MAE5763 - Modelos Lineares Generalizados - Resolução da Lista 3

Guilherme Marthe - 8661962

3/12/2020

1 Exercício 1

A base disponibilizada corresponde à uma análise para entender a associação entre o uso de 3 medicamentos diferentes ou mais no tratamento de doenças mentais. Essa variável é `polypharmacy` (=0 uso de no máximo 3 medicamentos diferentes; =1 uso de mais de 3 medicamentos diferentes). O estudo consiste em acompanhar 500 indivíduos ao longo de 7 anos observando o uso de medicamentos e as variáveis explicativas, que consistem em:

- (i) `mhv4`: número de consultas ambulatoriais relacionadas à saúde mental (nenhum, 'um a cinco', 'seis a quatorze' e 'maior do que quatorze')
- (ii) `inptmhv3`: número de internações hospitalares relacionadas à saúde mental ('nenhuma', 'uma' e 'mais do que uma')
- (iii) `gender`: gênero ('Feminino', 'Masculino'),
- (iv) `urban`: local de residência ('Urbana', 'Rural'))
- (v) `comorbid`: existência de comorbidade ('Não', 'Sim')
- (vi) `age`: idade em anos

Abaixo mostro as primeiras 10 linhas da base, que mostram as 7 observações do indivíduo com `id` = 1 e os dois primeiros anos para o segundo indivíduo.

polypharmacy	id	year	mhv4	inptmhv3	gender	urban	comorbid	age
No	1	2002	0	0	Female	Urban	Yes	4.67
No	1	2003	1-5	0	Female	Urban	Yes	5.67
No	1	2004	0	0	Female	Urban	No	6.00
No	1	2005	1-5	0	Female	Urban	Yes	7.08
No	1	2006	0	0	Female	Urban	Yes	8.00
No	1	2007	1-5	0	Female	Urban	Yes	9.92
No	1	2008	6-14	0	Female	Urban	Yes	10.67
No	2	2002	6-14	0	Male	Urban	Yes	7.58
No	2	2003	6-14	0	Male	Urban	Yes	8.08
No	2	2004	> 14	0	Male	Urban	No	9.83

1.1 Análise descritiva

Iniciaremos a análise descritiva analisando as tabelas de contingência da variável resposta, `polipharmacy` contra cada uma das variáveis categóricas. Um ponto importante antes de começarmos. Na inspeção da variável `urban`, que representa o local de residência dos indivíduos, foi notado que o indivíduo com id=490 não teve sua localização de residência marcada na base de dados no último ano do estudo. Notamos que em todos os anos do

estudo anteriores seu local de moradia era urbano. Por isso decidimos inserir o dado faltante com essa mesma informação.

	Usa 3 medicamentos ou mais		
Variável	**No**, N = 2,681[1]	**Yes**, N = 819[1]	**Total**, N=3500[1]
comorbid			
Com comorbidade	541 (15%)	80 (2.3%)	621 (18%)
Sem comorbidade	2,140 (61%)	739 (21%)	2,879 (82%)
urban			
Urban	1,944 (56%)	585 (17%)	2,529 (72%)
Rural	737 (21%)	233 (6.7%)	970 (28%)
(Não disponível)	0	1	1
gender			
Female	660 (19%)	138 (3.9%)	798 (23%)
Male	2,021 (58%)	681 (19%)	2,702 (77%)
inptmhv3			
0	2,610 (75%)	723 (21%)	3,333 (95%)
1	51 (1.5%)	70 (2.0%)	121 (3.5%)
> 1	20 (0.6%)	26 (0.7%)	46 (1.3%)
mhv4			
0	514 (15%)	44 (1.3%)	558 (16%)
1-5	797 (23%)	112 (3.2%)	909 (26%)
6-14	758 (22%)	244 (7.0%)	1,002 (29%)
> 14	612 (17%)	419 (12%)	1,031 (29%)

[1]Estatísticas apresentadas: número de casos (% sob o total). A soma das porcentagens sob uma mesma variável explicativa, excluindo a coluna total é 1

A variável que representa a coexistência de doenças, `comorbid` está concentrada na não presença de comorbidades. Podemos ver que nesse caso, 25% = 739/2879 (ou 21% do total, 739/3500) dos indivíduos-ano fazem o uso de 3 ou mais medicamentos. A chance de polifarmácia parece indicar que aqueles sem comorbidades, $739/2140 = 0.35$, tem mais chances de usar 3 ou mais medicamentos que os indivíduos com comorbidades, $80/541 = 0.15$.

Com relação ao local de residência (`urban`) existe uma prevalência de residências urbanas. As chances entre regiões urbanas e rurias de um indivíduo ser polifarmaco são parecdidas, $585/1944 = 0.3$ (urbano), $233/737 = 0.32$ (rural).

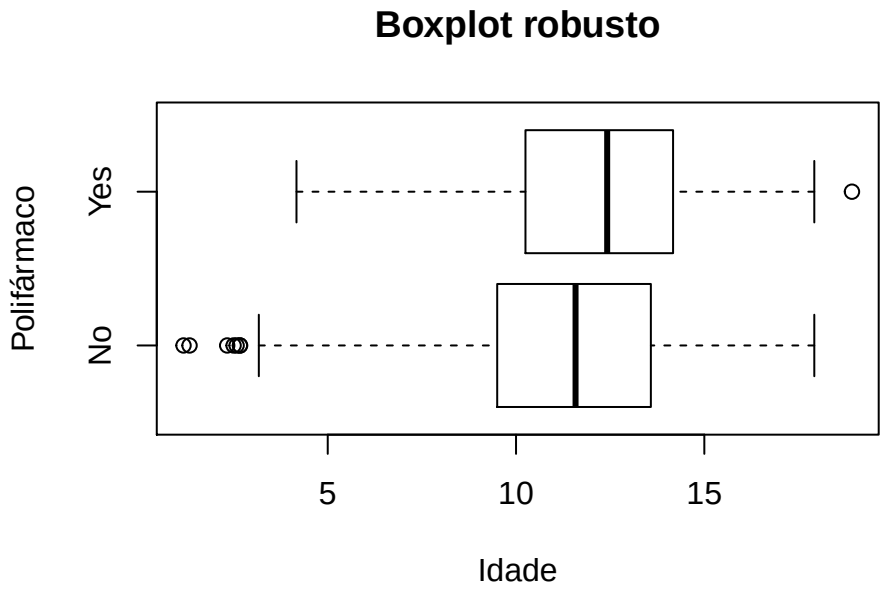
Quando estratificamos por gênero, podemos detectar que dentre os indivíduos do sexo masculino há proporção de individuos que usam muitos medicamentos maior que nos indivíduos do sexo feminino. A amostra pode ser descrita como desbalanceada com relação ao gênero uma vez que 77% dela é do sexo masculino.

Ao analisarmos a variável `inptmhv3`, que representa o número de internações por doenças mentais, podemos notar a crecente proporção de polifarmácia, partindo de uma chance de $723/2610 = 0.28$quando não existem internações para $70/51 = 1.37$ quando temos o histórico de uma internação e $26/20 = 1.3$ quando tem mais de

uma internação.

Por fim, a inspeção do número de consultas ambulatoriais relacionadas à saúde mental (variável `mhv4`) nos indica que ~40% da amostra está concentrada em indivíduos que realizaram 6 ou mais consultas. A chance de polifarmácia aumenta com o número de consultas aparentemente, uma vez que a chance quando não houveram consultas é $44/514 = 0.09$ e quando os pacientes tiveram 14 consultas ou mais $419/612 = 0.68$.

A inspeção do boxpot a seguir nos mostra que os indivídos que usam 3 medicamentos ou mais no tratamentos de doenças mentais têm, em mediana, idades mais elevadas e um indivíduo que pode ser considerado atípico por sair dos limites do gráfico. No caso de indivíduos que usam menos de 3 medicamentos, além de serem mais jovens, temos uma concetração de idades baixas possívelment atípicas quando comparadas com o resto da distribuição.



1.2 Ajuste de modelos

1.2.1 Equação de estimação generalizada Bernoulli

1.2.1.1 Seleção de modelo O primeiro modelo que iremos estimar é uma equação de estimação generalizada com uma estrutura probabilística de Bernoulli e uma estrutura de correlação auto-regressivo de ordem 1. O pacote sugerido para a análise, o **gee** não possui comparação de modelos via testes de comparação de modelos encaixados. Por isso, para realizar a seleção de modelos vamos depender de comparações univariadas via testes de Wald, uma vez que o quadrado do Z-robusto segue uma distribuição chi quadrado com um grau de liberdade.

Nessas comparações univariadas, quando a variável possui apenas dois fatores e fosse não significativa retiramos a variável. No caso da variável categórica ordinal, quando um dos fatores não era significativo ele foi juntado com um dos fatores abaixo ou acima dele de acordo com o contexto. A seguir está o código desse procedimento da seleção do modelo.

```
df <- df %>%
  mutate(urban = if_else(id == 409 & year == 2008,
                         factor('Urban', levels=c('Urban', 'Rural')),
                         urban))

df['resp'] =  if_else(df$polypharmacy == 'Yes', 1, 0)

fit1.poly = gee(resp ~ mhv4 + inptmhv3 + gender + urban + comorbid + log(age/100),
                id = id, family = binomial(link = "logit"),
                corstr = "AR-M", Mv = 1, data=df)
```

```
fit2.poly = gee(resp ~ mhv4 + inptmhv3 + gender + urban + log(age/100),
                id = id, family = binomial(link = "logit"),
                corstr = "AR-M", Mv = 1, data=df)

fit3.poly = gee(resp ~ mhv4 + inptmhv3 + gender + log(age/100),
                id = id, family = binomial(link = "logit"),
                corstr = "AR-M", Mv = 1, data=df)

df1 <- df %>%
  mutate(inptmhv3 = case_when(
    inptmhv3 == '1' ~ '>=1',
    inptmhv3 == '> 1' ~ '>=1',
    T ~ '0'
  ) %>% factor(., levels = c('0', '>=1')))

fit4.poly = gee(resp ~ mhv4 + inptmhv3 + gender + log(age/100),
                id = id, family = binomial(link = "logit"),
                corstr = "AR-M", Mv = 1, data=df1)

df2 <- df1 %>%
  mutate(mhv4 = case_when(
    mhv4 == '0' ~ '0-5',
    mhv4 == '1-5' ~ '0-5',
    T ~ as.character(mhv4)
  ) %>% factor(., c('0-5', '6-14', '> 14')))

fit5.poly = gee(resp ~ mhv4 + inptmhv3 + gender + log(age/100),
                id = id, family = binomial(link = "logit"),
                corstr = "AR-M", Mv = 1, data=df2)
```

Na tabela a seguir estão as estimativas dos coeficientes dos modelos, os valores-z robustos e a significância à 10%. O modelo final após a seleção univariada apresenta apenas coeficientes significativos (excluindo o intercepto).

