2.190

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.13686	0.11325	1.21	0.2269
habito1	-0.32489	0.10486	-3.10	0.0019 **
local1	-0.25044	0.11035	-2.27	0.0232 *
fetaria1	0.27522	0.14042	1.96	0.0500 .
fetaria2	-0.22473	0.15673	-1.43	0.1516
sexo1	0.03401	0.10843	0.31	0.7537
habito1:local1	0.10337	0.09887	1.05	0.2958
habito1:fetaria1	0.09747	0.12695	0.77	0.4426
habito1:fetaria2	0.14124	0.15212	0.93	0.3531
habito1:sexo1	-0.00417	0.10387	-0.04	0.9680
local1:fetaria1	0.11833	0.13722	0.86	0.3885
local1:fetaria2	-0.17062	0.15988	-1.07	0.2859
local1:sexo1	0.16698	0.11056	1.51	0.1310
fetaria1:sexo1	0.05883	0.14339	0.41	0.6816
fetaria2:sexo1	-0.10681	0.15921	-0.67	0.5023

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.614) family taken to be 1)

Null deviance: 299.58 on 286 degrees of freedom
Residual deviance: 269.87 on 272 degrees of freedom
AIC: 914.8

Number of Fisher Scoring iterations: 1

Theta: 0.6139
Std. Err.: 0.0983

2 x log-likelihood: -882.8440

Onde vemos que agora o desvio residual está bem mais próximo dos graus de liberdade. Selecionamos os efeitos significativos através do procedimento de AIC:

```
> rec.nb2 <- stepAIC(rec.nb1, trace = 0)
> summary(rec.nb2, cor = FALSE)
```

Call:

```
glm.nb(formula = infecoes ~ habito + local + sexo + local:sexo,
```

```
data = recrutas, init.theta = 0.575993095629647, link = log)
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-1.38	-1.12	-0.92	0.29	2.49

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.2187	0.1024	2.14	0.0327 *
habito1	-0.2967	0.0947	-3.13	0.0017 **
local1	-0.1901	0.1019	-1.86	0.0622 .
sexo1	0.0174	0.1019	0.17	0.8645
local1:sexo1	0.1863	0.1019	1.83	0.0676 .

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for Negative Binomial(0.576) family taken to be 1)
```

```
Null deviance: 289.91 on 286 degrees of freedom  
Residual deviance: 269.54 on 282 degrees of freedom  
AIC: 903.1
```

```
Number of Fisher Scoring iterations: 1
```

```
Theta: 0.5760  
Std. Err.: 0.0905
```

```
2 x log-likelihood: -891.1040
```

Novamente o desvio residual está bem próximo dos graus de liberdade. Na Figura 7 temos os gráficos de resíduos para esse modelo. Vemos que todas as suposições parecem satisfeitas. Os pontos ficam todo dentro do envelope simulado e não há aparentemente observações com influência desproporcional.

Começamos analisando a razão de taxas para a variável hábito. De acordo com a estimativa do coeficiente para essa variável, fixadas as outras variáveis, alguém que nada ocasionalmente tem uma taxa de risco de ter uma infecção no ouvido 45% menor do que alguém que nada frequentemente ($e^{2\beta_1} = 0.552$). Como a interação ficou no modelo, devemos analisar as taxas de risco de sexo dentro dos níveis de local ou vice-versa. Consideremos as pessoas que costumam nadar na praia. Nesse caso as mulheres tem taxa de risco 1.5 maior de ter infecção no ouvido ($e^{2(\beta_3+\beta_{34})} = 1.50$). Já na piscina as mulheres tem taxa de risco 30% menor de sofrerem infecção de ouvido ($e^{2(\beta_3-\beta_{34})} = 0.713$)

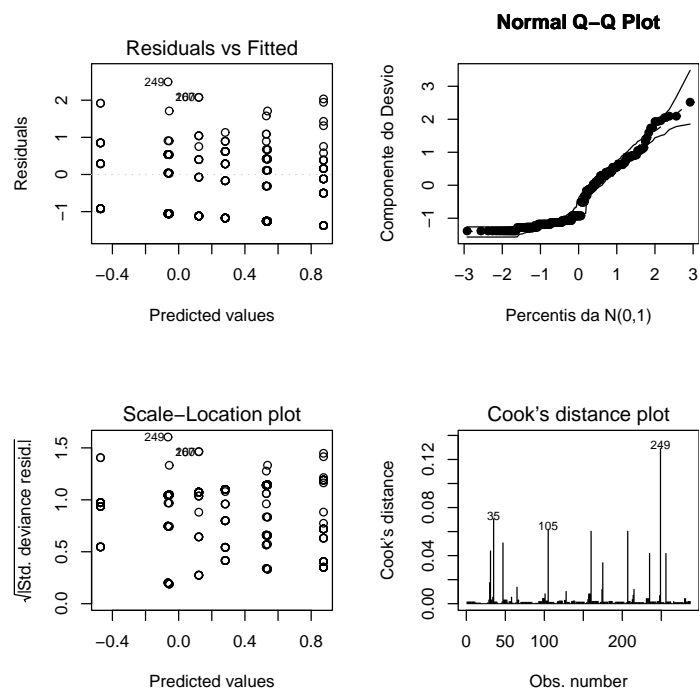


Figura 7: Gráficos de diagnóstico para o modelo binomial negativa

Sobre

A versão eletrônica desse arquivo pode ser obtida em <http://www.feferraz.net>

Copyright (c) 1999-2005 Fernando Henrique Ferraz Pereira da Rosa.
 É dada permissão para copiar, distribuir e/ou modificar este documento sob os termos da Licença de Documentação Livre GNU (GFDL), versão 1.2, publicada pela Free Software Foundation;
 Uma cópia da licença em está inclusa na seção intitulada "Sobre / Licença de Uso".