SME0823 - Modelos Lineares Generalizados - Lista 5

Danilo Augusto Ganancin Faria – Nº USP: 9609172

20 de dezembro de 2019

Resposta Positiva

Exercício 3

Os dados deste exercício correspondem a um estudo sobre a atividade das frotas pesqueiras de espinhel de fundo baseadas nas cidades de Santos e Ubatuba no litoral paulista.

O espinhel de fundo é definido como um método de pesca passivo. É um dos métodos que mais satisfazem às premissas da pesca responsável, com alta seletividade de espécies e comprimentos, alta qualidade do pescado, consumo de energia baixo e pouco impacto sobre o fundo do oceano.

A espécie de peixe considerada é o peixe-batata pela sua importância comercial e ampla distribuição espacial.

Uma amostra de $n = 156$ embarcações foi analisada, encontrou-se as seguintes variáves: `frota` (Santos ou Ubatuba), `ano` (95 a 99), `trimestre` (1 ao 4), `latitude` (de 23,25º a 28,25º), `longitude` (41,25º a 50,75º), `dias de pesca`, `captura` (quantidade de peixes batata capturados, em kg) e `cpue` (captura por unidade de esforço, Kg/dias de pesca), que é a variável resposta.

Para realizar a leitura dos dados utilizou-se o seguinte código:

```
# Código para ler os dados
dados <- read.table("pesca.txt", header = TRUE)
dados <- data.frame(dados)
```

Análise de dados preliminar

Inicialmente é apresentada uma análise descritiva dos dados para em seguida propor um modelo MLG a fim de tentar explicar a `cpue` média pelas variáveis explicativas.

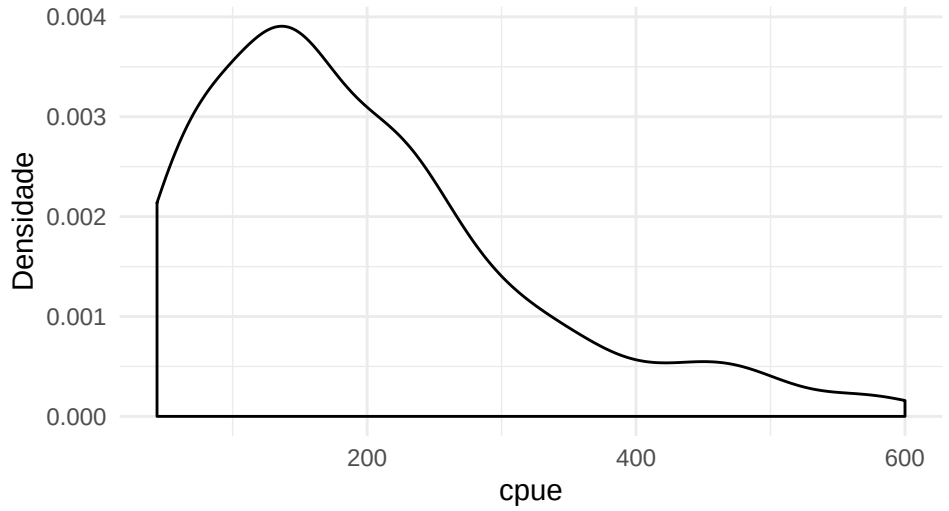

Figura 1: Densidade aproximada da cpue para todas as embarcações

É razoável fazer a suposição de que a distribuição da `cpue` possui assimetria à direita, conforme é ilustrado na Figura 1.

```
# Estatísticas descritivas das variáveis dias pesca, captura e cpue
summary(dados)
```

Tabela 1: Medidas resumo das variáveis diaspesca, captura e cpue

Variável	Min.	Med.	Máx.	Média	DP	Var.
diaspesca	1	9	18	8,397	3,6	12,963
captura	50	1200	6500	1623	1227,99	1507963
cpue	43,75	166,41	600	195,55	121,063	14656,17

Na Tabela 1 encontram-se algumas estatísticas descritivas para as variáveis `diaspesca`, `captura` e `cpue`. Nota-se que as embarcações passaram de 1 a 18 dias pescando porém, em média eles pescaram durante 4 dias. A quantidade média de peixes capturados foi de 1623 Kg. A `cpue` média foi de 121,03 Kg/dia de pesca.

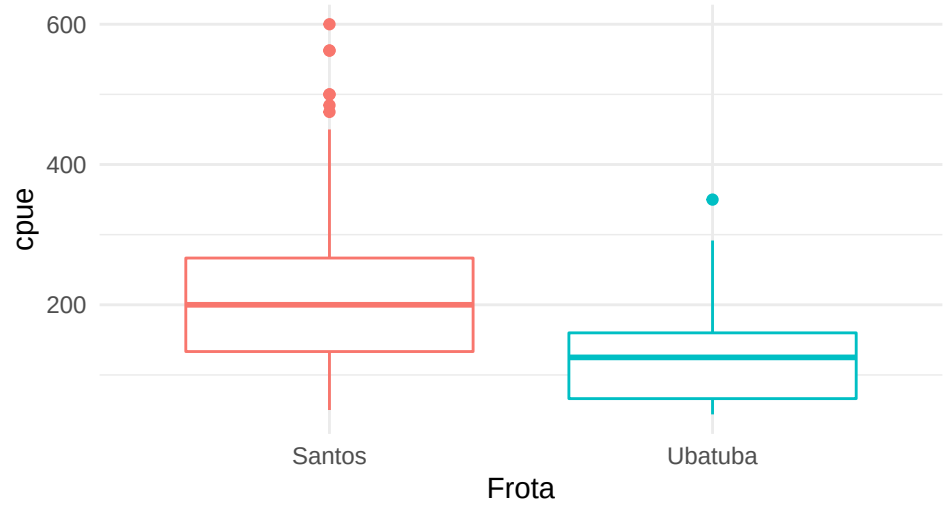

Figura 2: Gráficos de caixa da cpue pela frota

De acordo com a Figura 2, nota-se que a frota de Santos é superior à frota de Ubatuba.

