MAE5763 - Modelos Lineares Generalizados - Resolução da Lista 2

Guilherme Marthe - 8661962

3/11/2020

1 Exercício 6

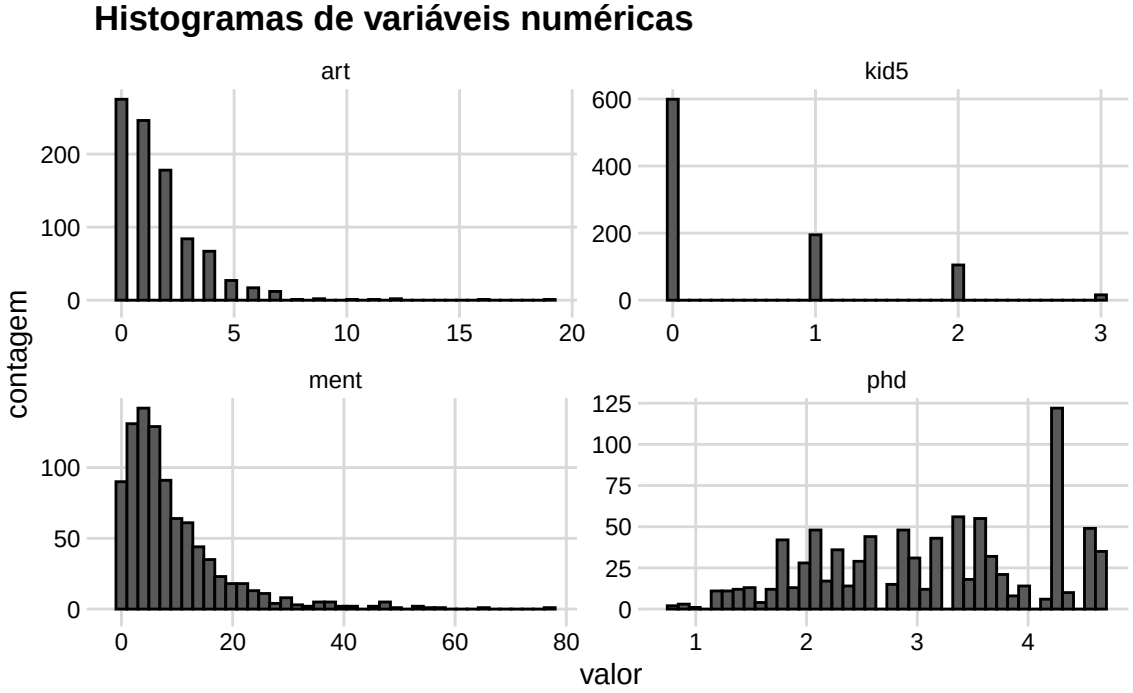
Iremos analisar a base de dados `bioChemists` fornecida pelo pacote `pscl`. As variáveis contidas nessa base de dados são:

- `art`: número de artigos publicados nos últimos 3 anos pelo doutor. Esta é a variável resposta.
- `fem`: sexo (masculino ou feminino)
- `mar`: estado civil (caso ou solteiro)
- `kids5`: número de filhos com até 5 anos
- `phd`: escore de prestígio do departamento onde o aluno fez doutorado
- `ment`: número de artigos publicados pelo orientador

1.1 Análise descritiva

1.1.1 Variáveis numéricas

Iniciaremos a análise por uma inspeção dos histogramas das variáveis numéricas. Assim, apresento o seguinte gráfico:



