Modelos de regressão para tabelas de contingência

Prof. Caio Azevedo

Comentários

- As variáveis definidoras das tabelas de contingência (de dupla entrada) podem ser:
  - Nominais ou ordinais.
  - Resposta ou explicativas (pelo menos uma tem de ser resposta).
- Aspectos de interesse: dependência, homogeneidade de distribuições, associação e concordância.
- Modelos probabilísticos geradores das tabelas: multinomial, produto de multinomiais independentes, produto de Poissons independentes, hipergeométricas multivariadas (não-centrais).

Tabela de contingência $r \times s$: multinomial

		Variável 1 (resposta)					
		C_{11}	C_{12}	...	$C_{1(s-1)}$	C_{1s}	Total
Variável 2	C_{21}	$Y_{11}(p_{11})$	$Y_{12}(p_{12})$	...	$Y_{1(s-1)}(p_{1(s-1)})$	$Y_{1s}(p_{1s})$	$Y_{1.}$
(resposta)	C_{22}	$Y_{21}(p_{21})$	$Y_{22}(p_{22})$	...	$Y_{1(s-1)}(p_{2(s-1)})$	$Y_{2s}(p_{2s})$	$Y_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	
	C_{2r}	$Y_{r1}(p_{r1})$	$Y_{r2}(p_{r2})$	...	$Y_{r(s-1)}(p_{r(s-1)})$	$Y_{rs}(p_{rs})$	$Y_{r.}$
Total	-	$Y_{.1}$	$Y_{.2}$	...	$Y_{.(s-1)}$	$Y_{.s}$	$y_{..}$

Somente o total geral é fixado.

Tabela de contingência $r \times s$: produto de multinomiais independentes

		Variável 1 (resposta)					
		C_{11}	C_{12}	...	$C_{1(s-1)}$	C_{1s}	Total
Variável 2	C_{21}	$Y_{11}(p_{11})$	$Y_{12}(p_{12})$	...	$Y_{1(s-1)}(p_{1(s-1)})$	$Y_{1s}(p_{1s})$	$y_{1.}$
(explicativa)	C_{22}	$Y_{21}(p_{21})$	$Y_{22}(p_{22})$	...	$Y_{1(s-1)}(p_{2(s-1)})$	$Y_{2s}(p_{2s})$	$y_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	
	C_{2r}	$Y_{r1}(p_{r1})$	$Y_{r2}(p_{r2})$	...	$Y_{r(s-1)}(p_{r(s-1)})$	$Y_{rs}(p_{rs})$	$y_{r.}$
Total	-	$Y_{.1}$	$Y_{.2}$	...	$Y_{.(s-1)}$	$Y_{.s}$	$y_{..}$

Os totais marginais por linha ou coluna são fixados.

Tabela de contingência $r \times s$: produto de Poisson independentes

		Variável 1 (resposta)					
		C_{11}	C_{12}	...	$C_{1(s-1)}$	C_{1s}	Total
Variável 2	C_{21}	$Y_{11}(\mu_{11})$	$Y_{12}(\mu_{12})$	...	$Y_{1(s-1)}(\mu_{1(s-1)})$	$Y_{1s}(\mu_{1s})$	$Y_{1.}$
(resposta)	C_{22}	$Y_{21}(\mu_{21})$	$Y_{22}(\mu_{22})$	...	$Y_{1(s-1)}(\mu_{2(s-1)})$	$Y_{2s}(\mu_{2s})$	$Y_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	
	C_{2r}	$Y_{r1}(\mu_{r1})$	$Y_{r2}(\mu_{r2})$	...	$Y_{r(s-1)}(\mu_{r(s-1)})$	$Y_{rs}(\mu_{rs})$	$Y_{r.}$
Total	-	$Y_{.1}$	$Y_{.2}$	...	$Y_{.(s-1)}$	$Y_{.s}$	$Y_{..}$

Nenhuma quantidade é fixada.

Modelos de regressão

- Em termos de modelo de regressão
  - Tabela 1 (multinomial): modelos de regressão para dados multinomiais com $m_i = m, i = 1$ (vistos aqui), modelo de Poisson (veremos adiante)).
  - Tabela 2 (produtos de multinomiais): modelos de regressão para dados multinomiais com $m_i \geq 1, i = 1, ...n$ (vistos aqui), modelo de Poisson (veremos adiante)).
  - Tabela 3 (produtos de Poisson): modelos de regressão para dados de Poisson $m = 1$ (vistos aqui).