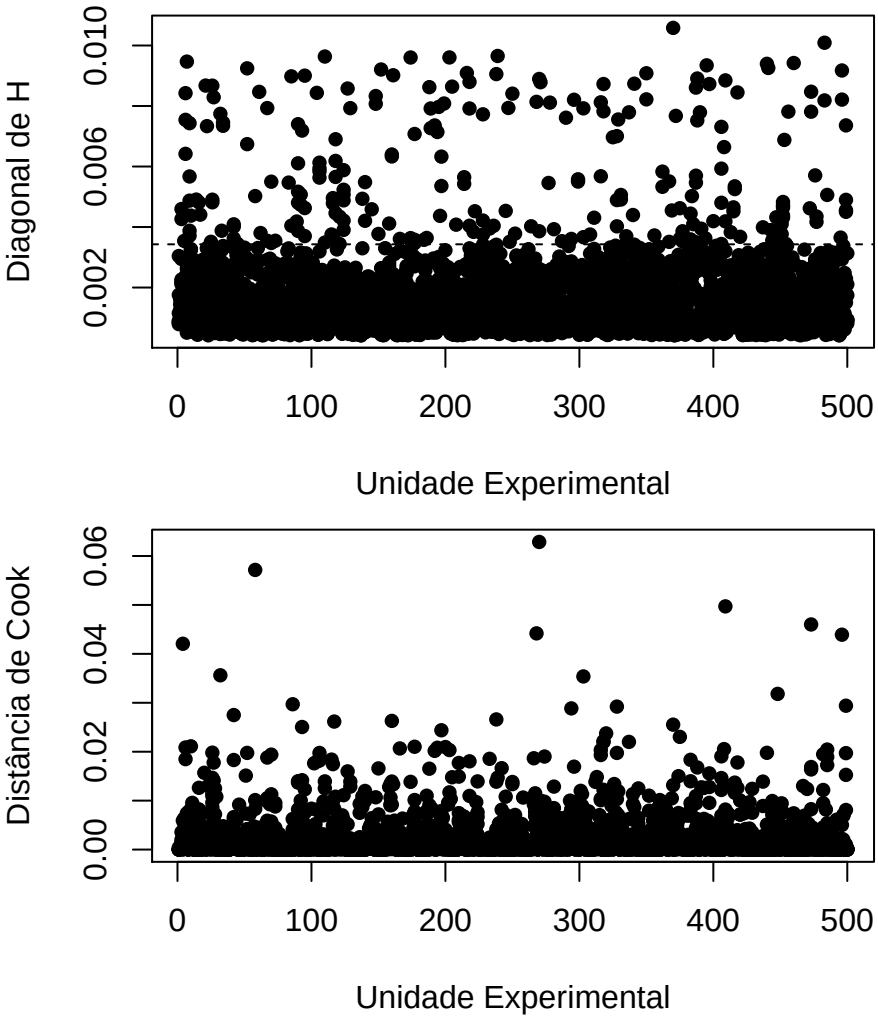
As estimativas dos parâmetros que já eram significativos não mudam de maneira brusca no modelo tentativo final. Os parêmetros de escala e de correlação não variam entre os ajustes.

1.2.1.2 Diagnóstico de modelo Partindo do modelo ajustado na seção anterior, iremos primeiramente realizar uma análise de pontos de alavanca e pontos influentes. Com relação aos pontos de alavancas, podemos ver no gráfico de índices dos pontos contra o valor da diagonal da matriz H correspondente, podemos ver que existem dois grupos de pontos. A faixa superior mostra que existe um número grande de observações remotas, porém não é algo focado em um grupo de pontos pequeno.

(#tab:unnamed-chunk-5)Seleção de modelos

Variável	fit1	fit2	fit3	fit4	fit5
(Intercept)	0.64 (1.25)	0.61 (1.2)	0.62 (1.24)	0.62 (1.24)	0.73 (1.51)
comorbidYes	-0.13 (-1.23)	-	-	-	-
genderMale	0.45 (2.28)*	0.46 (2.35)*	0.46 (2.33)*	0.45 (2.32)*	0.46 (2.37)*
inptmhv3> 1	0.27 (0.86)	0.28 (0.89)	0.29 (0.89)	-	-
inptmhv3>=1	-	-	-	0.35 (1.83)*	0.35 (1.82)*
inptmhv31	0.37 (1.68)*	0.37 (1.7)*	0.37 (1.7)*	-	-
log(age/100)	1.28 (5.96)*	1.28 (6)*	1.28 (6)*	1.28 (6)*	1.29 (6.02)*
mhv4> 14	0.89 (5.5)*	0.89 (5.58)*	0.9 (5.6)*	0.89 (5.59)*	0.78 (6.62)*
mhv41-5	0.13 (0.95)	0.14 (1.03)	0.15 (1.04)	0.15 (1.04)	-
mhv46-14	0.57 (3.76)*	0.59 (3.9)*	0.59 (3.91)*	0.59 (3.91)*	0.47 (4.73)*
urbanRural	0.05 (0.31)	0.05 (0.31)	-	-	-
escala	0.97	0.98	0.97	0.97	0.98
rho	0.58	0.58	0.58	0.58	0.58

Cada entrada consiste em estimativa (valor z robusto)
*: indica significância à 10% num teste de Wald com distribuição Chi-Quadrado



A influência dos pontos será medida através da distância de Cook. Utilizando uma regra de média + 3 vezes o desvio padrão das distâncias de cook das observações, podemos ver que temos um total de 77 indivíduos-ano que

(#tab:unnamed-chunk-8)Análise de influência

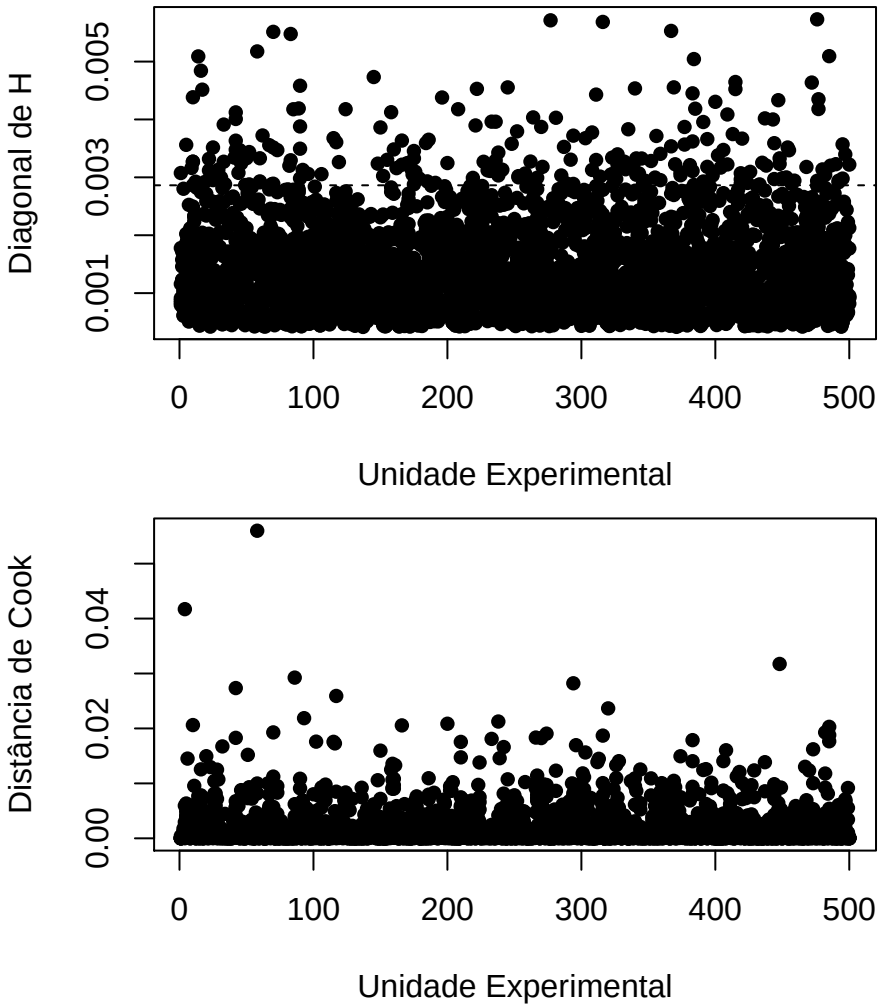
Variável	toda amostra	s/ 77 pontos influentes	toda amostra -inptmhv3	s/ 77 pontos influentes -inptmhv3
(Intercept)	0.73 (1.51)	1.03 (2.11)*	0.72 (1.5)	1.03 (2.13)*
genderMale	0.46 (2.37)*	0.64 (2.94)*	0.46 (2.36)*	0.64 (2.94)*
inptmhv3>=1	0.35 (1.82)*	0.17 (1.02)	-	-
log(age/100)	1.29 (6.02)*	1.5 (7.02)*	1.28 (6.02)*	1.5 (7.05)*
mhv4> 14	0.78 (6.62)*	0.73 (6.73)*	0.81 (7.06)*	0.74 (6.82)*
mhv46-14	0.47 (4.73)*	0.36 (3.98)*	0.48 (4.83)*	0.36 (3.99)*
escala	0.98	0.99	0.98	0.99
rho	0.58	0.65	0.58	0.65

Cada entrada consiste em estimativa (valor z robusto)
*: indica significância à 10% num teste de Wald com distribuição Chi-Quadrado