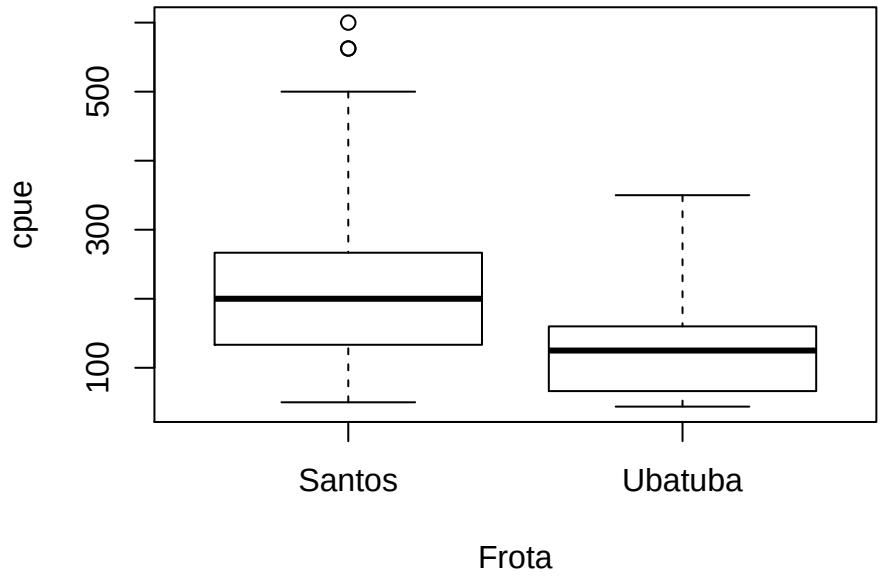

Figura 3: Gráficos de caixa robusto da cpue pela frota

A mesma superioridade já identificada da frota de Santos em comparação com a frota de Ubatuba é apresentada na Figura 3 conforme mostra os gráficos de caixa robustos.

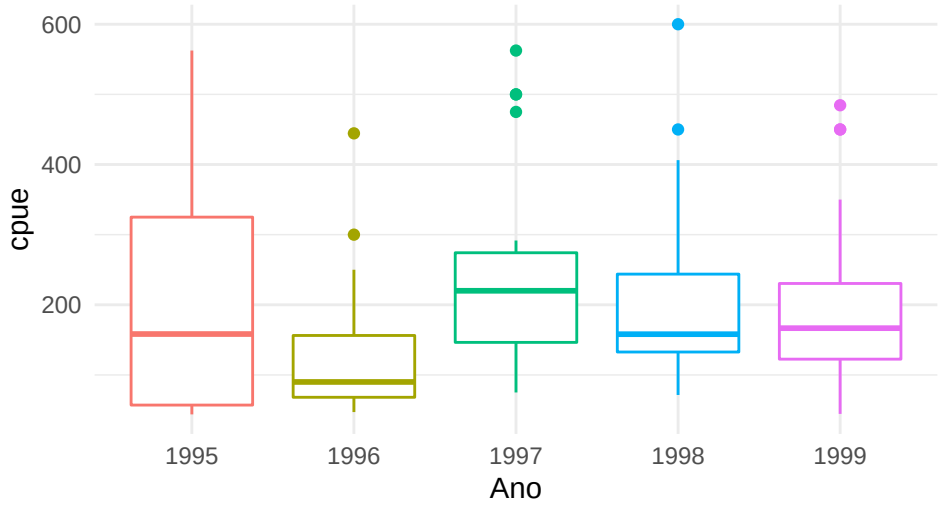

Figura 4: Gráficos de caixa da cpue pelo ano

Na Figura 4 observa-se que o ano de 1997 apresenta maior mediana em comparação com os demais, neste mesmo ano nota-se a presença de três observações *outliers*.

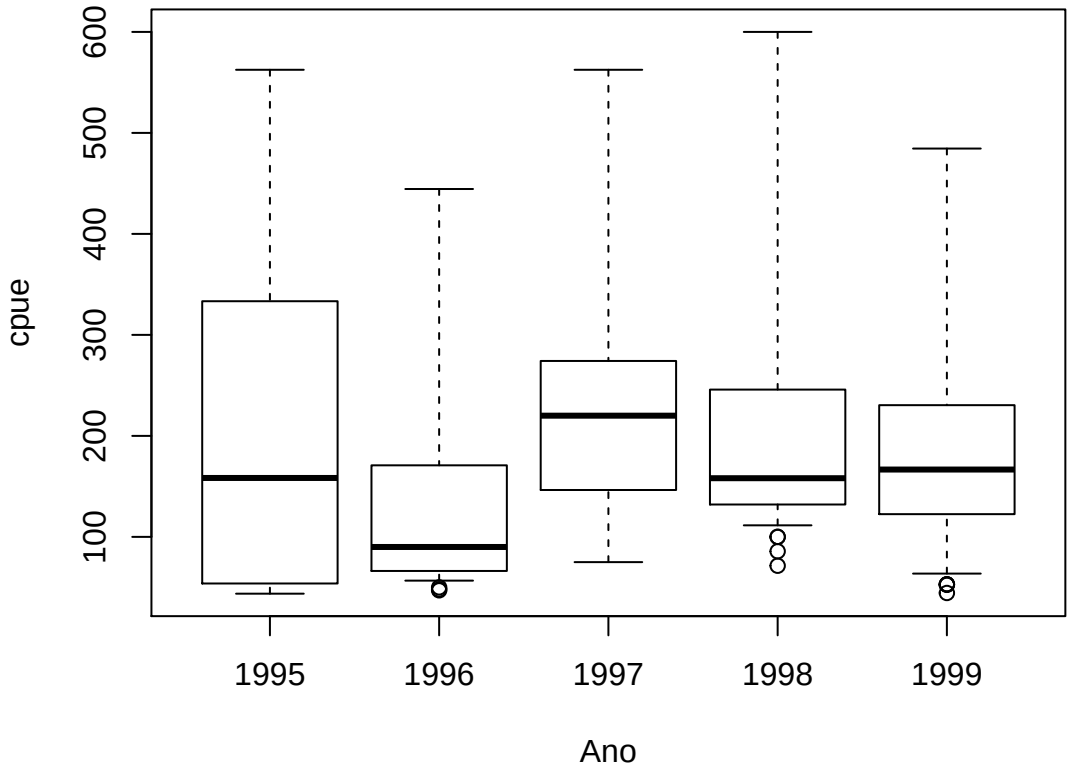

Figura 5: Gráficos de caixa robusto da cpue pelo ano

Conforme é ilustrado na Figura 5, as observações *outliers* agora encontram-se na parte inferior dos gráficos de caixa robustos. O ano de 1997 ainda possui mediana superior aos demais anos.

```
# Estatísticas descritivas da variável frota em função dos anos
dados %>%
  dplyr::select(frota, ano, cpue) %>%
  group_by(ano, frota) %>%
  dplyr::summarize(media = round(mean(cpue), 2), dp = round(sd(cpue), 2),
                   cv = round(dp/media * 100, 2), n = n()) %>%
  arrange(frota)
```

Tabela 2: Medidas resumo da variável frota de acordo com o ano

Frota	Estatística	1995	1996	1997	1998	1999
Santos	Média	229,37	193,19	262,67	210,29	197,22
	Des. Padrão	148,07	132,55	153,60	122,95	103,45
	Coef. Variação	64,55%	68,61%	58,48%	58,44%	52,48%
	n	19	8	17	27	46
Ubatuba	Média	47,08	96,09	210,56	174,43	140,85
	Des. Padrão	4,73	59,19	77,51	99,16	71,59
	Coef. Variação	10,05%	61,60%	36,81%	56,85%	50,83%
	n	3	12	6	5	13

Na Tabela 2 é possível conferir que somente no ano de 1996 a frota de Ubatuba possuia mais embarcações do que a frota de Santos, em 1995 e de 1997 a 1999, a frota de Santos contava com mais embarcações. O coeficiente de variação, que é dado pelo quociente do desvio padrão pela média, para a frota de Santos é praticamente constante, por outro lado, na frota de Ubatuba nos anos de 1995 e 1997 seu valor disto dos demais anos.