Relação entre os modelos Poisson e Multinomial

- Considera a Tabela do slide 5 e suponha que $Y_{ij} \overset{ind.}{\sim} P(\mu_{ij})$, $i = 1, 2, ..., r; j = 1, 2..., s$. Assim

$$\begin{aligned} Y_{11}, Y_{12}, ..., Y_{r(s-1)} | Y_{..} = m &\sim \text{multinomial}_{(ks-1)}(m, p_{ij}) \\ p_{ij} &= \frac{\mu_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{s} \mu_{ij}}. \end{aligned}$$

- Ou seja, é possível fazer inferência a respeito de p_{ij} através de μ_{ij}.

Relação entre os modelos Poisson e Multinomial

- Vamos considerar, adicionalmente, que:

$$\begin{aligned} \ln \mu_{ij} &= \alpha + \beta_i + \gamma_j + (\beta\gamma)_{ij} \\ \beta_1 &= \gamma_1 = (\beta\gamma)_{1j} = (\beta\gamma)_{i1} = 0 \forall i, j. \end{aligned}$$

- Logo, $\mu_{ij} = e^{\alpha+\beta_i+\gamma_j+(\beta\gamma)_{ij}}$ e, também:

$$p_{ij} = \frac{e^{\beta_i+\gamma_j+(\beta\gamma)_{ij}}}{\sum_{i=1}^{r}\sum_{j=1}^{s} e^{\beta_i+\gamma_j+(\beta\gamma)_{ij}}}.$$

Relação entre os modelos Poisson e Multinomial

- Assim, podemos ajustar o modelo multinomial:

$$
\begin{aligned}
Y_{11}, Y_{12}, ..., Y_{r(s-1)} | Y_{..} = m &\sim \text{multinomial}_{(ks-1)}(m, \boldsymbol{p}) \\
\boldsymbol{p} &= (p_{11}, p_{12}, ...p_{r(s-1)})' \\
p_{ij} = \frac{e^{\beta_i + \gamma_j + (\beta\gamma)_{ij}}}{\sum_{i=1}^{r}\sum_{j=1}^{s} e^{\beta_i + \gamma_j + (\beta\gamma)_{ij}}}, \beta_1 &= \gamma_1 = (\beta\gamma)_{1j} = (\beta\gamma)_{i1} = 0, \forall i, j
\end{aligned}
$$

- Ou o modelo Poisson:

$$
\begin{aligned}
Y_{ij} &\overset{ind.}{\sim} \text{Poisson}(\mu_{ij}) \\
\ln \mu_{ij} &= \alpha + \beta_i + \gamma_j + (\beta\gamma)_{ij} \\
\beta_1 &= \gamma_1 = (\beta\gamma)_{1j} = (\beta\gamma)_{i1} = 0 \forall i, j.
\end{aligned}
$$

Relação entre os modelos Poisson e Multinomial

- Paula (2024) prova que as estimativas de máxima verossimilhança de $\beta_i, \gamma_j, (\alpha\beta)_{ij})$, $i = 1, 2, ..., r, j = 1, 2, ..., s$ são equivalente entre ambos os modelos.
- Uma das vantagens de usar o modelo Poisson é sua menor complexidade em relação ao modelo multinomial e (eventualmente) maior simplicidade na verificação do ajuste do modelo. Contudo, neste caso, temos um modelo saturado ($n \equiv p$), a contrário do que acontece com o modelo multinomial.

Exemplo 12: classificação de indivíduos segundo renda e grau de satisfação no emprego

- A Tabela a seguir apresenta o resultado de uma pesquisa com 901 indivíduos (Agresti (1990), pgs. 20-21) classificados segundo a renda anual e o grau de satisfação no emprego. Denotamos por Y_{ij} o número de indivíduos pertencentes à classe de renda i com grau de satisfação j.
- Por simplicidade, consideraremos que essas duas variáveis são categorizadas nominais.
- Exercício: resolver considerando ambas como ordinais.

Cont.

	grau de satisfação			
renda (US$)	alto	bom	médio	baixo
$<$6000	20	24	80	82
6000-15000	22	38	104	125
15000-25000	13	28	81	113
$>$25000	7	18	54	92

Um dos objetivos é ver como o grau de satisfação e renda estão relacionados (bem como avaliar como as pessoas se distribuem ao longo dos grupos: combinações entre os níveis das variáveis).

Cont.