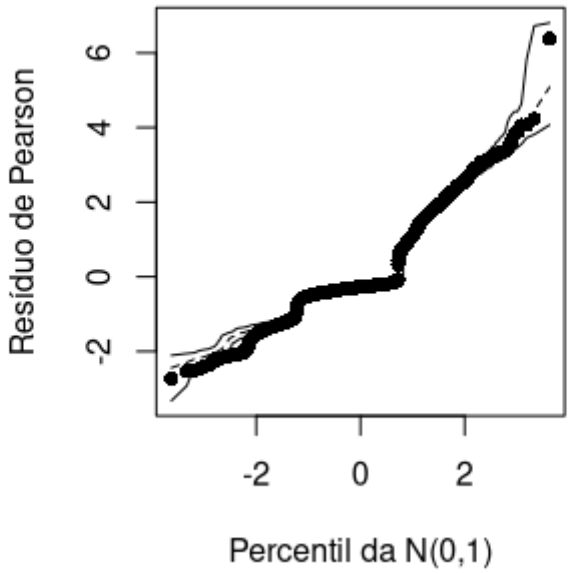
podem ser considerados influentes. Ao ajustarmos o modelo sem esses pontos, podemos ver que as conclusões inferenciais mudam pois a significância do fator `inptmhv3>=1` mudou, fazendo com que essa variável não seja mais significante ao nível de 10%. O ajuste dos modelos com a amostra inteira e sem os pontos influentes removendo essa variável não apresentam uma alteração relevante nas conclusões inferenciais do modelo.

```
di = diag_values$DistCook
cut = mean(di) + 3*sd(di)

fitgee_cook_cut = gee(resp ~ mhv4 + inptmhv3 + gender + log(age/100),
                id = id, family = binomial(link = "logit"),
                corstr = "AR-M", Mv = 1, data=df2[di < cut,])

fitgee2 = gee(resp ~ mhv4 + gender + log(age/100),
                id = id, family = binomial(link = "logit"),
                corstr = "AR-M", Mv = 1, data=df2 %>% filter(id!=237))

fitgee_cook_cut2 = gee(resp ~ mhv4 + gender + log(age/100),
                id = id, family = binomial(link = "logit"),
                corstr = "AR-M", Mv = 1, data=df2[di < cut,])

gee_cook_list <- list('toda amostra' = fitgee,
                      's/ 77 pontos influentes' = fitgee_cook_cut,
                      'toda amostra -inptmhv3' = fitgee2,
                      's/ 77 pontos influentes -inptmhv3' = fitgee_cook_cut2)
```

O modelo que iremos finalizar então consiste na amostra interira sem a variável `inptmhv3` uma vez que ela não se mostrou estável com relação à influência da amostra. Abaixo temos a análise de resíduos deste novo modelo. A análise do gráfico de envelope mostra que o ajuste está satisfatório com relação à normalidade dos resíduos padronizados de Pearson.



Um ponto importante sobre o programa que gerou o gráfico de envelope para o resíduo de Pearson foi necessário retirar o grupo de id=237 pois o cálculo do envelope sempre travava neste grupo. Além disso, o tempo para gerar o envelope foi muito alto, algo em torno de 2H, por isso não foi prático gerar uma visualização mais nítida do gráfico de probabilidades.



(#tab:unnamed-chunk-11)Estimativas intervalares para o modelo de EEG final

variável	estimativa	e.p. robusto	exp(estimativa)	lim. inf. p/ exp(estimativa)	lim. sup. p/ exp(estimativa)
(Intercept)	0.72	0.48	2.06	0.80	5.33
mhv46-14	0.48	0.10	1.61	1.33	1.96
mhv4> 14	0.81	0.11	2.25	1.80	2.82
genderMale	0.46	0.20	1.59	1.08	2.34
log(age/100)	1.28	0.21	3.59	2.37	5.44

1.2.1.3 Interpretação de parâmetros e estimativas intervalares

Na tabela a seguir mostramos as estimativas intervalares para as razões de chance estimadas pelo modelo resultante das análise nas sessões anteriores.

Eles têm as seguintes interpretações:

- `(Intercept)`: no caso o intercepto representa o caso de referência sob o qual todos os outros fatores serão comparados. Estes são um paciente do sexo feminino, com idade 0 (não realista) e que teve de 0 a 5 de consultas ambulatoriais relacionadas à saúde mental.
- `mhv46-14`: a razão de chances de polifarmácia entre um indivíduo que teve 6 a 14 consultas ambulatoriais com relação à um indivíduo que teve 0 a 5 de consultas é 1.61, ou seja a chance é de se usar 3 medicamentos ou mais é 60% maior no grupo que teve 6 a 14 consultas ambulatoriais.
- `mhv4> 14`: a razão de chances de seu usar 3 medicamentos ou mais do grupo que foi mais do que 14 vezes em consultas ambulatoriais no ano é 2.25, ou seja, 125% maior que o grupo com 0 a 5 consultas ambulatoriais no ano.
- `genderMale`: indivíduos do sexo masculino têm uma de chance de polifarmácia 59% maior que os do sexo feminino, indicato pela estimativa de razão de chances de 1.59.
- `log(age/100)`: para idade podemos chegar que o impacto do aumento em 1% da idade na razão de chances de polifarmácia de um indivíduo é 1.01^0.21 - 1= 0.209%.

1.2.2 Modelo Misto

1.2.2.1 Seleção de modelo

O processo de seleção do modelo misto seguiu de maneira similar ao modelo de equações de estimação generalizadas. A diferença está na possibilidade de testar a remoção de 2 ou mais parâmetros e testar sua auseência através de um teste de razão de verossimilhança. Abaixo mostro o código que utilizei para ajustar esses modelos e uma tabela mostrando as estimativas de cada modelo e testes revelevantes.

```
fit_mixed_1 = gamlss(resp ~ mhv4 + inptmhv3 + gender + urban + comorbid +
                       log(age/100) + random(as.factor(id)),
                     family = BI, data = df,
                     control = gamlss.control(trace = F))

fit_mixed_2 = gamlss(resp ~ mhv4 + inptmhv3 + gender +
                       log(age/100) + random(as.factor(id)),
                     family = BI, data = df,
                     control = gamlss.control(trace = F))

fit_mixed_3 = gamlss(resp ~ mhv4 + inptmhv3 + gender +
                       log(age/100) + random(as.factor(id)),
                     family = BI, data = df2,
                     control = gamlss.control(trace = F))
```

	fit1	fit2	fit3
μ (Intercept)	2.285*	2.284*	2.387*
	(0.517)	(0.516)	(0.484)
μ mhv41-5	0.162	0.154	
	(0.240)	(0.238)	
μ mhv46-14	0.850*	0.842*	0.721*
	(0.226)	(0.222)	(0.147)
μ mhv4> 14	1.260*	1.254*	1.125*
	(0.221)	(0.217)	(0.140)
μ inptmhv31	0.857*	0.855*	
	(0.245)	(0.245)	
μ inptmhv3> 1	0.349	0.348	
	(0.392)	(0.392)	
μ genderMale	0.624*	0.612*	0.626*
	(0.142)	(0.141)	(0.141)
μ urbanRural	0.026		
	(0.122)		
μ comorbidYes	0.091		
	(0.171)		
μ log(age/100)	2.555*	2.536*	2.537*
	(0.219)	(0.218)	(0.218)
μ inptmhv3>=1			0.729*
			(0.213)
Desvio	1751.66	1753.26	1753.48
AIC	2512.69	2509.37	2506.81
Teste RV*	-	1.598	0.226
Valor p	-	0.553	0.77
gl	-	2.461	1.394
Num. obs.	3500	3500	3500

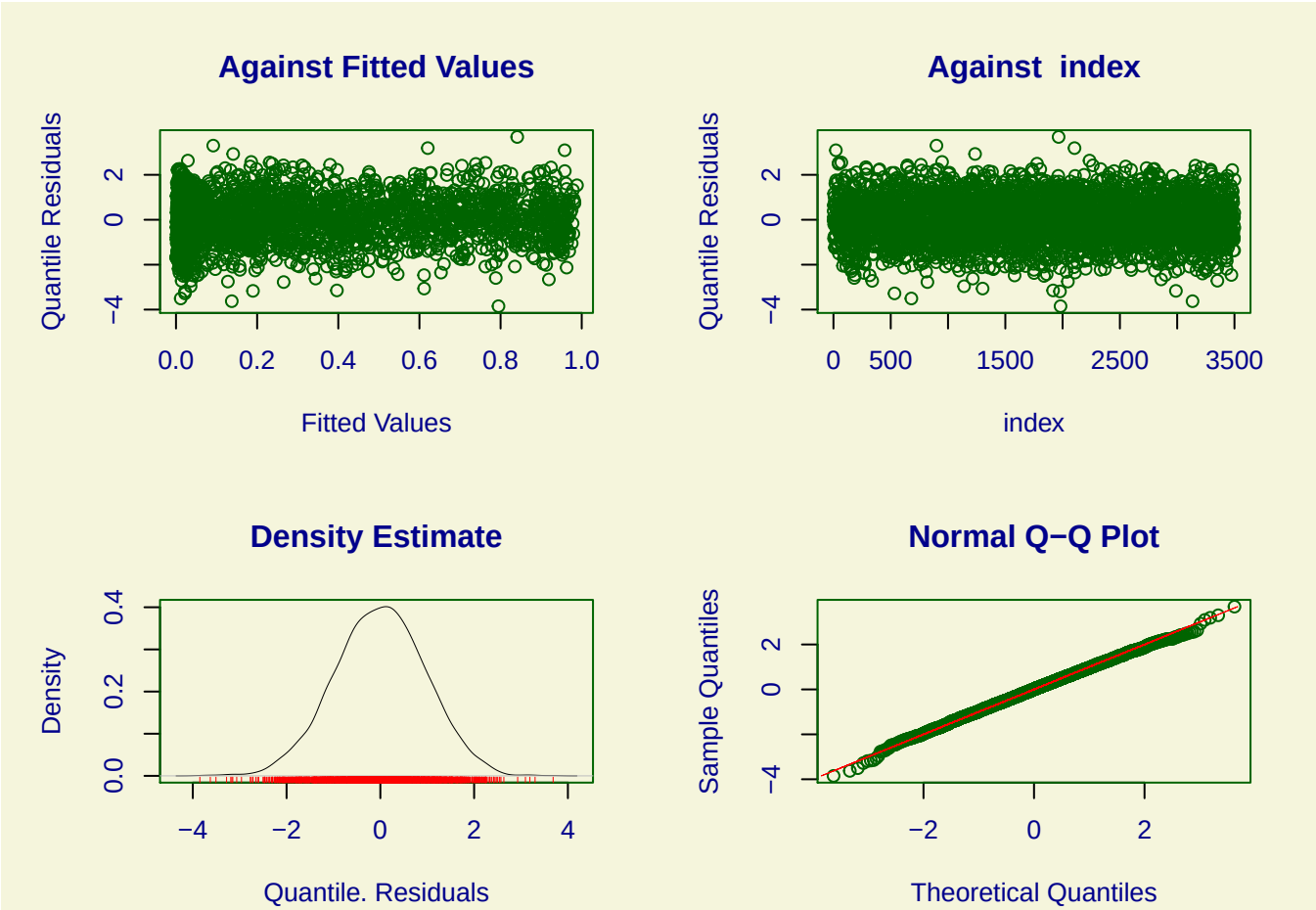
*: Teste de RV se refere ao modelo em questão contra o modelo anterior.