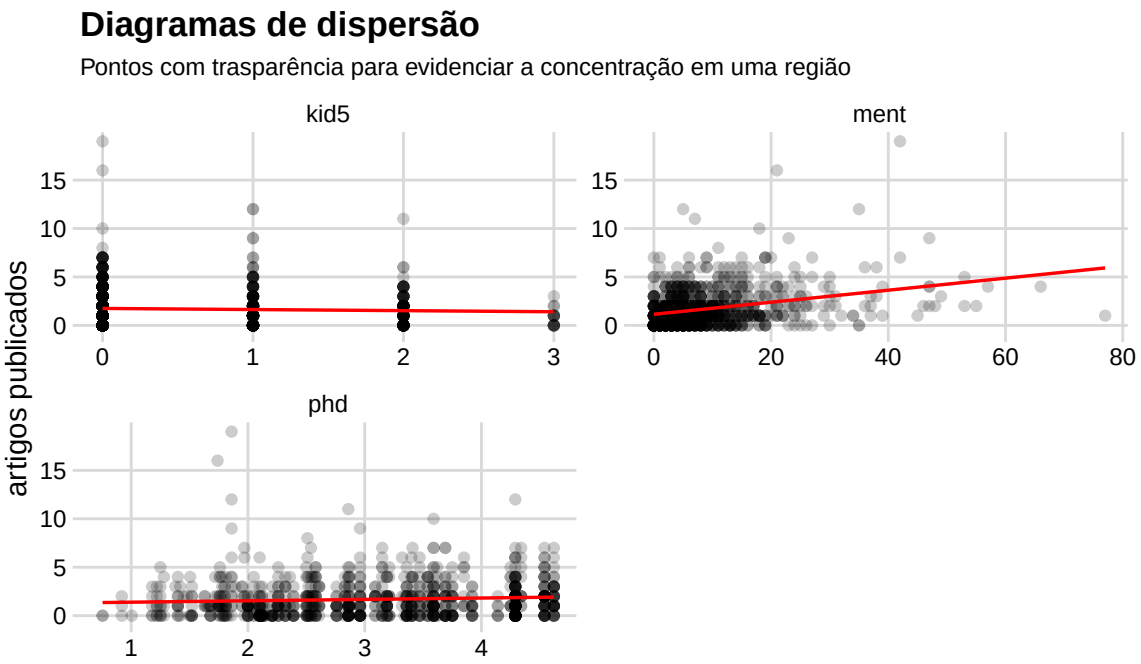
Alguns comentários:

- A variável resposta `art` é assimétrica à direita e positiva (>= a zero).
- Além disso, `art` é inteira e uma variável de contagem, como pode ser indicado pelo primeiro gráfico. Naturalmente nossa análise terá o enfoque em dados de contagem devido à isso.

- Com relação à variável `kid5` vemos a prevalência de doutores sem filhos.
- A massa de artigos escritos pelos orientadores (`ment`) fica concentrada entre 0 a 40 e, junto com a variável resposta, possui uma distribuição assimétrica à direita.
- O score de prestígio do departamento `phd` é aparentemente bem distribuído entre 1 e 5, porém parece ter uma moda acima do escore 4 (apesar de o histograma ser uma aproximação em faixas da variável contínua, existe uma concentração aí).

Abaixo temos os gráficos de dispersão das variáveis explicativas contra a variável resposta. Junto nesse gráfico temos uma regressão linear simples que está lá apenas para uma vaga referência da relação entre as duas variáveis. Como podemos ver, existe uma aparente correlação entre o número de artigos publicados pelo mentor e o publicado pelo aluno. Com as outras variáveis não temos uma relação marginal tão forte, aparentemente.