- temos que o modelo gerador dos dados é:

$$\begin{aligned} \boldsymbol{Y} = (Y_{11}, Y_{12}, ..., Y_{33})' &\sim \text{multinomial}_{15}(901, \boldsymbol{p}) \\ \boldsymbol{p} &= (p_{11}, p_{12}, ..., p_{33}) \end{aligned}$$

- Teste de qui-quadrado para testar H_0 : não há dependência vs H_1 : há dependência: 11,99 (p= 0,2140).
- Vamos utilizar dois modelos para modelar o problema.

Cont.

- Modelo 1 (resposta: contagens em cada grupo, explicativas: gs e renda):

$$
\begin{aligned}
Y_{ij} &\overset{ind.}{\sim} \text{Poisson}(\mu_{ij}) \\
\ln \mu_{ij} &= \alpha + \beta_i + \gamma_j + (\beta\gamma)_{ij} \\
\beta_1 &= \gamma_1 = (\alpha\beta)_{1j} = (\alpha\beta)_{j1} = 0 \\
i(\text{renda}) &= 1(< 6000), 2(6000 - 15000), 3(15000 - 25000), \\
j(\text{gs}) &= 1(\text{alto}), 2(\text{bom}), 3(\text{médio}), 4(\text{baixo}) \\
&\quad 4(> 25000),
\end{aligned}
$$

Cont.

- Modelo 2 (não consideraremos a distribuição multinomial, alternativamente, consideraremos um produto de multinominais, sendo gs (j) a resposta e renda a covariável (i)):

$$
\begin{aligned}
\boldsymbol{Y}_i &= (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})' \sim \text{multinomial}_4(m_i, \boldsymbol{p}_i) \\
\boldsymbol{p} &= (p_{i1}, p_{i2}, p_{i3}, p_{i4})', \sum_{j=1}^{4} y_{ij} = m_i, \sum_{j=1}^{4} p_{ij} = 1, \forall i, \\
m_1 &= 206, m_2 = 289, m_3 = 235, m_4 = 171 \\
p_{ij} &= \frac{e^{\eta_{ij}}}{\sum_{j=1}^{4} e^{\eta_{ij}}} = \frac{e^{\eta_{ij}}}{p_i}; \eta_{ij} = \alpha_j + \beta_{ij}, \alpha_1 = \beta_{i1} = 0, \forall i
\end{aligned}
$$

Cont.

Modelo 1

	grau de satisfação			
renda (US$)	alto	bom	médio	baixo
$<$6000	e^{α}	$e^{\alpha+\gamma_2}$	$e^{\alpha+\gamma_3}$	$e^{\alpha+\gamma_4}$
6000-15000	$e^{\alpha+\beta_2}$	$e^{\alpha+\beta_2+\gamma_2+(\alpha\beta)_{22}}$	$e^{\alpha+\beta_2+\gamma_3+(\alpha\beta)_{23}}$	$e^{\alpha+\beta_2+\gamma_4+(\alpha\beta)_{24}}$
15000-25000	$e^{\alpha+\beta_3}$	$e^{\alpha+\beta_3+\gamma_2+(\alpha\beta)_{32}}$	$e^{\alpha+\beta_3+\gamma_3+(\alpha\beta)_{33}}$	$e^{\alpha+\beta_3+\gamma_4+(\alpha\beta)_{34}}$
$>$25000	$e^{\alpha+\beta_4}$	$e^{\alpha+\beta_4+\gamma_2+(\alpha\beta)_{42}}$	$e^{\alpha+\beta_4+\gamma_3+(\alpha\beta)_{43}}$	$e^{\alpha+\beta_4+\gamma_4+(\alpha\beta)_{44}}$

Cont.

Modelo 2

	grau de satisfação			
renda (US$)	alto	bom	médio	baixo
$<$6000	e^0/p_1	e^{α_2}/p_1	e^{α_3}/p_1	e^{α_4}/p_1
6000-15000	e^0/p_2	$e^{\alpha_2+\beta_{22}}/p_2$	$e^{\alpha_3+\beta_{23}}/p_2$	$e^{\alpha_4+\beta_{24}}/p_2$
15000-25000	e^0/p_3	$e^{\alpha_2+\beta_{32}}/p_3$	$e^{\alpha_3+\beta_{33}}/p_3$	$e^{\alpha_4+\beta_{34}}/p_3$
$>$25000	e^0/p_4	$e^{\alpha_2+\beta_{42}}/p_4$	$e^{\alpha_3+\beta_{43}}/p_4$	$e^{\alpha_4+\beta_{44}}/p_4$

em que $p_i = \sum_{j=1}^{4} e^{\eta_{ij}}$.