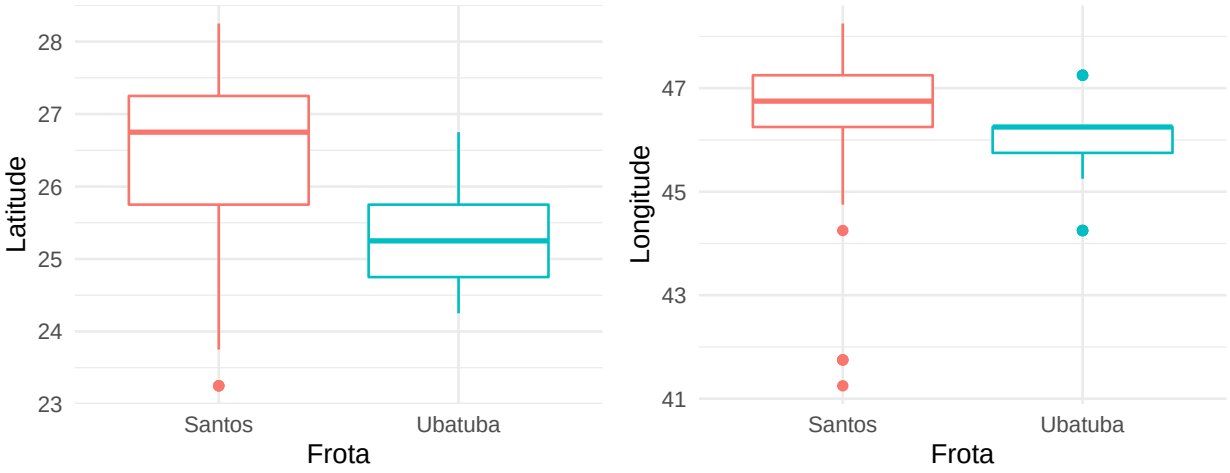

Figura 6: Gráficos de caixa da latitude e longitude pela frota

Pela Figura 6 nota-se que a frota da cidade de Santos tem preferência por pescar em latitudes e longitudes mais elevadas do que a frota de Ubatuba.

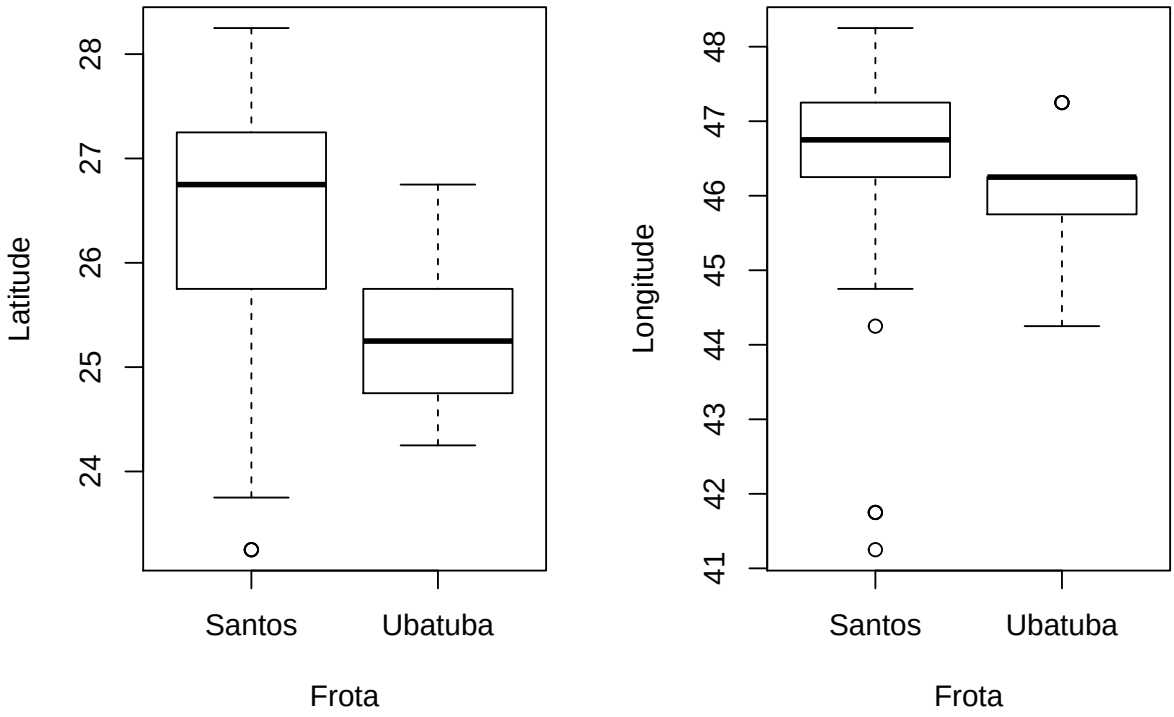

Figura 7: Gráficos de caixa robusto da latitude e longitude pela frota

Assim como foi destacado na Figura 6, os gráficos de caixa robustos da Figura 7 também ilustra a preferência da frota de Santos em pescar nas latitudes e longitudes maiores do que a frota de Ubatuba.