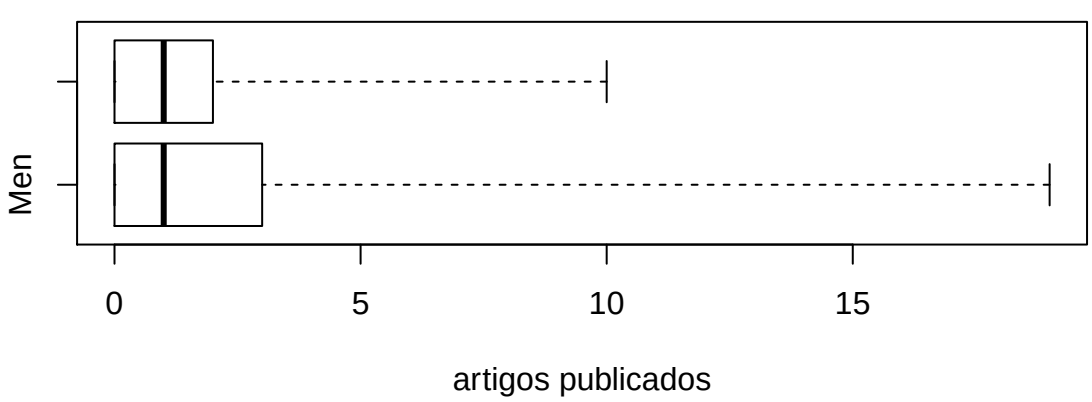
\



1.1.2 Variáveis do tipo fator

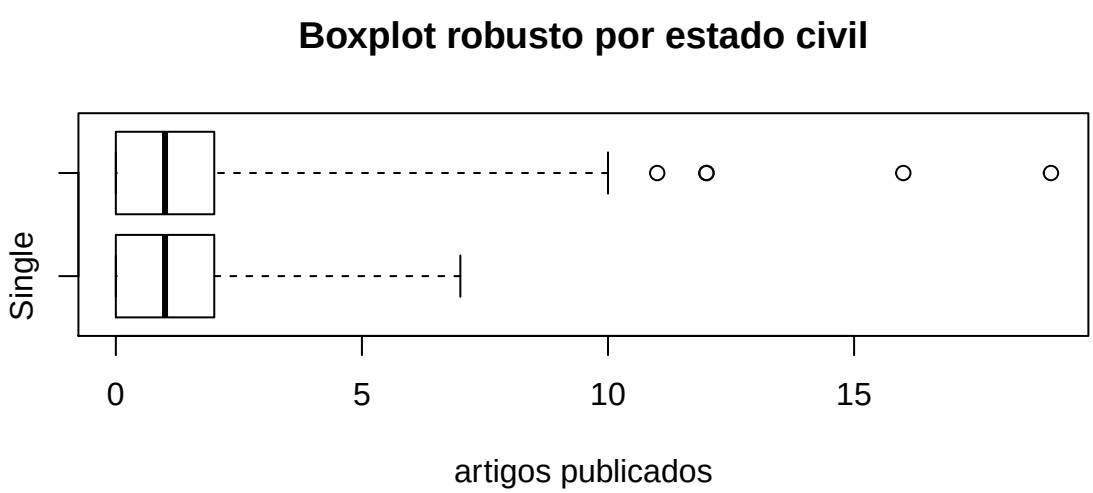
Para as variáveis do tipo fator iremos apresentar seus boxplots robustos contra a variável resposta `art`. Abaixo temos o caso da variável sexo. Notadamente não há uma diferença entre as medianas dos grupos, porém o sexo masculino parece possuir uma distribuição mais assimétrica à direita.

```
robustbase::adjbox(art ~ fem ,
                   data = bio, varwidth=F,
                   horizontal=T, xlab='artigos publicados',
                   main='Boxplot robusto por sexo')
```



Com relação ao estado civil, a variável resposta parece ter a mediana muito próxima entre o grupo dos casados contra os solteiros. O grupo dos casados, todavia, parece ter uma distribuição mais assimétrica, junto com o fato de 4 observações parecem estar acima da tendência robusta para esses dados.

```
robustbase::adjbox(art ~ mar,
                   data = bio, varwidth=F,
                   horizontal=T, xlab='artigos publicados',
                   main='Boxplot robusto por estado civil')
```



1.1.2.1 Tabelas de contigência Primeiramente, com relação à variável `sex`, apresentamos abaixo uma tabela de contingência que sumaria as contagens com relação ao total. Em geral existe uma concentração de zeros da variável resposta. Além disso a amostra parece balanceada com relação ao sexo. Note que combinamos em termos de faixa a variável resposta para que a tabela não fique muito extensa.

art	Men	Women	Total
0	14.9%	15.2%	30.1%
1	14.8%	12.1%	26.9%
2	9.5%	9.9%	19.5%
3	5.9%	3.3%	9.2%
[4,19]	9.0%	5.5%	14.4%
Total	54.0%	46.0%	100.0%

Com respeito ao estado civil, aproximadamente dois terços da amostra é casada. Essa prevalência de doutores casados parece ser consistente ao longo dos diversos níveis do número de artigos publicados.

art	Single	Married	Total
0	10.8%	19.2%	30.1%
1	9.2%	17.7%	26.9%
2	6.6%	12.9%	19.5%
3	2.4%	6.8%	9.2%
[4,19]	4.8%	9.6%	14.4%
Total	33.8%	66.2%	100.0%

1.1.3 Outros comentários sobre a variável resposta

Mostro abaixo as médias e variâncias amostrais estratificadas pelas variáveis nominais, incluindo a amostra como um todo. Como podemos ver, em geral a variância é maior que a média, indicando que, apesar dos dados serem de contagem, existe uma sobredispersão dos dados. Se sobredispersão for excessiva, isso pode indicar que uma modelagem via Poisson não é adequada para ajustar os dados. Mas iremos checar isso mais adiante através de medidas de qualidade do ajuste que realizaremos.