Cont.

- No modelo 1, $\nexists$ interação $\leftrightarrow (\beta\gamma)_{ij} = 0$.
- No modelo 2, $\nexists$ interação $\leftrightarrow \beta_{ij} = 0$ (equivalente a dizer que as distribuições do gs, ao longo dos níveis dee renda são idêntica).

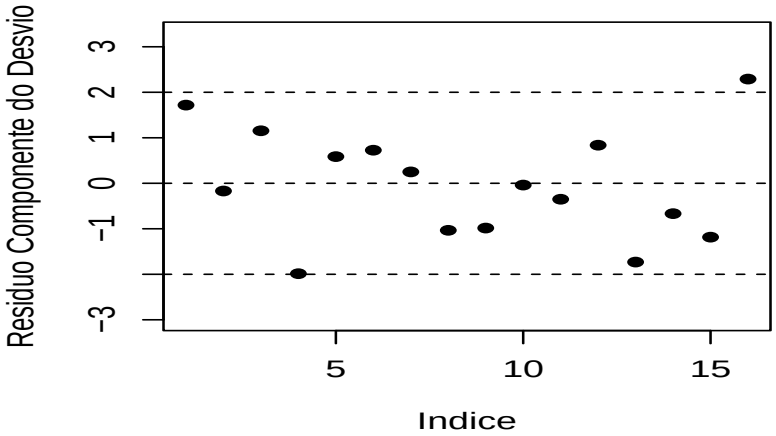
Modelo 1 (estimativas dos parâmetros)

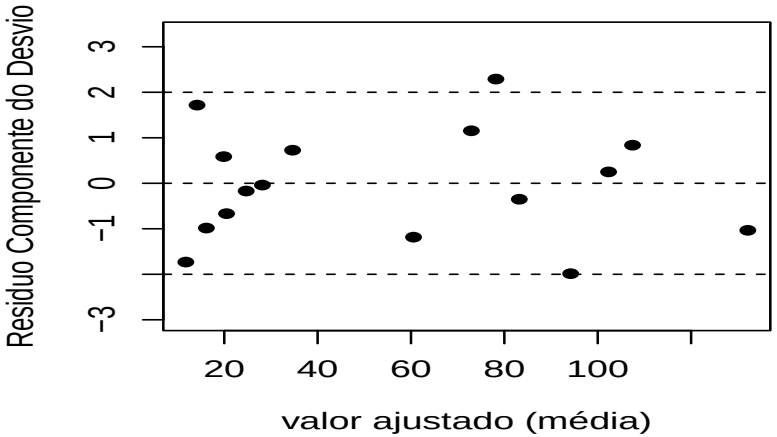
Par.	Est.	EP	IC(95%)	Estat. Z_t	p-valor
α	3,00	0,22	[2,56;3,43]	13,40	0,0000
β_2	0,10	0,31	[-0,51;0,70]	0,31	0,7577
β_3	-0,43	0,36	[-1,13;0,27]	-1,21	0,2266
β_4	-1,05	0,44	[-1,91;-0,19]	-2,39	0,0168
γ_2	0,18	0,30	[-0,41;0,78]	0,60	0,5470
γ_3	1,39	0,25	[0,90;1,88]	5,55	0,0000
γ_4	1,41	0,25	[0,92;1,90]	5,66	0,0000
$(\beta\gamma)_{22}$	0,36	0,40	[-0,43;1,16]	0,90	0,3676
$(\beta\gamma)_{23}$	0,58	0,45	[-0,30;1,47]	1,29	0,1956
$(\beta\gamma)_{24}$	0,76	0,54	[-0,29;1,82]	1,42	0,1570
$(\beta\gamma)_{32}$	0,17	0,34	[-0,50;0,84]	0,49	0,6261
$(\beta\gamma)_{33}$	0,44	0,39	[-0,32;1,21]	1,14	0,2553
$(\beta\gamma)_{34}$	0,66	0,47	[-0,27;1,58]	1,39	0,1651
$(\beta\gamma)_{42}$	0,33	0,34	[-0,34;0,99]	0,96	0,3373
$(\beta\gamma)_{43}$	0,75	0,38	[-0,00;1,51]	1,95	0,0508
$(\beta\gamma)_{44}$	1,16	0,46	[0,25;2,08]	2,51	0,0122