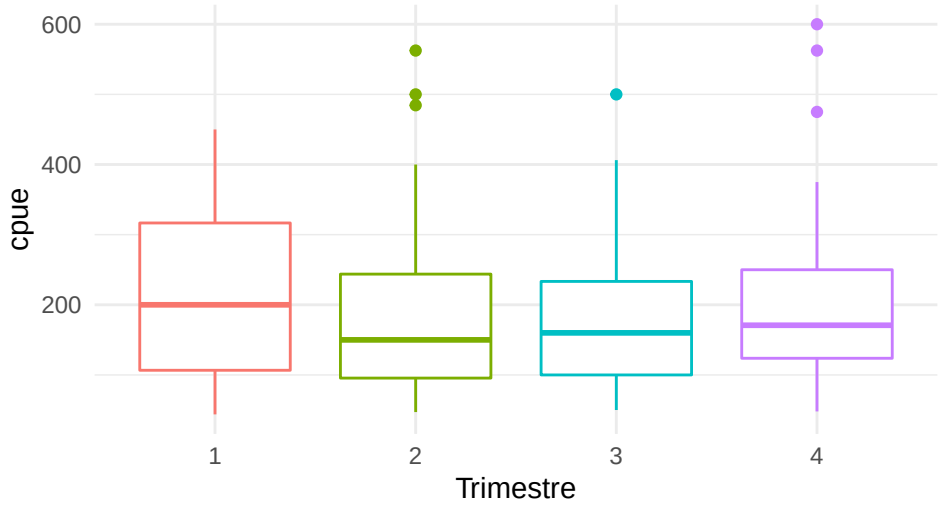

Figura 8: Gráficos de caixa da cpue pelo trimestre

Em relação aos trimestres, o primeiro trimestre apresenta mediana levemente maior do que os demais e o segundo e quarto trimestre possuem três observações *outliers*, conforme ilustra a Figura 8.

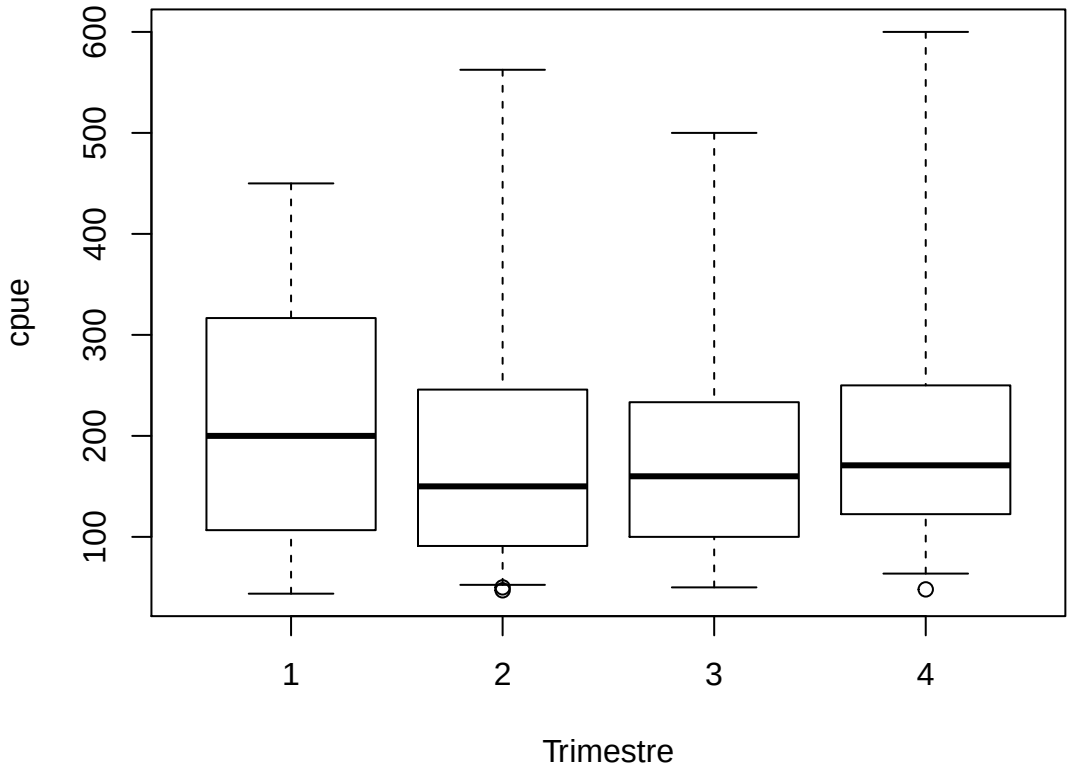

Figura 9: Gráficos de caixa robusto da cpue pelo trimestre

Analogamente, os gráficos de caixa robustos também indicam que o primeiro trimestre apresenta mediana levemente maior do que os demais trimestres, veja Figura 9.

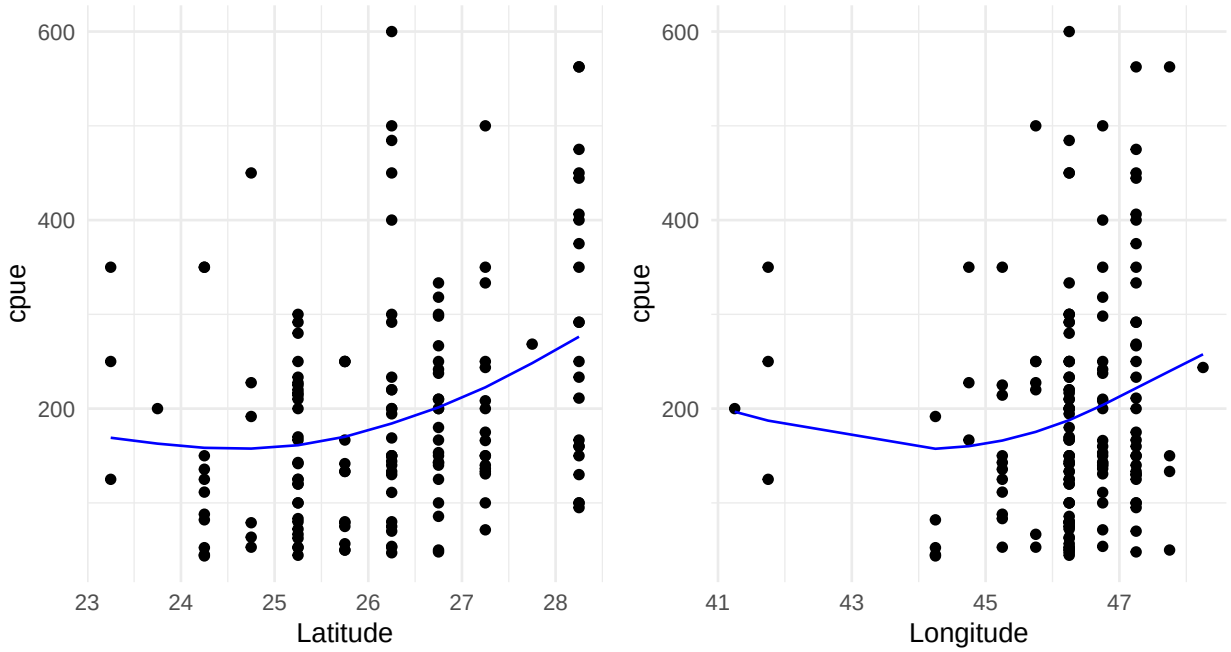

Figura 10: Gráficos de dispersão da cpue pela latitude e longitude

O gráfico de dispersão da `cpue` pela latitude indica indícios de um sutil crescimento da `cpue` de acordo com o aumento da latitude, por outro lado, não se pode dizer o mesmo da `cpue` em relação à longitude, há inicialmente uma tendência decrescente e em seguida crescente da `cpue` com o aumento da longitude.