		estatísticas	
variavel	estratificação	média	variância
	amostra inteira	1.692896	3.709742
estado civil	Men	1.882591	4.748865
	Women	1.470309	2.406854
	Married	1.744224	4.074967
sexo	Single	1.592233	2.989030

1.2 Ajuste e seleção de modelos

Iremos iniciar nossa investigação ajustando um modelo linear generalizado com distribuição poisson e outro modelo com resposta binomial negativa. Um fato importante para comentar é que durante os primeiros ajustes notamos um claro padrão entre os resíduos e a ordem das observações. Como não estamos cientes de alguma ordenação relevante entre as observações, nós embaralhamos as linhas da base de dados.

```
bio <- slice(bio, sample(1:n()))
fit_po_psel = gamlss(art ~ ., family = PO, data = bio)
fit_nbi_psel = gamlss(art ~ ., family = NBI, data = bio)
```

O próximo passo foi realizar uma seleção um sub-modelo de variáveis através do método de Akaike, que consiste na minimização do AIC (no caso usaremos o GAIC, uma generalização do critério de informação de Akaike). Iniciaremos com o modelo contendo todos os efeitos principais e em seguida retiramos uma variável por vez objetivando a minimização do AIC.

```
fit_nbi = stepGAIC(fit_nbi_psel)
fit_po = stepGAIC(fit_po_psel)
```

Os ajustes pré e pós seleção de ambos os modelos sem encontram na tabela abaixo. Alguns pontos importantes de serem explicitados.

- independentemente da resposta, a variável `phd` que indica o prestígio do departamento não era significante ao nível de 10% e foi retirada do modelo pelo processo de seleção via AIC.
- as estimativas dos coeficientes são muito similares entre os modelos poisson e binomial negativa após a seleção de variáveis.
- o GAIC é naturalmente menor entre os modelos pré e pós seleção.
- notamos que todas as variáveis que são selecionadas são significativas à 10%.

- note que o ajuste binomial negativo permite também o ajuste de uma parâmetro de dispersão `sigma`. Na distribuição Poisson esse parâmetro é igual à 1 pela própria construção da distribuição e parametrização via família exponencial. Isso está intimamente relacionado com o fato de que, para a distribuição de Poisson a média coincide com a variância. O mesmo não é verdade para a binomial Negativa, permitindo que o parâmetro de dispersão seja estimado pelos dados.

```
md_list <- list('pré seleção' = fit_po_psel,
                'pós seleção' = fit_po,
                'pré seleção' = fit_nbi_psel,
                'pós seleção' = fit_nbi)

texreg(md_list, stars=c(0.01, 0.05, 0.1),
       center = T,
       include.nagelkerke=F,
       custom.gof.rows =
         list('Desvio' = map(md_list, deviance) %>% map(round, digits=2)),
       custom.header = list('Poisson'=1:2, 'Bin. Negativa'=3:4),
       custom.gof.names = c('Num. Obs.', 'GAIC'),
       float.pos = 'h'
)
```

	Poisson		Bin. Negativa	
	pré seleção	pós seleção	pré seleção	pós seleção
μ (Intercept)	0.30^{***}	0.35^{***}	0.26^{*}	0.30^{***}
	(0.10)	(0.06)	(0.14)	(0.08)
μ femWomen	-0.22^{***}	-0.23^{***}	-0.22^{***}	-0.22^{***}
	(0.05)	(0.05)	(0.07)	(0.07)
μ marMarried	0.16^{**}	0.15^{**}	0.15^{*}	0.15^{*}
	(0.06)	(0.06)	(0.08)	(0.08)
μ kid5	-0.18^{***}	-0.18^{***}	-0.18^{***}	-0.18^{***}
	(0.04)	(0.04)	(0.05)	(0.05)
μ phd	0.01		0.02	
	(0.03)		(0.04)	
μ ment	0.03^{***}	0.03^{***}	0.03^{***}	0.03^{***}
	(0.00)	(0.00)	(0.00)	(0.00)
σ (Intercept)			-0.82^{***}	-0.82^{***}
			(0.12)	(0.12)
Desvio	3302.11	3302.35	3121.92	3122.1
Num. Obs.	915	915	915	915
GAIC	3314.11	3312.35	3135.92	3134.10