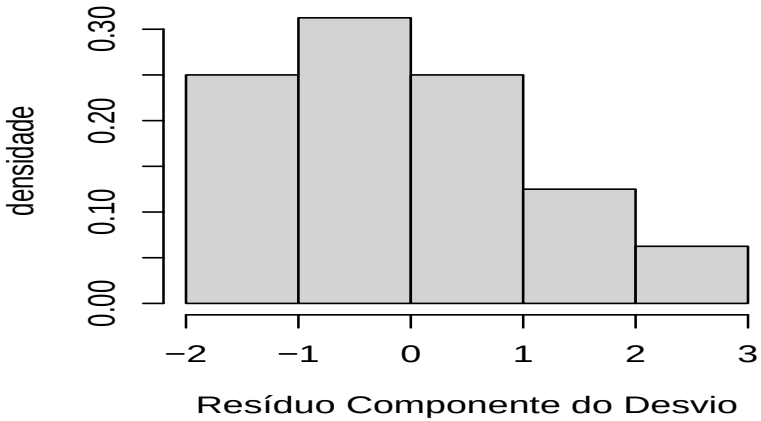
Comentários

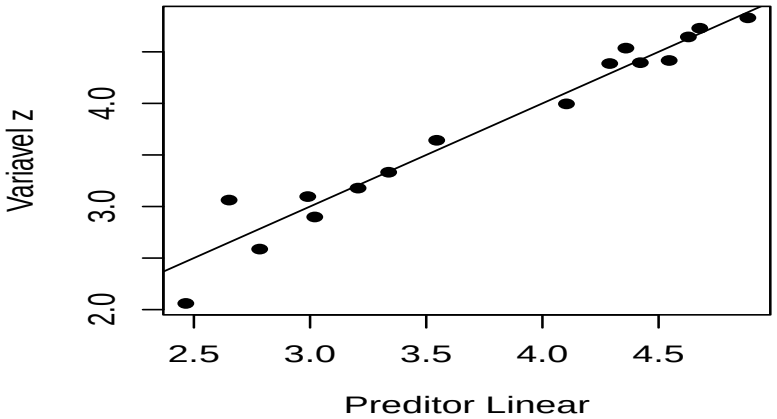
- Não é possível realizar análise de diagnóstico pois se trata de um modelo saturado (n=p).
- Temos indícios de ausência de interação pois: TRV (p= 0,2112) e Teste do tipo Wald ($\boldsymbol{C\beta} = \boldsymbol{M}$) (p=0,2278).
- Ajustaremos um modelo sem interação.