Os gráficos de caixa robustos são indicados quando os dados são assimétricos. Foi possível observar um aumento no número de observações *outliers* na parte inferior dos gráficos e uma redução na parte superior desses gráficos.

Ajustes dos modelos lineares generalizados

Conforme foi constatado, vide Figura 1, que a distribuição da `cpue` possui assimetria à direita, é plausível propor e ajustar modelos MLG com resposta gama ou normal inversa.

- Modelo gama

```
# Ajuste do modelo MLG gama completo com função de ligação log
dados$frota <- as.factor(dados$frota)
dados$ano <- as.factor(dados$ano)
dados$trimestre <- as.factor(dados$trimestre)

fit.mod1 <- glm(cpue ~  frota +  ano + trimestre + latitude +
                  longitude, family = Gamma(link = log), data = dados)
summary(fit.mod1)
xtable(summary(fit.mod1), digits = 2)
```

Inicialmente realizou-se o ajuste do modelo MLG gama com todas as covariáveis, as estimativas dos parâmetros podem ser conferidas na Tabela 3.

Tabela 3: Estimativas do modelo de regressão gama ajustado

Efeito	Estimativa	Erro padrão	valor-t	Pr(>|t|)
Intercepto	5,20	2,25	2,31	0,02*
frotaUbatuba	-0,22	0,13	-1,68	0,09
ano1996	-0,19	0,18	-1,05	0,30
ano1997	0,36	0,17	2,12	0,04*
ano1998	0,10	0,16	0,63	0,53
ano1999	0,04	0,14	0,30	0,77
trimestre2	-0,14	0,15	-0,93	0,35
trimestre3	-0,29	0,15	-2,00	0,05*
trimestre4	-0,22	0,15	-1,48	0,14
latitude	0,18	0,07	2,56	0,01*
longitude	-0,10	0,07	-1,32	0,19

Na Tabela 3 verifica-se que ao nível de 5% de significância, o intercepto e as covariáveis `ano`, `trimestre` e `latitude` são significativas para o modelo ajustado.

Seleção dos modelos MLG gama

```
# Seleção do melhor modelo MLG gama segundo a técnica stepwise
stepAIC(fit.mod1)
```

```
# Ajuste do modelo MLG gama indicado pela técnica stepwise
fit.mod2 <- glm(cpue ~ frota + ano + latitude + longitude, family = Gamma(link = log),
                data = dados)
summary(fit.mod2)
xtable(summary(fit.mod2), digits = 2)
```

Com o intuito de selecionar o melhor modelo, utilizou-se a técnica de seleção de modelos *stepwise*, as estimativas dos parâmetros podem ser verificadas na Tabela 4.

Tabela 4: Estimativas do modelo de regressão gama ajustado segundo a técnica *stepwise*

Efeito	Estimativa	Erro padrão	valor-t	Pr(>|t|)
Intercepto	5,99	2,26	2,66	0,01*
frotaUbatuba	-0,28	0,13	-2,10	0,04*
ano1996	-0,15	0,19	-0,81	0,42
ano1997	0,33	0,17	1,90	0,06
ano1998	0,12	0,16	0,72	0,47
ano1999	0,07	0,15	0,48	0,63
latitude	0,17	0,07	2,30	0,02*
longitude	-0,11	0,08	-1,46	0,15

Note que a covariável `trimestre` foi retirada do modelo. De acordo com a Tabela 4, o intercepto e as covariáveis `frota`, e `latitude` são significativas para o modelo ao nível de 5% de significância.

```
# Ajuste do modelo MLG gama final com a interação frota * ano
fit.mod3 <- glm(cpue ~  frota + ano + latitude + longitude + frota * ano,
                family = Gamma(link = log), data = dados)
summary(fit.mod3)
xtable(summary(fit.mod3), digits = 2)
```

Com a intenção de melhorar ainda mais o modelo indicado pelo *stepwise* acrescentamos a interação entre as covariáveis `frota` * `ano`.

Tabela 5: Estimativas do modelo de regressão gama final

Efeito	Estimativa	Erro padrão
Intercepto	6,90	2,30
frotaUbatuba	-1,36	0,37
ano1996	-0,06	0,24
ano1997	0,14	0,19
ano1998	-0,04	0,17
ano1999	-0,01	0,16
latitude	0,20	0,07
longitude	-0,15	0,08
frotaUbatuba:ano1996	0,81	0,46
frotaUbatuba:ano1997	1,45	0,45
frotaUbatuba:ano1998	1,50	0,45
frotaUbatuba:ano1999	1,11	0,40

- Modelo normal inversa

```
# Ajuste do modelo MLG normal inversa completo com função de ligação log
fit.mod11 <- glm(cpue ~  frota +  ano + trimestre + latitude +
                 longitude, family = inverse.gaussian(link = log), data = dados)
summary(fit.mod11)
xtable(summary(fit.mod11), digits = 2)
```

Tabela 6: Estimativas do modelo de regressão normal inversa ajustado

Efeito	Estimativa	Erro padrão	valor-t	Pr(>|t|)
Intercepto	4,86	2,27	2,14	0,03*
frotaUbatuba	-0,25	0,12	-2,01	0,05*
ano1996	-0,16	0,17	-0,95	0,35
ano1997	0,47	0,19	2,49	0,01*
ano1998	0,21	0,16	1,27	0,21
ano1999	0,12	0,15	0,80	0,43
trimestre2	-0,14	0,16	-0,89	0,37
trimestre3	-0,28	0,15	-1,81	0,07
trimestre4	-0,23	0,16	-1,47	0,14
latitude	0,18	0,07	2,41	0,02*
longitude	-0,09	0,08	-1,18	0,24

Verifica-se na Tabela 6 que as covariáveis `frota`, `ano` e `latitude` e o intercepto são significativos para o modelo ao nível de 5% de significância.