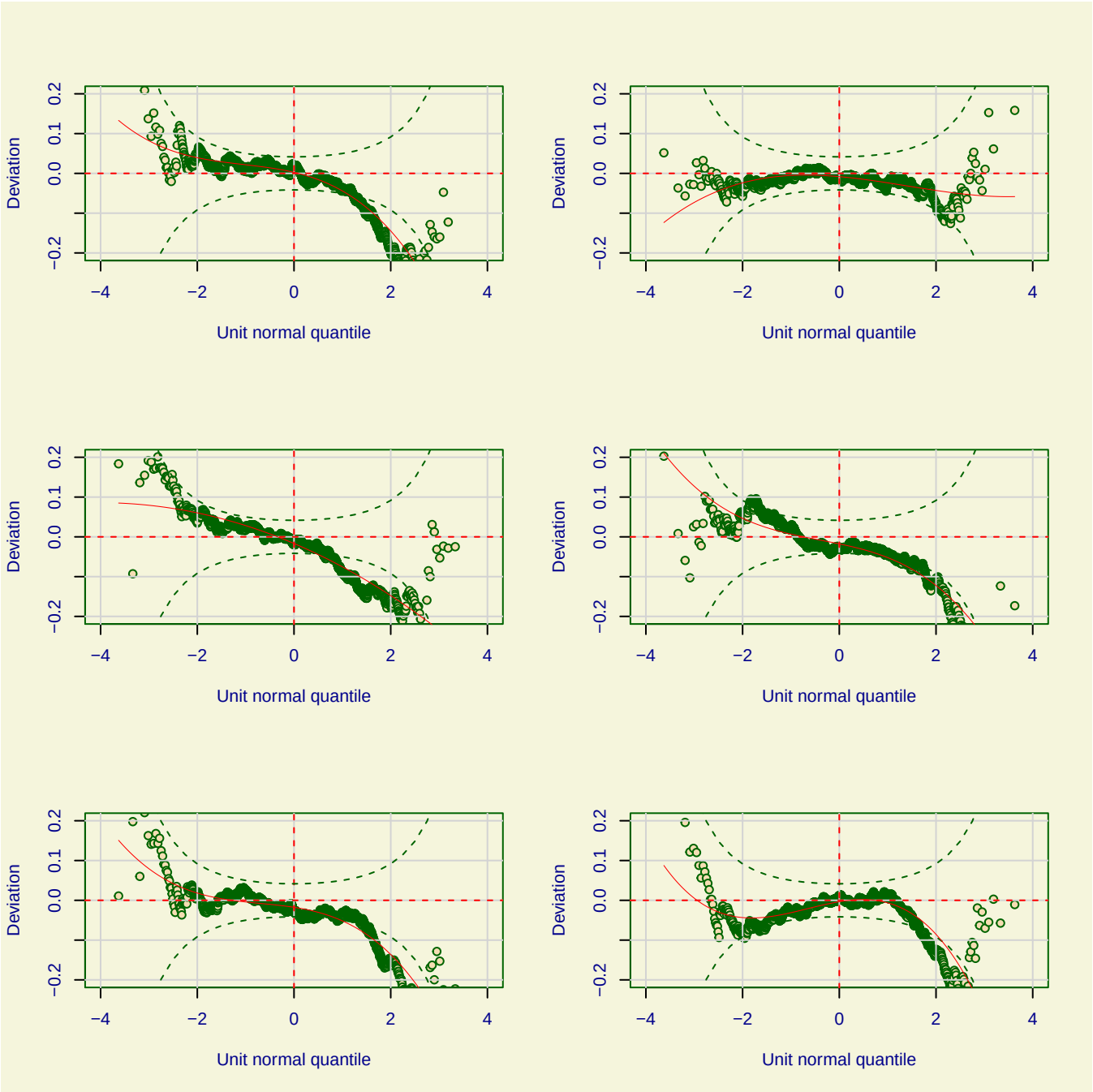
```
lr_2_v_1 <- LR.test(fit_mixed_2, fit_mixed_1, print = F)
lr_3_v_2 <- LR.test(fit_mixed_3, fit_mixed_2, print = F)

md_list <- list(fit1 = fit_mixed_1, fit2 = fit_mixed_2, fit3 = fit_mixed_3)
```

Como podemos ver, o modelo 3 é o que tem menor AIC e por isso vamos avançar com ele.

1.2.2.2 Análise de resíduos A seguir podemos checar a análise de resíduos para o modelo misto estimado via o `gamlss`. Alguns comentários sobre ele:

- o gráfico de valores ajustados contra os resíduos quantílicos indica que a variabilidade foi controlada, uma vez que não notamos padrões na dispersão e duas retar paralelas parecem colocar os pontos numa faixa uniforme.
- a inspeção do gráfico normal de probabilidades e da densidade estimada dos resíduso quantílicos indica que a normalidade parece ter sido alançada pelo ajuste.
- as réplicas do gráfico de wormplot indicam um ajuste insatisfatório, uma vez que há a sobreposição dos pontos com o exterior da região de confiança do gráfico.


Against Fitted Values
Quantile Residuals
−4
0
2
0.0
0.2
0.4
0.6
0.8
1.0
Fitted Values
Against index
Quantile Residuals
−4
0
2
0
500
1500
2500
3500
index
Density Estimate
Density
0.0
0.2
0.4
−4
−2
0
2
4
Quantile. Residuals
Normal Q−Q Plot
Sample Quantiles
−4
0
2
−2
0
2
Theoretical Quantiles



1.2.2.3 Interpretação de parâmetros e estimativas intervalares Abaixo temos as estimativas intervalares para o modelo misto e a interpretação dos parâmetros estimados.

Eles têm as seguintes interpretações:

- `(Intercept)`: no caso o intercepto representa o caso de referência sob o qual todos os outros fatores serão comparados. Estes são um paciente do sexo feminino, com idade 0 (não realista) e que teve de 0 a 5 de consultas ambulatoriais relacionadas à saúde mental sem nenhuma internação hospitalar no período.
- `mhv46-14`: a razão de chances de polifarmácia entre um indivíduo que teve 6 a 14 consultas ambulatoriais com relação à um indivíduo que teve 0 a 5 de consultas é 2.06, ou seja a chance é de se usar 3 medicamentos ou mais é 106% maior no grupo que teve 6 a 14 consultas ambulatoriais para o mesmo indivíduo.
- `mhv4> 14`: considerando o indivíduo fixo, a razão de chances de seu usar 3 medicamentos ou mais do grupo que foi mais do que 14 vezes em consultas ambulatoriais no ano é 3.08, ou seja, 208% maior que o grupo

(#tab:unnamed-chunk-16)Estimativas intervalares para o modelo misto final

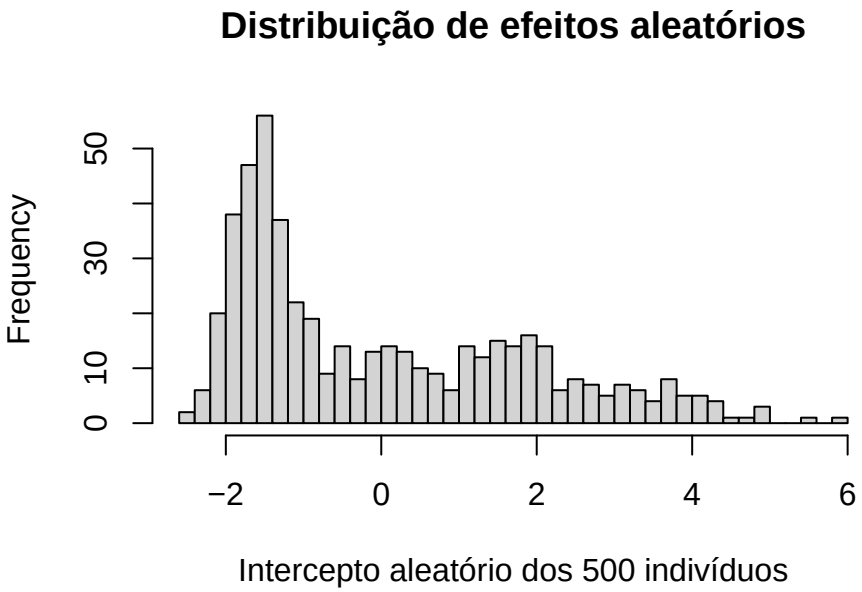
variável	estimativa	e.p.	exp(estimativa)	lim. inf. p/ exp(estimativa)	lim. sup. p/ exp(estimativa)
(Intercept)	2.39	0.48	10.88	4.22	28.07
mhv46-14	0.72	0.15	2.06	1.54	2.74
mhv4> 14	1.12	0.14	3.08	2.34	4.05
inptmhv3>=1	0.73	0.21	2.07	1.37	3.15
genderMale	0.63	0.14	1.87	1.42	2.46
log(age/100)	2.54	0.22	12.64	8.25	19.37

com 0 a 5 consultas ambulatoriais no ano.

- `inptmhv3>=1`: ao comparar um mesmo indíviduo que tenha tido 1 ou mais internações com o mesmo indivíduo no caso não tivesse internações, a razão de chances de polifarmácia é 2.07, ou seja 107% maior para esse indivíduo quando ele teve uma ou mais internações.
- `genderMale`: supondo que um indivíduo fixo possa trocar de sexo, indivíduos do sexo masculino têm uma de chance de polifarmácia 87% maior que os do sexo feminino, indicato pela estimativa de razão de chances de 1.87.
- `log(age/100)`: para idade podemos chegar que o impacto do aumento em 1% da idade na razão de chances de polifarmácia de um indivíduo (fixo) é 1.01^2.54 - 1= 2.56%.

1.2.2.4 Qual o papel do efeito aleatório neste tipo de modelo? O papel do efeito aleatório é, quando contrastado com a abordagem via as equações de estimação generalizadas, é propor uma estimativa condicional ao grupo de medidas repetidas em questão. A estimativa via EEG estima parâmetros à nivel populacional, tentando controlar a variabilidade por meio da estrutura de correlação intra grupo. Já a modelagem mista permite a possibilidade de inserir as características do grupo como co variáveis que estimam um efeito aleatório. Isso permite extrair a variação desses efeitos para o nível do grupo (id no caso deste exercício) e avaliar essas diferenças entre os grupos.