Gráficos de diagnóstico: modelo 1

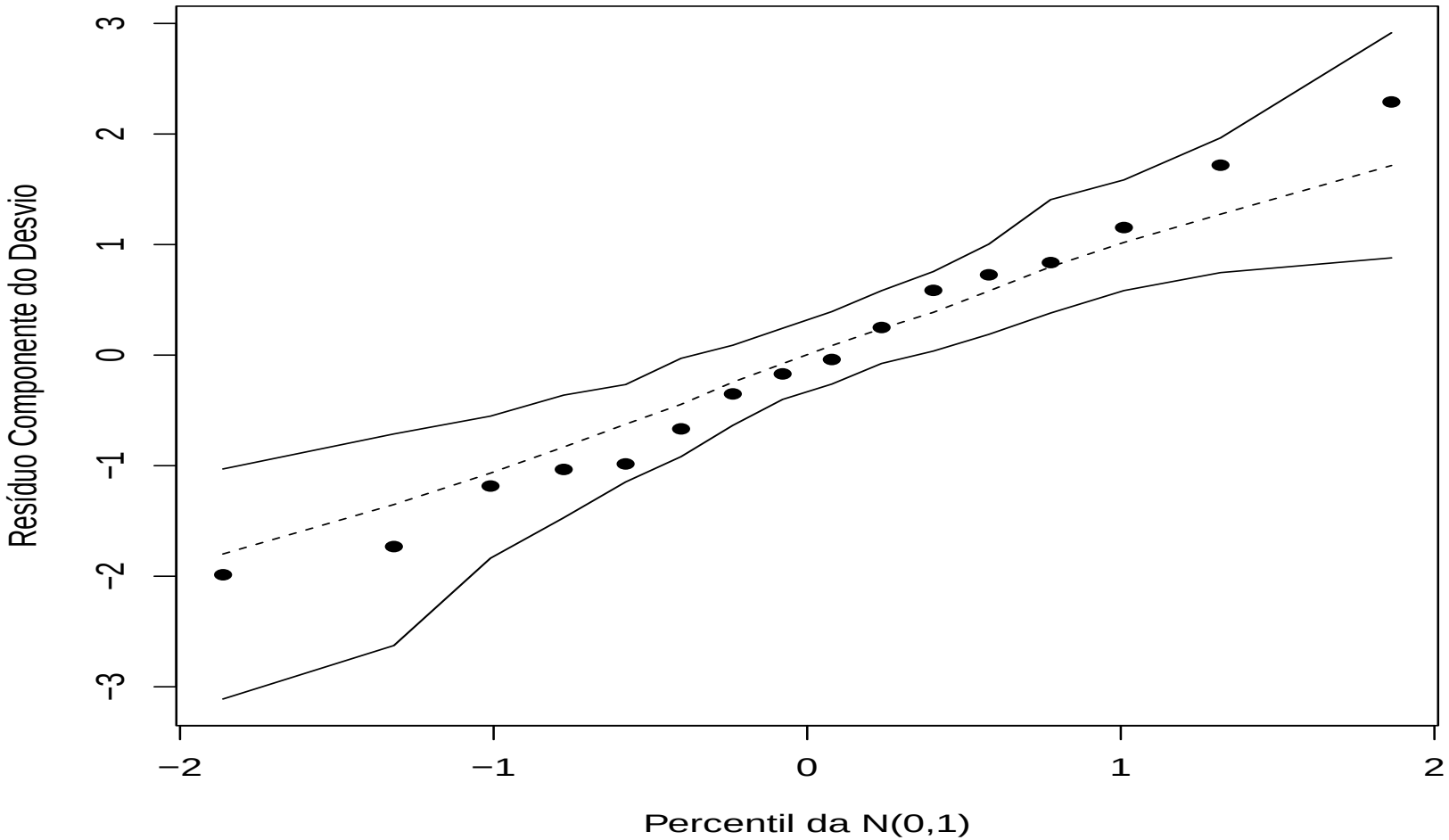




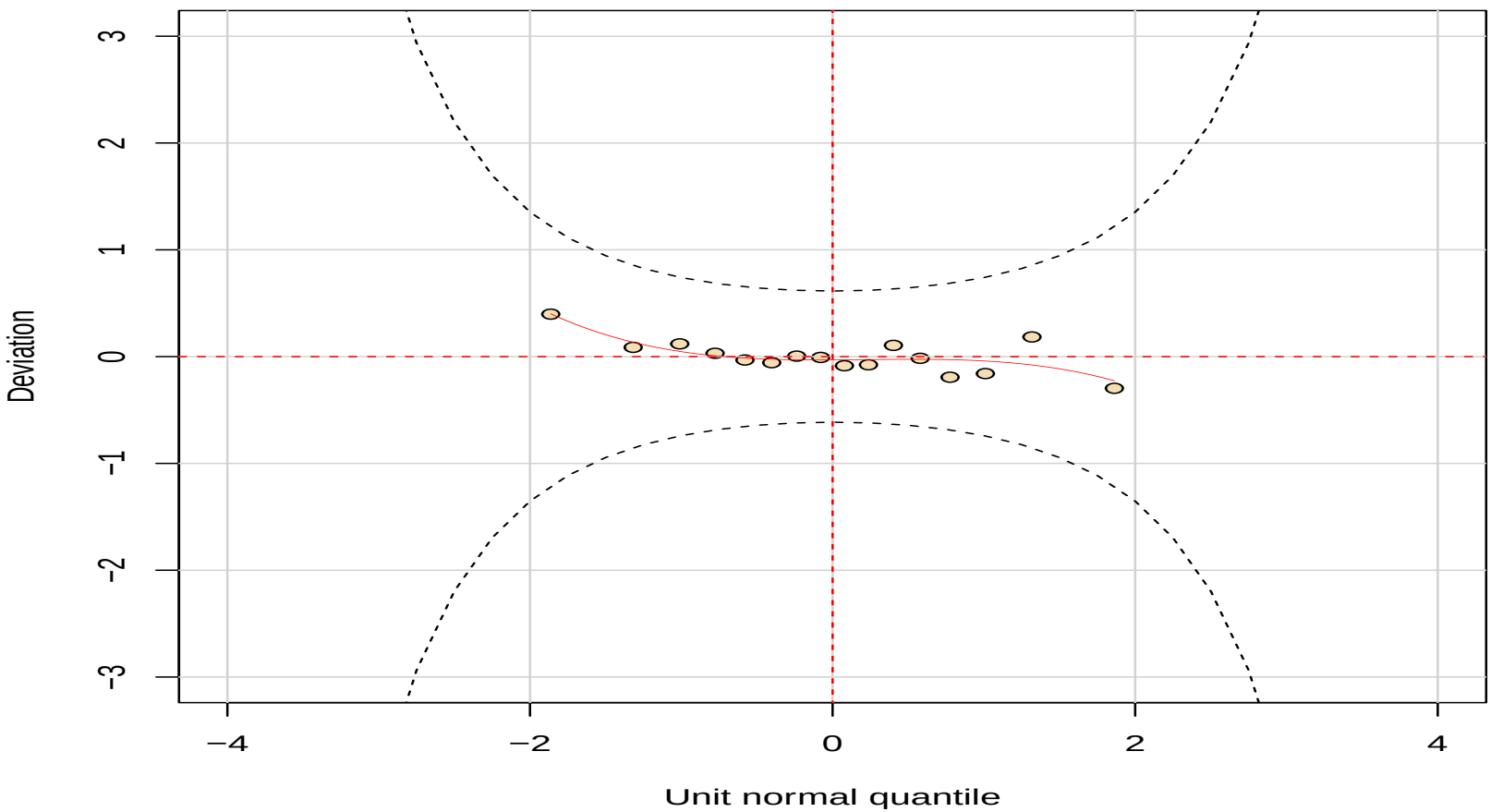




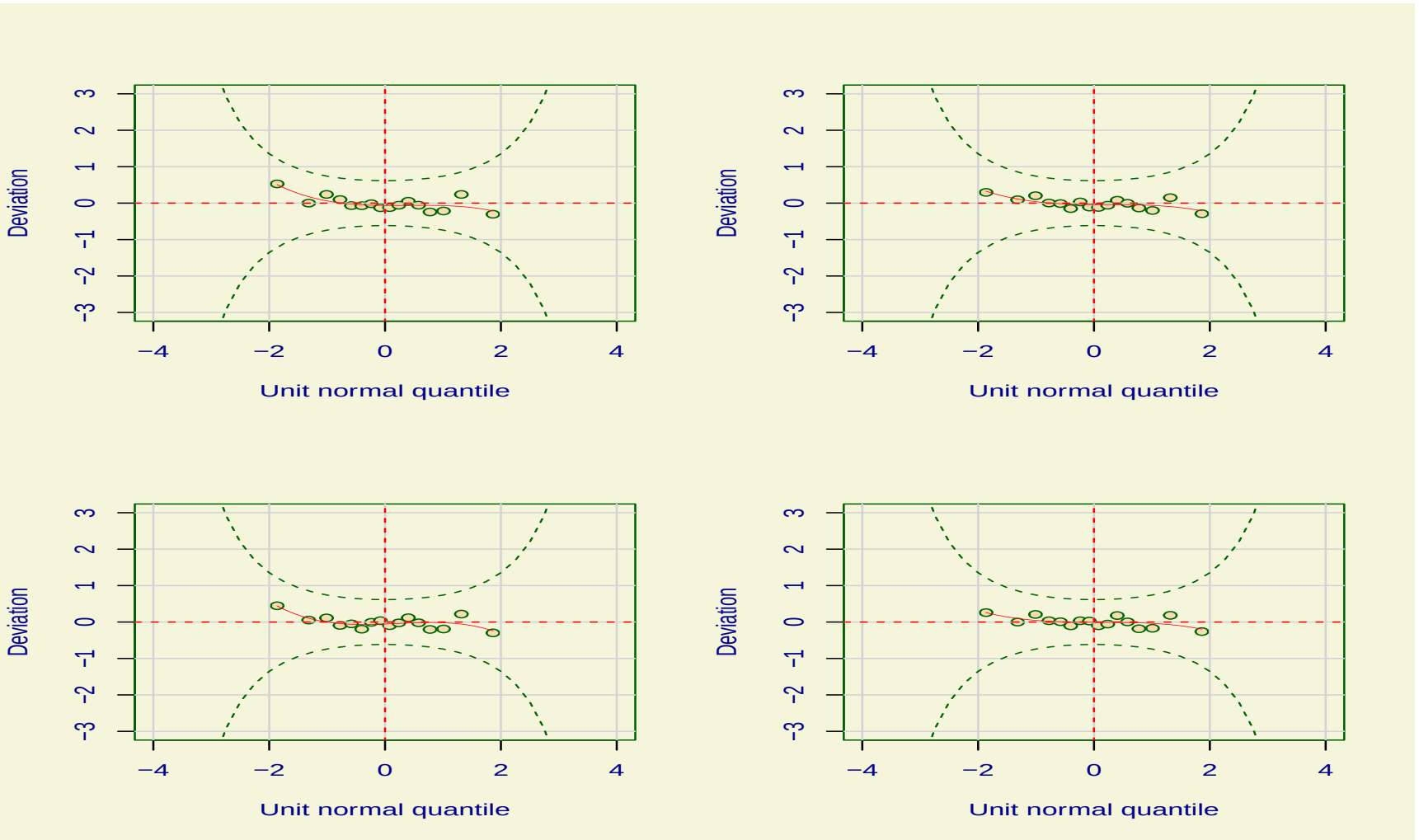
QQ plot com envelope: modelo 1



Worm plot; modelo 1



Worm plotse: modelo 1



Cont.

Par.	Est.	EP	IC(95%)	Estat. Z_t	p-valor
α	2,65	0,14	[2,38 ;2,93]	18,81	$< 0{,}0001$
β_2	0,34	0,09	[0,16;0,52]	3,71	0,0002
β_3	0,13	0,10	[-0,06;0,32]	1,38	0,1676
β_4	-0,19	0,10	[-0,39;0,02]	-1,80	0,0719
γ_2	0,55	0,16	[0,24;0,87]	3,48	$< 0{,}0001$
γ_3	1,64	0,14	[1,37;1,91]	11,80	$< 0{,}0001$
γ_4	1,89	0,14	[1,63;2,16]	13,90	$< 0{,}0001$

Comentários

- O modelo apresentou um ajusta razoável mas, há indícios de presença de superdispersão.
- A interação se mostrou não significativa.
- Os fatores se mostraram significativos, em que há indícios que há um aumento no número de pessoas, à medida que o grau de satisfação diminui. Por outro lado há, aparentemente, mais pessoas na segunda faixa de renda, e a mesma quantidade nas demais.