Seleção dos modelos MLG normal inversa

```
# Seleção do melhor modelo MLG normal inversa segundo a técnica stepwise
stepAIC(fit.mod11)
```

```
# Ajuste do modelo MLG normal inversa segundo a técnica stepwise
fit.mod22 <- glm(cpue ~ frota + ano + latitude, family = inverse.gaussian(link = log),
                 data = dados)
summary(fit.mod22)
xtable(summary(fit.mod22), digits = 2)
```

Analogamente ao que foi feito com o modelo MLG gama acima, utilizou-se a técnica de seleção de modelos *stepwise* para selecionar o melhor modelo, as estimativas dos parâmetros podem ser verificadas na Tabela 7.

Tabela 7: Estimativas do modelo de regressão normal inversa ajustado segundo a técnica *stepwise*

Efeito	Estimativa	Erro padrão	valor-t	Pr(>|t|)
Intercepto	3,17	1,15	2,75	0,01*
frotaUbatuba	-0,41	0,12	-3.51	0,00*
ano1996	-0,08	0,17	-0,50	0,62
ano1997	0,45	0,19	2,37	0,02*
ano1998	0,24	0,16	1,47	0,14
ano1999	0,15	0,14	1,04	0,30
latitude	0,08	0,04	1,78	0,08

Conforme consta na Tabela 7, o intercepto e as covariáveis `frota` e `ano` ao nível de 5% de significância, são significativas para o modelo.

```
# Ajuste do modelo MLG normal inversa final
fit.mod33 <- glm(cpue ~ frota + ano + latitude + frota * ano,
                 family = inverse.gaussian(link = log), data = dados)
summary(fit.mod33)
xtable(summary(fit.mod33), digits = 2)
```

Com a mesma intenção de melhorar ainda mais o modelo indicado pelo *stepwise* acrescentamos a interação entre as covariáveis `frota` * `ano`.

Tabela 8: Estimativas do modelo de regressão normal inversa final

Efeito	Estimativa	Erro padrão
Intercepto	3,41	1,1
frotaUbatuba	-1,34	0,24
ano1996	-0,06	0,25
ano1997	0,16	0,21
ano1998	-0,03	0,18
ano1999	-0,03	0,17
latitude	0,07	0,04
frotaUbatuba:ano1996	0,66	0,34
frotaUbatuba:ano1997	1,26	0,37
frotaUbatuba:ano1998	1,34	0,35
frotaUbatuba:ano1999	1,01	0,28

Diagnóstico dos modelos ajustados

- AIC

Uma das formas de se avaliar um modelo MLG é calculando seu critério de informação de Akaike (AIC), sabe-se quanto menor é seu valor, melhor é o ajuste do modelo aos dados.

```
# Critério AIC para cada um dos MLG ajustados
xtable(AIC(fit.mod1, fit.mod2, fit.mod3, fit.mod11, fit.mod22, fit.mod33))
```

Tabela 9: Critério AIC dos modelos MLG ajustados

Modelo	AIC
fit.mod1	1874,19
fit.mod2	1872,75
fit.mod3	1866,45
fit.mod11	1879,14
fit.mod22	1875,68
fit.mod33	1867,99

Pela Tabela 9, o modelo que melhor se ajustou aos dados do espinhel de fundo segundo o critério AIC é o modelo `fit.mod3`. Por outro lado, o modelo `fit.mod33` também fez um bom ajuste.

```
# Teste da Anova para justificar a inclusão da interação entre os fatores frota e ano
anova(fit.mod1, fit.mod3, test = "Chisq")
xtable(anova(fit.mod1, fit.mod3, test = "Chisq"), digits = 2)
```

Tabela 10: Estimativas do teste da Anova para justificar a inclusão do fator de interação entre frota e ano

Modelo	G.l. Resíduo	Desvio residual	G.l.	Desvio	Pr(>Chi)
fit.mod1	145,00	47,04			
fit.mod3	144,00	44,32	1,00	2,72	0,00

Conforme mostra a Tabela 10 acima, o modelo `fit.mod3` em que foi incluído a interação `frota` $*$ `ano` é preferível ao modelo sem a inclusão, segundo o teste da Anova.

```r
# Ilustração da intereção entre os fatores frote e ano
x <- model.matrix(fit.mod3)
x <- as.data.frame(x)
x[,7] <- rep(26, nrow(x))
x[,8] <- rep(46, nrow(x))
attach(dados)
yhat = predict(fit.mod3, newdata = x, type = "response")

interaction.plot(ano, trace.factor = frota, yhat, ylab = "cpue estimada", xlab = "Ano",
                 col = c("blue", "red", "green4", "darkorange", "black", "brown", "gold",
                         "purple", "gray", "yellow4", "pink"),
                 lty = 1, lwd = 2, trace.label = "Frota", xpd = FALSE)
```

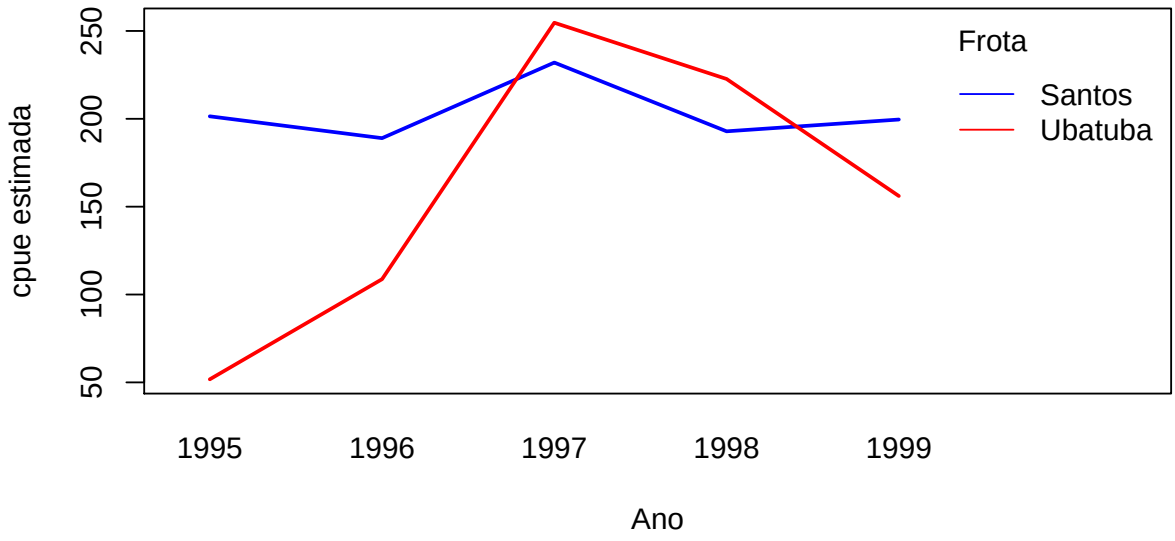

Figura 11: Estimativas da cpue média para as frotas de Santos e Ubatuba segundo o ano fixando-se a latitude em 26° e a longitude em 46° através do modelo gama