Neste exercício, a distribuição de efeitos aleatórios resultou no seguinte histograma:



Isto indica que alguns indivíduos tem uma probabilidade de polifarmácia naturalmente maior quando em comparação com outros, implicando que o resultado da associação entre as variáveis para um determinado

(#tab:unnamed-chunk-18)Comparação de chances entre 2 indíviduos com covaríaveis hipotéticamente idênticas

id	intercepto aleatório	intercepto	mhv46-14	log(age/100)	genderMale	chance	probabilidade
18	-1.94	2.39	0.72	2.54	0.63	0.64	0.39
217	3.14	2.39	0.72	2.54	0.63	2.12	0.68

indivíduo pode diferir para outro. Abaixo mostro, para dois indivívidos, supondo as covariáveis identicas (6 a 14 visitas ao ambulatorio, sexo masculino, 10 anos de idade e nenhuma internação), os parâmetros, a chance e probabilidade de polifarmácia. Entre esses dois indívidos, as chance são bem diferentes e estudar as associações pode mudar as conclusões de um estudo com base em qual individuo (e suas características) estamos focando.

2 Exercício 2

A base de dados deste exercício consiste em um experimento longitudinal desenvolvido na Austrália com 79 vacas que foram aleatorizadas segundo três dietas e foi observado semanalmente a quantidade de proteína no leite de cada animal. O objetivo é verificar as diferenças entre o conteúdo proteico semanal sob as três dietas. As variáveis presentes na base são:

- `protein`: quantidade de proteínas
- `time`: semana
- `Cow`: identificação do animal
- `Diet`: dieta utilizada sendo cevada (*barley*), cevada+tremoços (*barley+lupins*) e tremços (*lupins*)

Abaixo mostramos as 10 primeiras linhas do conjunto de dados:

protein	Time	Cow	Diet
3.63	1	B01	barley
3.57	2	B01	barley
3.47	3	B01	barley
3.65	4	B01	barley
3.89	5	B01	barley
3.73	6	B01	barley
3.77	7	B01	barley
3.90	8	B01	barley
3.78	9	B01	barley
3.82	10	B01	barley

2.1 Análise descritiva

O primeiro passo para esta análise é a inspeção da variável resposta protein. Abaixo temos algumas estatísticas descritivas para a amostra em nosso poder. Para todas as dietas e para a amostra inteira, podemos ver que a média e a mediana de proteínas é bem parecida, indicando que pode não haver uma assimentria acentuada neste conjunto de dados. A comparação entre os quartis 1 e 3 junto com os desvios padrões indica que não há uma diferença entre a variação entre esses grupos.

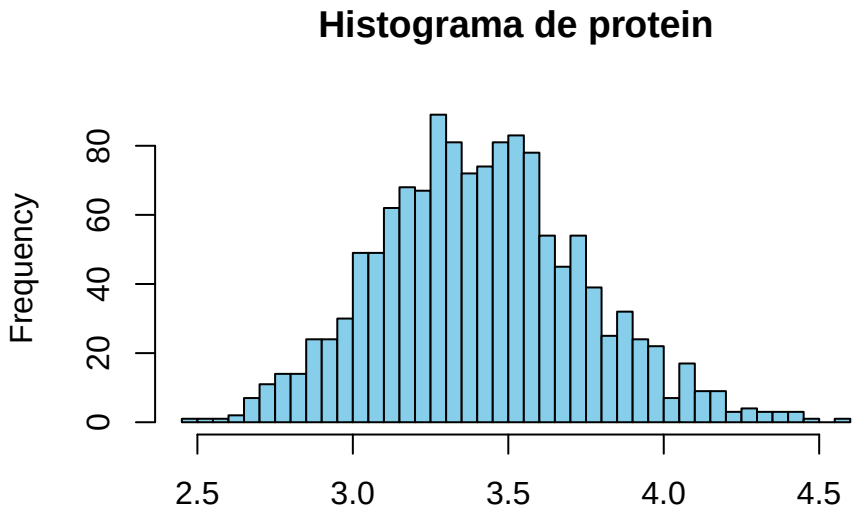
Outro ponto importante de se observar é que a análise das estatísticas descritivas mostra que o conjunto de dados com a diaeta de cevada (*barley*) parece ter a maior quantidade de proteínas, seguido por cevada+tremoços (*barley+lupins*) e por último tremoços (*lupins*), tanto em média quanto em mediana. [

O gráfico de histograma abaixo e as densidades confirmam algumas das sugestões provindas das análises descritivas. A variável resposta parece não ser muito assimétrica e a as densidades parecem indicar a mesma ordem de

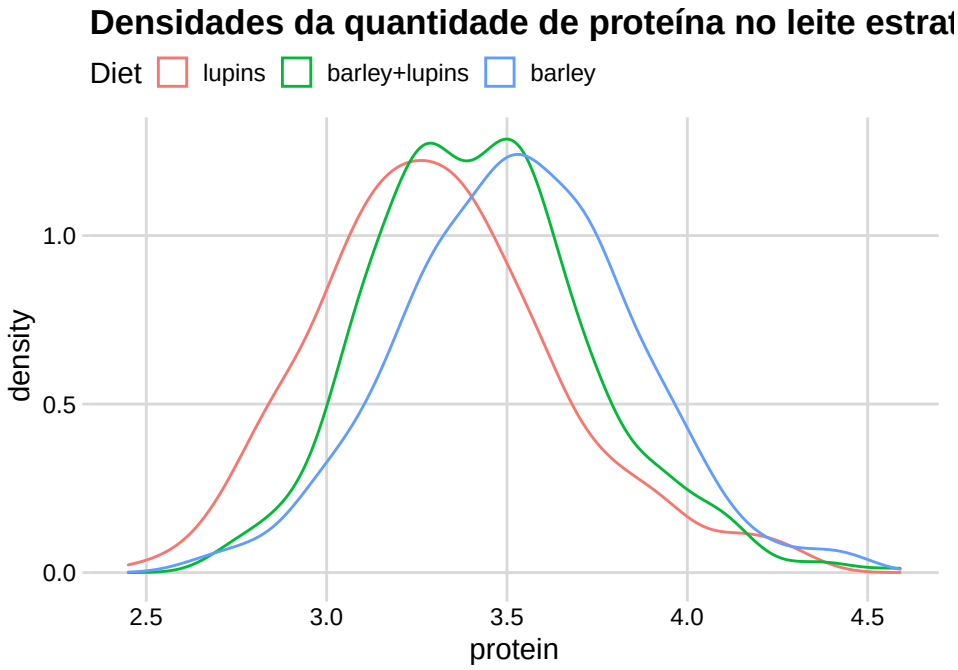
(#tab:unnamed-chunk-20)Estatísticas descritivas estratificadas por dieta

Diet	média	mediana	Quartil 1	Quartil 3	Desv. Pad.	Coef. Var.
barley	3.53	3.53	3.32	3.74	0.32	0.09
barley+lupins	3.43	3.42	3.21	3.60	0.30	0.09
lupins	3.31	3.29	3.08	3.51	0.34	0.10
toda amostra	3.42	3.41	3.20	3.63	0.33	0.10

localização dietas com relação à quantidade proteica no leite. Curiosamente, a densidade da estratificação da dieta por cevada+tremoços (*barley+lupins*) parece ter uma ligeira bimodalidade.

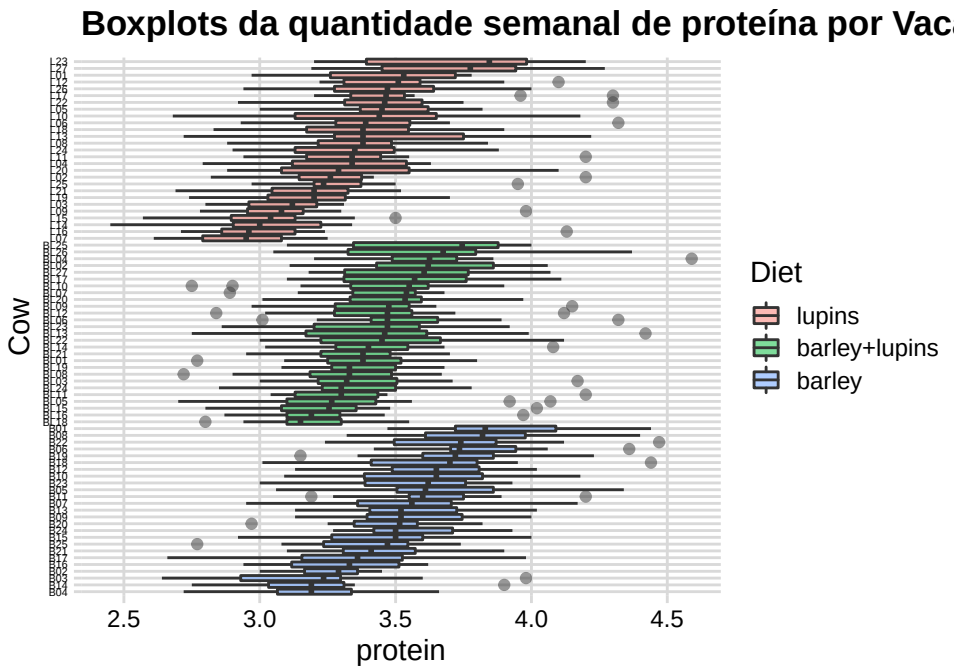


.