- Um modelo binomial negativo deveria ser utilizado.
- Exercício: fazer uma análise de influência.

Modelo 2

- Proporções amostrais por linhas (em função da renda):

	grau de satisfação			
renda (US$)	alto	bom	médio	baixo
$<$6000	0,10	0,12	0,39	0,40
6000-15000	0,08	0,13	0,36	0,43
15000-25000	0,06	0,12	0,34	0,48
$>$25000	0,04	0,11	0,32	0,54

As proporções parecem semelhante, para uma mesmo grau de satisfação, ao longo das linhas.

Modelo 2 (estimativas dos parâmetros)

Par.	Est.	EP	IC(95%)	Estat. Z_t	p-valor
α_2	0,18	0,30	[-0,41;0,78]	0,6022	0,5470
α_2	1,39	0,25	[0,90;1,88]	5,5452	0,0000
α_2	1,41	0,2494	[0,92;1,90]	5,6578	0,0000
β_{22}	0,36	0,40	[-0,43;1,16]	0,9009	0,3676
β_{23}	0,17	0,34	[-0,50;0,84]	0,4872	0,6261
β_{24}	0,33	0,34	[-0,34;0,99]	0,9595	0,3373
β_{32}	0,58	0,45	[-0,30;1,47]	1,2941	0,1956
β_{33}	0,44	0,39	[-0,32;1,21]	1,1377	0,2553
β_{34}	0,75	0,38	[<-0,00;1,51]	1,9535	0,0508
β_{42}	0,76	0,54	[-0,29;1,82]	1,42	0,1570
β_{43}	0,66	0,47	[-0,27;1,58]	1,39	0,1651
β_{44}	1,165	0,46	[0,25;2,08]	2,51	0,0122

Modelo 2 (estimativas dos parâmetros)

- Aparentemente, não há interação (há independência) entre os fatores. Com efeito, um teste do tipo Wald ($\boldsymbol{C\beta} = \boldsymbol{M}$) resultou em p= 0,2278.
- Assim, pode analisar as duas distribuições multinomais (renda e gs) de forma independente.
- Com efeito, as proporções estimadas (por MV) coincidem com as proporções amostrais (o que sempre acontece), ou seja (próximo slide):

Modelo 2 (estimativas dos parâmetros)

- Cont.

	Categoria			
Variável	1	2	3	4
gs	0,07	0,12	0,35	0,46
renda	0,23	0,32	0,26	0,19

- Essas estimativas concordam com as conclusões obtidas via modelo 1.