Na Figura 10 observou-se, com o aumento da latitude aumenta-se também a `cpue`, e que em relação à longitude ocorre o contrário. Dessa forma, espera-se maiores valores da `cpue` para altas latitudes e baixas longitudes. O gráfico da Figura 11 ilustra os valores esperados da `cpue` fixando latitude e longitude nos valores 26° e 46°, respectivamente. Assim, até 1996 os valores preditos para a frota de Ubatuba nessas latitude e longitude são bem menores do que os valores preditos para a frota de Santos. Entretanto, a partir de 1997 as diferenças entre os valores preditos para as duas frotas diminuem. Os valores preditos para a frota de Santos variam pouco no período de 1995 a 1999, diferentemente dos valores preditos para a frota de Ubatuba.

- Gráfico de envelope

Esta metodologia é utilizada com o intuito de verificar possíveis afastamentos das suposições feitas para o modelo, em especial para o componente aleatório e para a parte sistemática bem como a existência de observações discrepantes com alguma interferência desproporcional ou inferencial nos resultados dos ajustes.

```
# Envelope para o MLG gama final com função de ligação log
fit.model <- fit.mod3
source("http://www.ime.usp.br/~giapaula/envel_gama")
```

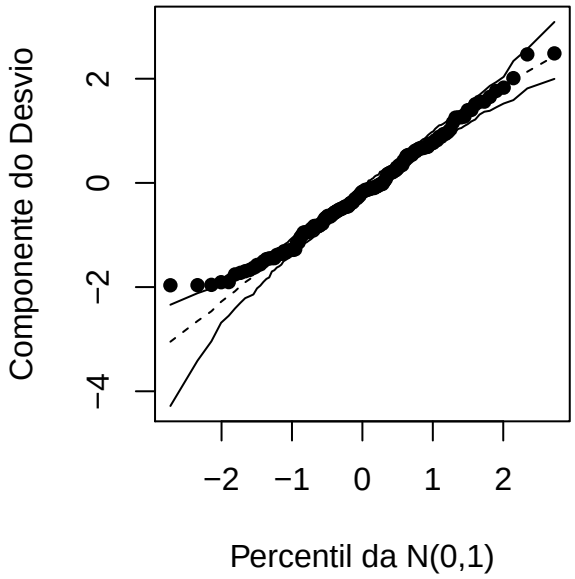

Figura 12: Gráfico de envelope para o modelo MLG gama que melhor se ajustou aos dados

Na Figura 12 referente ao envelope simulado, é possível constatar que duas observações estão fora do envelope porém, não apresenta indícios de que a distribuição gama seja inadequada para explicar a `cpue`, ou seja, o modelo foi ajustado adequadamente.

- Gráficos de diagnóstico

Nas figuras a seguir são apresentados os gráficos de diagnóstico. O gráfico da Medida h *versus* Valor ajustado permite identificar os pontos de alavanca. Com o gráfico da Distância de Cook *versus* Índice é possível identificar os pontos influentes. Já o gráfico do Resíduo Componente do Desvio *versus* Índice, identifica os chamados pontos aberrantes ou *outliers*. E, por fim, o gráfico do Resíduo Componente do Desvio *versus* Valor ajustado avalia a função de ligação escolhida.

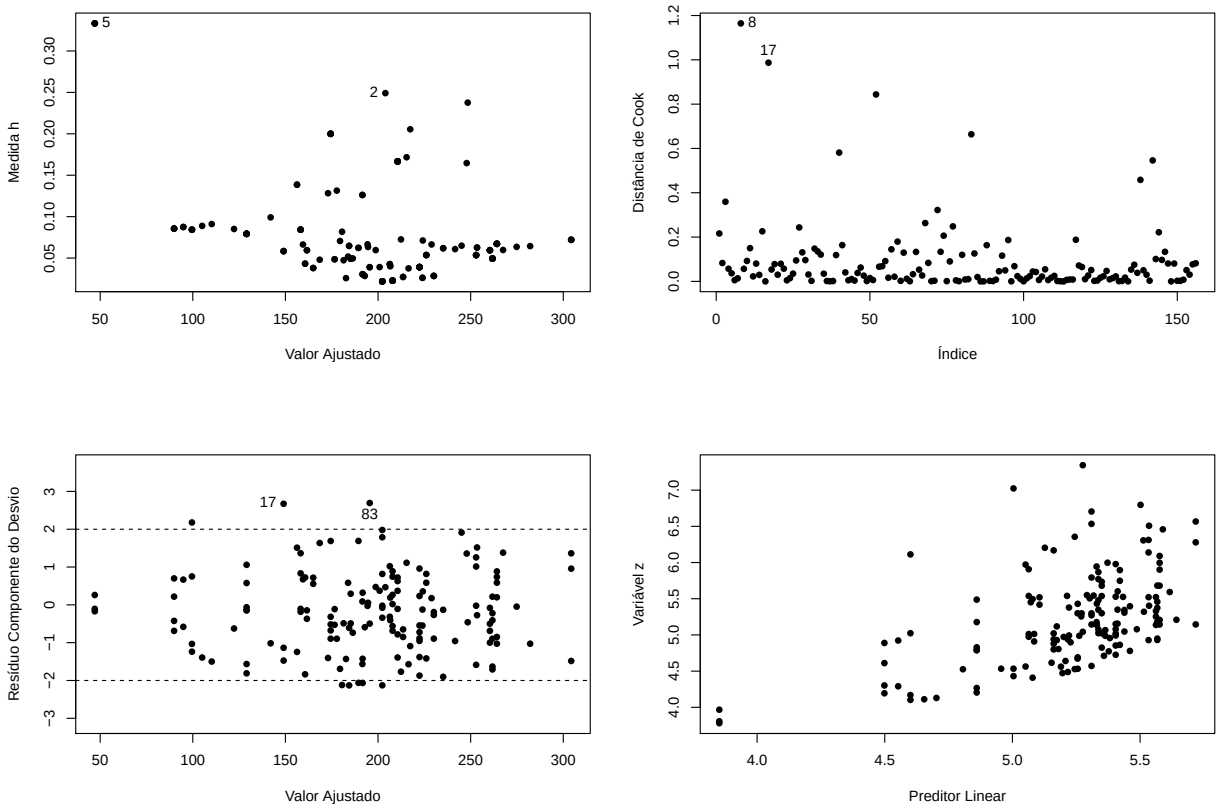

Figura 13: Gráfico de diagnóstico referente ao modelo MLG gama que melhor se ajustou aos dados

```
# Identificando as observações influentes
xtable(dados[c(8,17),], digits = 2)
```

Tabela 11: Informações das embarcações de números 8 e 17

Embarcação	Frota	Ano	Latitude	Longitude	Diaspesca	Captura	CPUE
8	Ubatuba	1998	24,25°	45,25°	10	3500	350
17	Santos	1999	24,75°	46,25°	5	2250	450

Os gráficos de diagnósticos não apresentam pontos de alavanca ou *outliers*, nem indicações de que a função de ligação utilizada é inadequada. Porém, no gráfico de pontos influentes duas observações aparecem com destaque, são as embarcações de números 8 e 17.