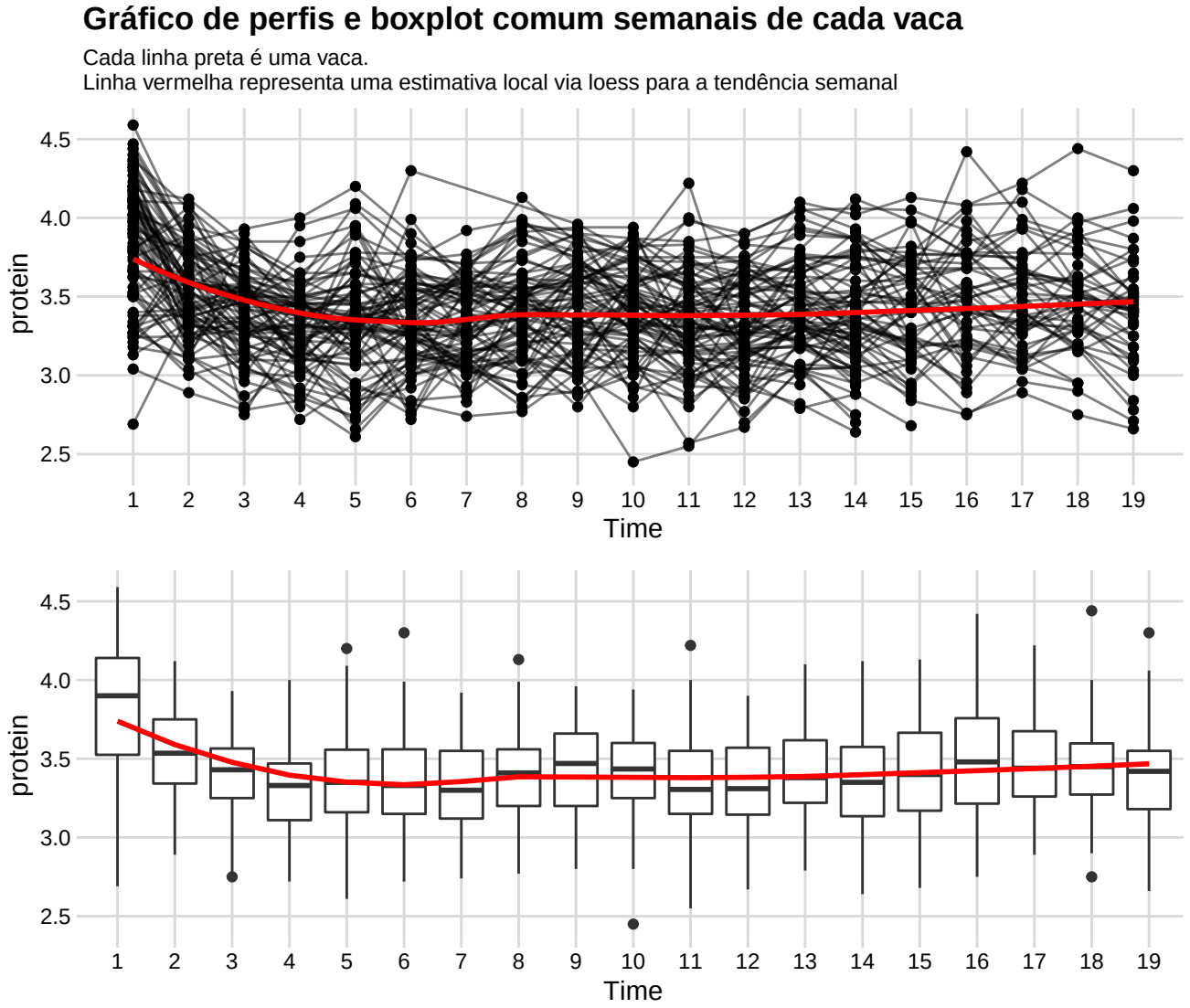


Quando analisamos as medidas semanais por cada vaca por meio de um boxplot comum, podemos ver que existe uma variação relevante entre as vacas. Por isso as recomendações de estimar o efeito da dieta levando em conta a vaca em questão, via EEG ou modelo misto faz muito sentido para esse conjunto de dados. Algumas vacas tem medidas atípicas tanto inferiores quanto superiores em relação ao intervalo interquartil do boxplot comum. Todavia, o grupo da dieta de cevada+tremoços (*barley+lupins*) parece ter mais pontos atípicos que as outra

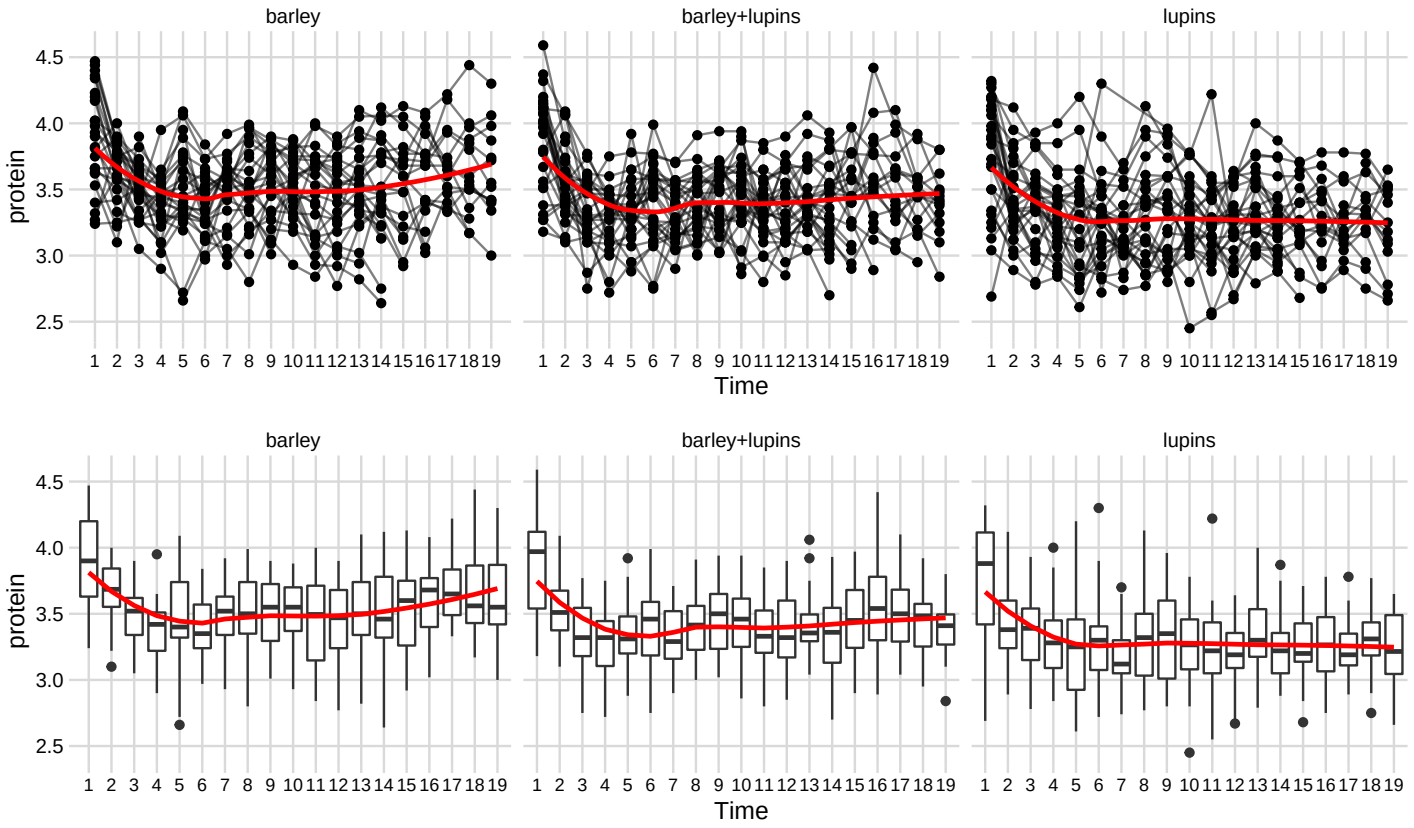
dietas. A dieta de tremoços (*lupins*) só tem observações atípicamente mais altas.



Analisando os perfis semanais da quantidade proteíca no leite, podemos ver, com o auxílio da estimativa loess e pelos boxplots semanais, que no começo do estudo a quantidade de proteina parecia ser consistentemente maior. Isso parece indicar a necessidade de uma modelagem não linear para a variável tempo.



Abaixo temos a estratificação do gráfico de perfis e boxplots semanais pela dieta da vaca. A ideia dessa visualização é tentar identidicar padrões diferentes na evolução das semanas do experimento e tentar checar se existe alguma diferença importante entre as dietas na evolução semanal. Talvez a maior diferença que é possível notar, tirando os níveis de localização já evidenciados nas outras análises, seja uma leve tendência crescente na evolução semanal das vacas que receberam cevada (*barley*) como dieta no fim do experimento. Isso pode indicar uma possível interação essencial entre as variáveis dieta e tempo.



2.2 Modelagem dos dados - comparação de modelagem mista e de equação de estimação generalizadas

Nesta fase da análise iremos ajustar dois modelos, um de equação de estimação generalizada e outro misto para comparar os ajustes e testar a interação entre as semanas do estudo e a e a dieta da vaca.

- modelo de equações de estimação generalizado

$$Y_{ijk}|\text{Semana=j, Dieta=k,} \sim Q(\pi_{ijk}; y_{ijk}), \;\; 0 < \pi_{ijk} < 1, \;\; Var(Y_{ijk}) = \sigma^2 \pi_{ijk}(1 - \pi_{ijk})$$

$$Cor(Y_{ijk}, Yij'k) = \rho^{|j-j'|} \quad \text{ou} \quad 0 \quad \text{c.c.}$$

$$\log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \alpha + \delta_k + \beta_2 \cdot j + \beta_3 \cdot j^2 + \Delta_k \cdot j$$

Onde Y_{ijk} representa a quantidade de proteína no leite da vaca i, na semana do estudo j e que recebeu a ração k. \$i $= 1, \ldots$,79 \$ representam as vaca, $j = 1, ..., 19$ as semanas, $k = 1, 2, 3$ as dietas sendo 1 a dieta de cevada (*barley*) a referência que será incorporada no intercepto, ou seja $\delta_1 = 0$, k=2 é a dieta cevada+tremoços (*barley+lupins*) e k=3 a dieta de tremoços (*lupins*). Por fim Δ_k representa a interação entre a dieta da vaca e a semana do estudo.

- modelo misto

$$Y_{ijk}|b_i, \text{Semana=j, Dieta=k,} \overset{ind}{\sim} \text{Bernoulli}(\pi_{ij})$$

$$\log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \alpha + b_i + \delta_k + \beta_2 \cdot j + \beta_3 \cdot j^2 + \Delta_k \cdot j$$
$$b_i \overset{iid}{\sim} \text{Normal}(0, \sigma_b^2)$$

Onde Y_{ijk} representa a quantidade de proteína no leite da vaca i, na semana do estudo j e que recebeu a ração k. Os \$i = 1, . . . ,79 \$ representam as vaca, $j = 1, ..., 19$ as semanas, $k = 1, 2, 3$ as dietas sendo 1 a dieta de cevada (*barley*) a referência que será incorporada no intercepto, ou seja $\delta_1 = 0$, k=2 é a dieta cevada+tremoços (*barley+lupins*) e k=3 a dieta de tremoços (*lupins*). Está presente no modele também um efeito aleatório b_i para cada vaca. Por fim Δ_k representa a interação entre a dieta da vaca e a semana do estudo.