Pela Tabela 11 pode-se conferir que a embarcação de número 8 é da frota de Ubatuba e obteve uma `cpue` de 350 numa latitude de 24,25º e longitude de 45,25º no ano de 1998. Já a embarcação número 17 é da frota de Santos, obteve uma `cpue` de 450 numa latitude de 24,75°e longitude de 46,25° em 1999. As duas embarcações alcançaram `cpues` bastante altas em latitudes relativamente baixas, confirmando a tendência apresentada pelo modelo MLG gama.

```
xtable(influencePlot(fit.mod3, xlab = "Valor Ajustado", ylab = "Resíduo Studentizado ",
                     scale = 10), digits = 2)
```

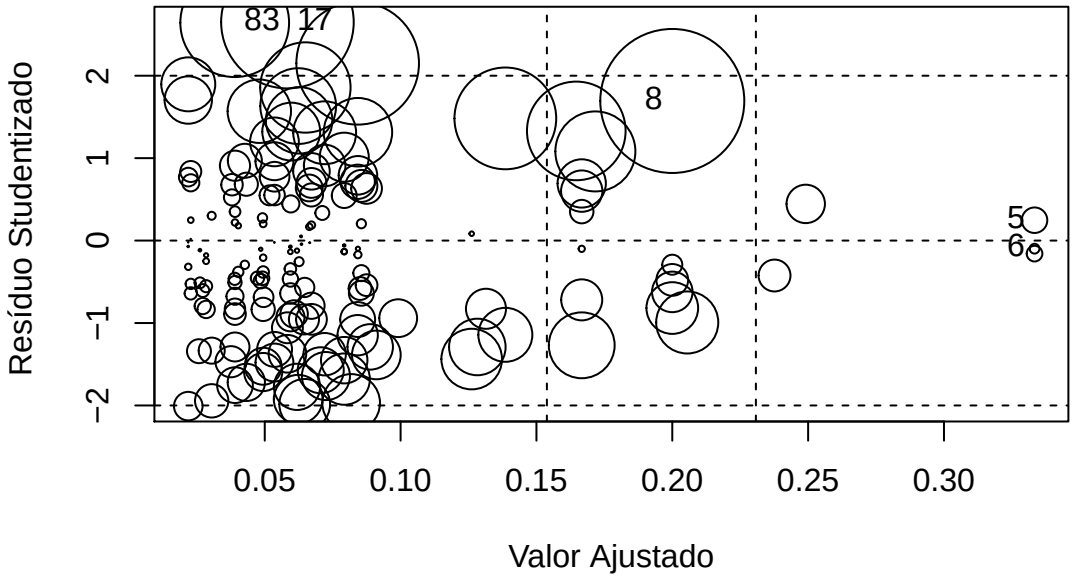

Figura 14: Gráfico de influência para o modelo MLG gama ajustado

Tabela 12: Estimativas das medidas de influência geradas pelo método *influencePlot*

Embarcação	Resíduo Studentizado	Valor Ajustado	Distância de Cook
5	0,25	0,33	0,00
6	-0,10	0,33	0,00
8	1,69	0,20	0,08*
17	2,65	0,06	0,07*
83	2,64	0,04	0,05

Para enfatizar a influência das embarcações de números 8 e 17, observe o gráfico da Figura 14 e as estimativas da Tabela 12, com destaque para as estimativas da distância de Cook onde tais embarcações obtiveram maiores valores.

Nota: Sabe-se que outras análises poderiam ser feitas com o objetivo de melhorar o ajuste do modelo, como por exemplo, a retirada dos pontos influentes e realização de um novo ajuste, ajustar o modelo com outras funções de ligação, porém para este contexto o modelo ajustado está muito bem adequado e satisfatório.

Conclusão

Após realizar todas as análises, o modelo que melhor se ajustou aos dados do espinhel de fundo segundo o critério AIC foi o modelo MLG gama com a inclusão da interação entre os fatores frota e ano.

O modelo estimado para explicar a `cpue` média segundo as covariáveis `frota`, `ano`, `trimestre`, `latitude` e `longitude` é dado por:

$$\begin{aligned}\hat{y}(\mathbf{x}) = 6,90 &- 1,36 frotaUbatuba - 0,06 ano1996 + 0,14 ano1997 - 0,04 ano1998 - 0,01 ano1999 + 0,20 latitude \\ &- 0,15 longitude + 0,81 frotaUbatuba * ano1996 + 1,45 frotaUbatuba * ano1997 \\ &+ 1,50 frotaUbatuba * ano1998 + 1,11 frotaUbatuba * ano1999\end{aligned}$$

em que $\mathbf{x} = (frota, ano, latitude, longitude, frota * ano)^T$.

Modelo final escolhido:

- Componente aleatório: $\mathbf{y}_{ijk} \overset{ind.}{\sim} Gama(\mu_{ijk}, \phi)$.
- Componente sistemático: $\eta_{ijk} = \log(\mu_{ijk}) = \alpha + \beta_j + \gamma_k + \delta_1 * latitude_{ijk} + \delta_2 * longitude_{ijk} + \theta_{jk}$, em que y_{ijk} denota a `cpue` observada para a i-ésima embarcação da j-ésima frota e no k-ésimo ano, enquanto θ_{jk} denota a interação entre a frota e o ano, com $i = 1, \ldots, 156$, $j \in \{Santos, Ubatuba\}$ e $k \in \{1995, 1996, 1997, 1998, 1999\}$. Como o modelo é casela de referência temos as restrições $\beta_1 = 0$, $\gamma_1 = 0$, $\theta_{1k} = 0\ \forall k$, e $\theta_{j1} = 0,\ \forall j$.
- Função de ligação: $g(\mu_{ijk}) = \eta_{ijk} = \log(\mu_{ijk})$.

"Essencialmente, todos os modelos estão errados, mas alguns são úteis" - George E. P. Box.