Para o ajuste dos modelos, devido às dificuldades no uso dos procedimentos de diagnóstico providos no enunciado para as EEG, partimos para o uso do pacote `geepack` que possui algumas funcionalidades que usaremos, nomeadamente procedimento ANOVA e geração de resíduos. Abaixo estão os código de ajuste desses modelos.

```
library(geepack)

df <- df %>%
mutate(Cow = as.character(Cow) %>% as.factor(.))


fit_milk_gee1 = geeglm(protein ~ Diet + Time + I(Time^2),
                       id = Cow,
                       family = Gamma(link = "log"),
                       corstr = "ar1", data=df)

fit_milk_gee2 = geeglm(protein ~ Diet + Time + I(Time^2) +  Time:Diet,
                        id = Cow,
                        family = Gamma(link = "log"),
                        corstr = "ar1",
                        data=df)


fit_milk_mix1 = gamlss(formula = protein ~ Diet + Time + I(Time^2) + random(as.factor(Cow)),
                        family = GA,
                        data=df,
                        control = gamlss.control(trace = F))

fit_milk_mix2 = gamlss(formula = protein ~ Diet + Time + I(Time^2) + Time:Diet
                        + random(as.factor(Cow)),
                        family = GA,
                        data=df,
                        control = gamlss.control(trace = F))
```

A tabela abaixo mostra o resultado dos ajustes. A conclusão sobre inclusão da interação é diferente para cada modelo. No modelo misto, quase todos os efeitos principais continuam com conclusões inferenciais similares, mas um dos novos parâmetros não é siginificante. Para o modelo de EEG temos que dois dos efeitos principais de dieta se tornaram não siginificantes com a inclusão das interações, porém só um dos parâmetros de interação é siginificante à um nível de 10%. Para fazer uma análise mais precisa vamos fazer um teste de modelo encaixado para cada um dos tipos de modelo. No caso do modelo misto será um teste de razão de verossimilhança e no caso do GEE será um teste de Wald.

As conclusões para os testes são diferentes em cada modelo. No modelo misto, se estivermos dispostos à um nível de significância de 5%, a inclusão da interação parece ser corroborada pelo ajuste. Já no caso do modelo de

	ef. princ. EEG	+ int EEG	ef. princ. misto	+ int misto
(Intercept)	1.3728**	1.3560**		
	(0.0152)	(0.0169)		
Dietbarley+lupins	−0.0283*	−0.0124		
	(0.0142)	(0.0205)		
Dietlupins	−0.0624**	−0.0307		
	(0.0160)	(0.0240)		
Time	−0.0237**	−0.0218**		
	(0.0026)	(0.0028)		
Time^2	0.0009**	0.0009**		
	(0.0001)	(0.0001)		
Dietbarley+lupins:Time		−0.0017		
		(0.0018)		
Dietlupins:Time		−0.0035		
		(0.0021)		
μ (Intercept)			1.3266**	1.3180**
			(0.0084)	(0.0099)
μ Dietbarley+lupins			−0.0269**	−0.0241*
			(0.0050)	(0.0102)
μ Dietlupins			−0.0604**	−0.0385**
			(0.0050)	(0.0102)
μ Time			−0.0174**	−0.0164**
			(0.0020)	(0.0021)
μ Time^2			0.0008**	0.0008**
			(0.0001)	(0.0001)
σ (Intercept)			−2.5977**	−2.5999**
			(0.0260)	(0.0261)
μ Dietbarley+lupins:Time				−0.0003
				(0.0010)
μ Dietlupins:Time				−0.0024*
				(0.0010)
Scale parameter: gamma	0.0086	0.0086		
Scale parameter: SE	0.0005	0.0004		
Correlation parameter: alpha	0.7279	0.7252		
Correlation parameter: SE	0.0312	0.0309		
Num. obs.	1337	1337	1337	1337
Num. clust.	79	79		

$^{**}p < 0.01$; $^{*}p < 0.05$

Statistical models

EEG não é possível rejeitar a hipótese nula que corresponde ao modelo sem interação.

```
LR.test(fit_milk_mix1,fit_milk_mix2)
```

```
##  Likelihood Ratio Test for nested GAMLSS models.
##  (No check whether the models are nested is performed).
##
##        Null model: deviance= 127.9979 with  74.80504 deg. of freedom
##  Altenative model: deviance= 121.9624 with  76.72389 deg. of freedom
##
##  LRT = 6.035428 with 1.918848 deg. of freedom and p-value= 0.04514493
```

```
anova(fit_milk_gee1,fit_milk_gee2)
```

Df	X2	P(>|Chi|)
2	2.84493	0.241119

2.3 Comparação de ajustes

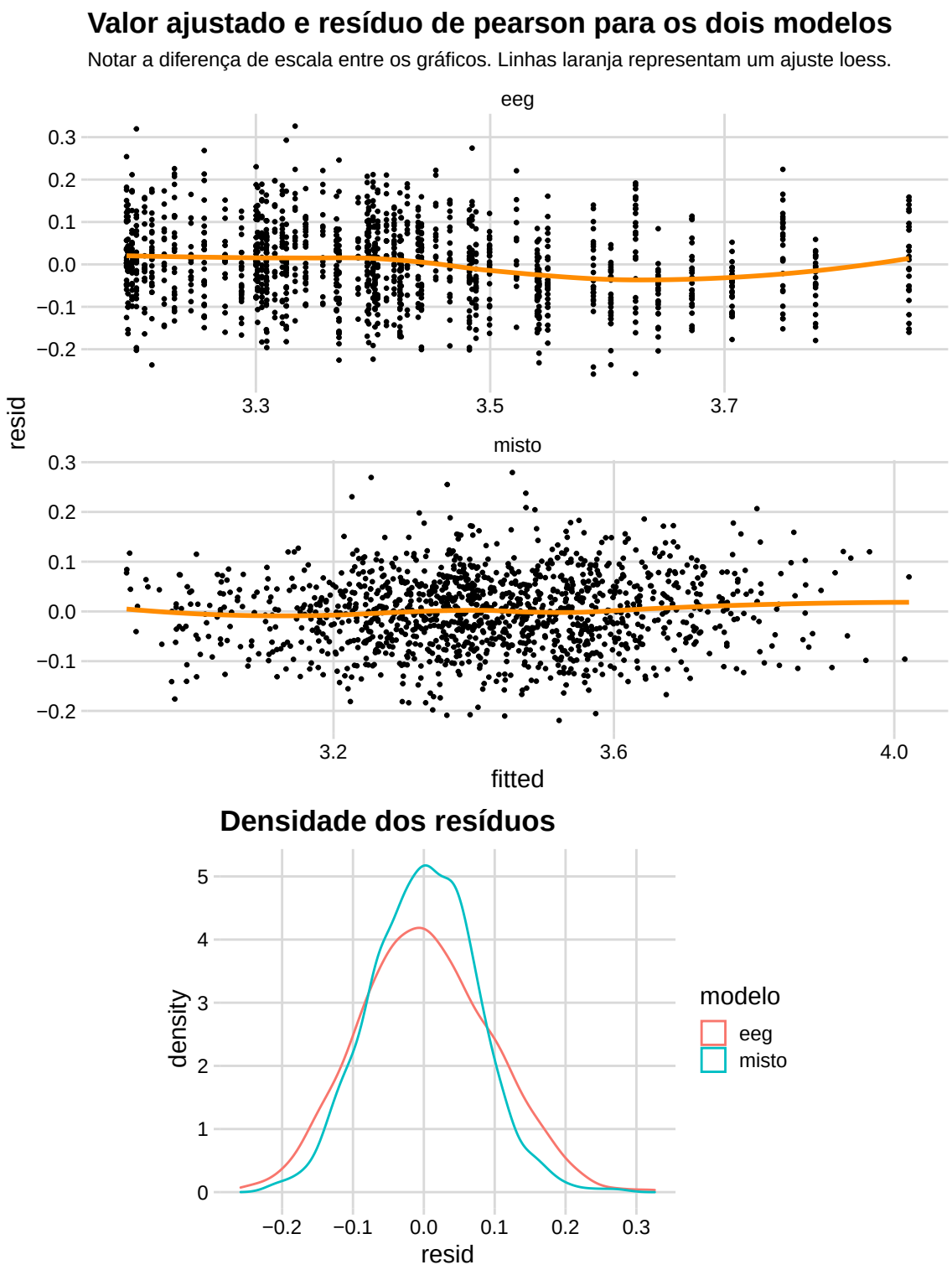
Vamos então comparar os modelos que culminaram da análise anterior, isto é o modelo misto com interação e o modelo de EEG sem a interação. Para possibilitar a comparação dos resíduos na mesma escala, vamos utilizar em ambos o resíduo de pearson para inspecioná-los.

```
md_mix <- fit_milk_mix2
md_gee <- fit_milk_gee1
```

```
md_res <- data.frame(
  fitted__misto = fitted(md_mix),
  fitted__eeg = fitted(md_gee),
  #resid__misto = resid(md_mix),
  resid__misto = (md_mix$y-fitted(md_mix) )*(1/sqrt(fitted(md_mix)^2)),
  resid__eeg = (md_gee$y-fitted(md_gee) )*(1/sqrt(fitted(md_gee)^2))
)
```

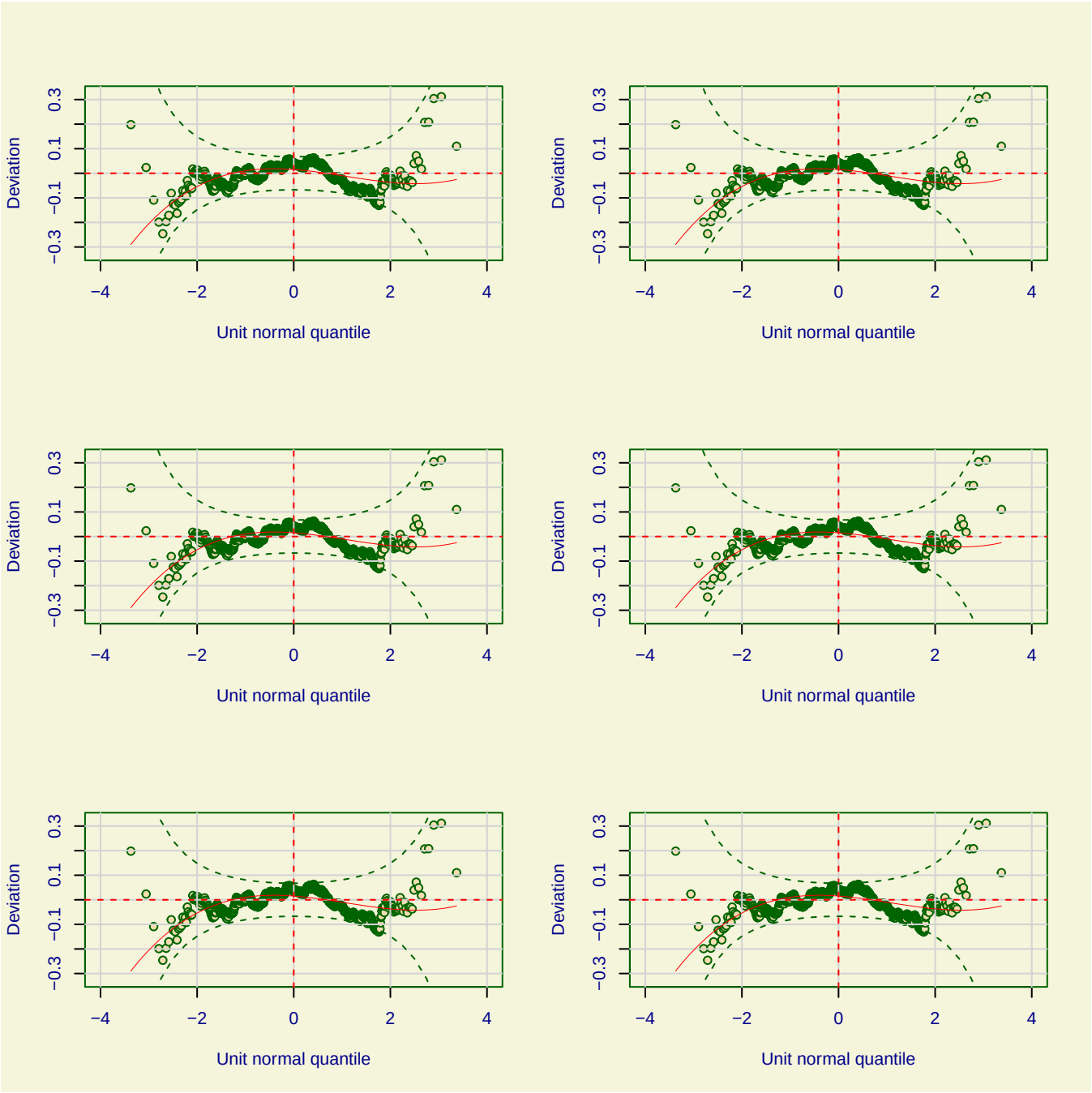
Os gráficos a seguir apresentam uma análise de resíduos para ambos os modelos. Alguns comentários sobre eles:

- devido às diferenças estruturais entre os modelos, a escala dos resíduos parecem ser bem diferentes.
- pela EEG se caracterizar por um modelo marginal, seus valores ajustados correspondem à uma "média" sobre todas os grupos levando em consideração uma estrutura de correlação interna e dividindo a variabilidade entre o parâmetro de escala e os parâmtros da matriz trabalho, fazendo com que suas estimativas sejam à nível médio, desconsiderando grupos não idênticos. Isto parece fazer com que haja menos variabilidade das estimativas.
- o ajuste do modelo misto parece ser apropriado, porêm com uma leve tendência de heterocedasticidade no início superior do gráfico.
- o wormplot do modelo misto, este sendo o sobre o resíduo quantílico, também indica um bom ajuste uma vez que não há sobreposição com a região externa das extremidades de confianca.
- as densidades dos resíduos parecem sugerir que os resíduos dos modelo de EEG é mais concentrado que os do modelo misto.




```
gamlss::rqres.plot(md_mix, howmany = 6, main = 'skjkfdsdkjf')
```

2.3.0.1 Wormplot para o modelo Misto



2.4 Comentários sobre os resultados

A diferença na construção teórica de ambos os modelos parece estar presente na análise dos resíduos. Como os modelos EEG têm a proposta de ser um modelo marginal, sendo, informalmente, uma média ao longo de grupo potencialmente heterogêneos, isso aparentemente se traduz na inspreção dos resíduos.

Já o modelo misto propõe inserir certo protagonismo aos grupos da análise ao inserir efeitos aleatórios que variam entre eles por meio de uma estrutura probabilística condicional à esses efeitos. Um ponto relevante consiste no fato de que esses efeitos aleatórios induzem necessariamente correlações positivas no nível do grupo. Neste sentido, o modelo condicional misto permite a individualização das estimativas de efeitos, uma vez que a estrutura probabilistica permite isso.

Existem então vantagens e desvantagens à cada um desses modelos. Como vantagens podemos elencar que a abordagem via EEG marginal parece ser mais direta quanto à investigação de associações populacionais e a abordagem mista condicional permite a avaliação para grupos específicos destas associações. A imposição

de estruturas de correlação positivas pode ser uma limitação aos modelos mistos. No caso de EEG por serem baseados em abordagens de quasi-verossimelhanças, ferramentas baseadas em verossimilhança tradicional precisam ser adaptadas para o seu uso. Uma outra vantagem para os modelos EEG é presença natural de um parâmetro de disperção que pode possibilitar sobre ou sub dispersão, isso também devido à sua construção via quasi-verossimilhança.

3 Exercício 3

3.1 Cálculo da Variância

O primeiro passo para entender a relação entre o poder de se testar $H_0 : \mu_1 - \mu_2 = 0$ contra $H_1 : \mu_1 - \mu_2 \neq 0$ através da estatística de Wald ξ_W proposta é o cálculo da variância da diferença entra $\bar{Y}_1 - \bar{Y}_2$:

$$\text{Var}(\bar{Y}_1 - \bar{Y}_2) = \text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_2) - 2\text{Cov}(\bar{Y}_1, \bar{Y}_2)$$

A variância da média amostral do grupo $j = 1, 2$ é:

$$\text{Var}(\bar{Y}_j) \overset{ind}{=} \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(Y_{ij}) = \frac{n\phi^{-1}V_j}{n^2} = \frac{\phi^{-1}V_j}{n}$$

Onde V_j é a função de variância da família exponencial sob a qual Y_{ij} pertence.

A covariância entre as médias amostrais pode ser simplificada com as seguintes manipulações:

$$\begin{aligned}
\text{Cov}(\bar{Y}_1, \bar{Y}_2) &= \text{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_{i1}, \frac{1}{n}\sum_{k=1}^{n} y_{k2}\right) \\
&= \frac{1}{n^2}\text{Cov}\left(\sum_{i=1}^{n} y_{i1}, \sum_{k=1}^{n} y_{k2}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n} \text{Cov}\left(y_{i1}, y_{k2}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Cov}\left(y_{i1}, y_{i2}\right)
\end{aligned}$$

Onde a última igualdade parte do fato de que a correlação é nula quando $i \neq k$ para $i, k = 1, ..., n$, conforme a construção no enunciado do exercício. Note que existe uma relação entre covariância e correlação:

$$\text{Corr}\left(y_{i1}, y_{i2}\right) = \rho = \frac{\text{Cov}\left(y_{i1}, y_{i2}\right)}{\sqrt{\text{Var}(\bar{y}_{i1})\text{Var}(\bar{y}_{i2})}}$$

E então, $\text{Cov}\left(y_{i1}, y_{i2}\right) = \rho\sqrt{\text{Var}(\bar{y}_{i1})\text{Var}(\bar{y}_{i2})} = \rho\phi^{-1}\sqrt{V_1 V_2}$. Concluindo então que

$$\begin{aligned}
\text{Cov}(\bar{Y}_1, \bar{Y}_2) &= \frac{1}{n^2}\sum_{i=1}^{n} \text{Cov}\left(y_{i1}, y_{i2}\right) \\
&= \frac{1}{n^2} n\rho\phi^{-1}\sqrt{V_1 V_2} \\
&= \rho n^{-1}\phi^{-1}\sqrt{V_1 V_2}
\end{aligned}$$

Em conclusão, temos que

$$\begin{aligned}\mathrm{Var}(\bar{Y}_1 - \bar{Y}_2) &= \frac{\phi^{-1}V_1}{n} + \frac{\phi^{-1}V_2}{n} - 2\rho n^{-1}\phi^{-1}\sqrt{V_1 V_2} \\ &= n^{-1}\phi^{-1}\left(V_1 + V_2 - 2\rho\sqrt{V_1 V_2}\right)\end{aligned}$$

3.2 Estudo do poder com relação à ρ

Utilizando o resultado da seção anterior, o parâmetro de não centralidade da distribuição de ξ_W sob H_1 : $\mu_1 - \mu_2 \neq 0$ é:

$$\lambda = \frac{n\phi(\mu_1 - \mu_2)^2}{2\left(V_1 + V_2 - 2\rho\sqrt{V_1 V_2}\right)}$$

Como λ representa o parâmetro de não centralidade, valores maiores de λ implicam uma maior probabilididade de rejeitar a hipótese nula quando esta é falsa. Na notação do exercício, e definindo $\lambda_1 > \lambda_2$, temos que $P_n(\lambda_1, \rho) > P_n(\lambda_2, \rho)$. Vamos agora estudar como λ muda com ρ.

Definindo ρ_1 e ρ_2 tal que $\rho_1 > \rho_2$ e respeitando $-1 \leq \rho_l \leq 1$ para $l = 1, 2$, temos que $\lambda_{\rho_1} > \lambda_{\rho_1}$ uma vez que ρ se encontra no denominador de λ. Assim para n, $\mu_1 - \mu_2$ e ϕ fixos, temos:

$$P_n(\lambda, \rho_1) > P_n(\lambda, \rho_2)$$

Por fim, não é difícil encontrar $n_2 > n_1$ de tal forma que:

$$P_{n_1}(\lambda, \rho_1) \leq P_{n_2}(\lambda, \rho_2)$$

Intuitivamente, o que esses resultados mostram é que, quanto maior a correlação de um par de amostras i, mais fácil é detectar uma diferença, uma vez que há mais informação transferida pela dependência entre o par de amostras. Se eu não tenho muita informação, ou seja, a correlação é baixa ou até negativa, o aumento causado na variação das estimativas deverá ser sopesado por uma amostra maior.