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

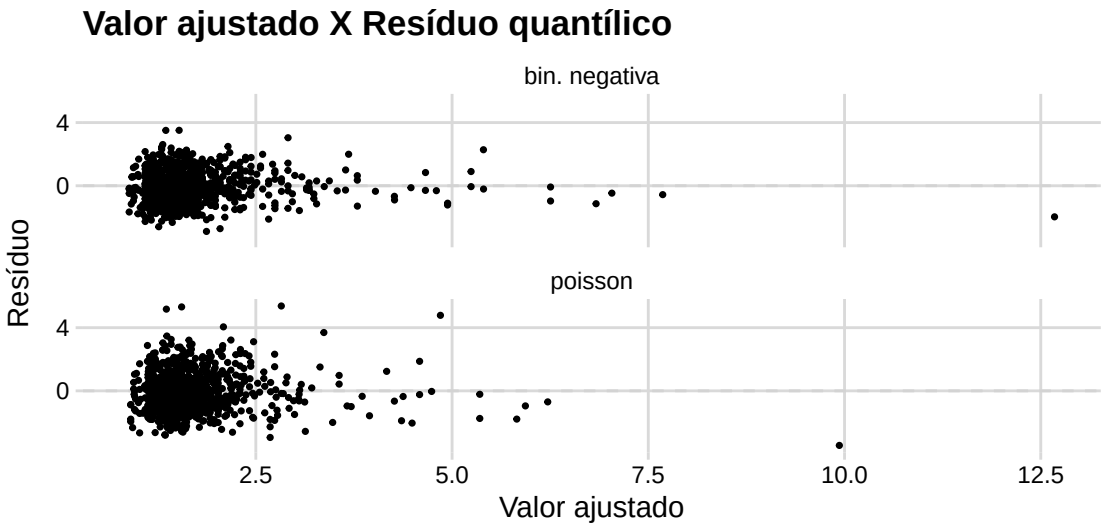
Statistical models

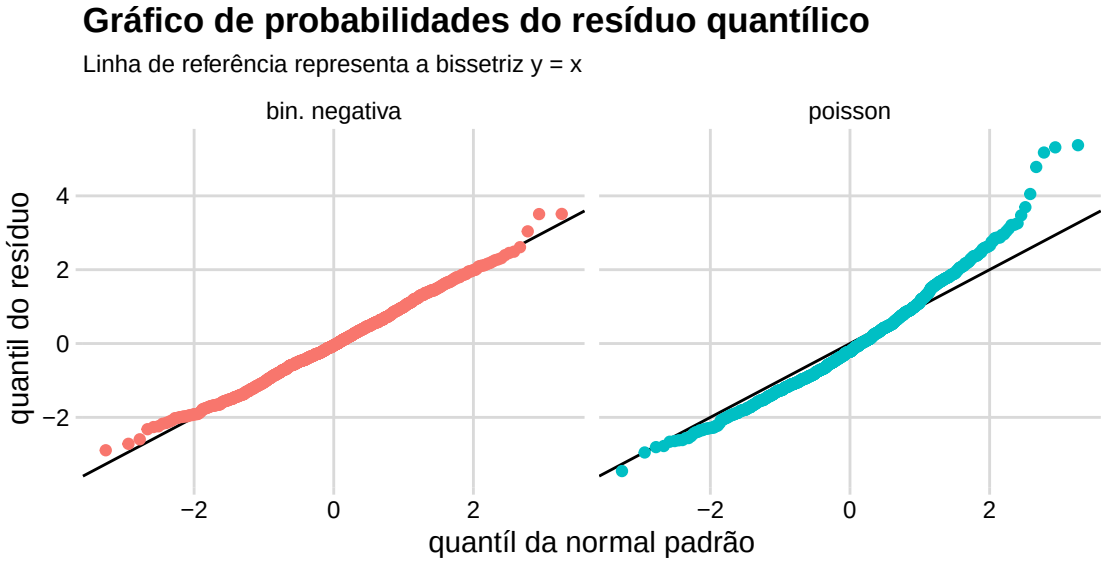
1.3 Análise de resíduos

Seguiremos com a análise dos modelos ajustados após a seleção de via método de Akaike. Como é possível observar no gráfico de valores ajustados contra o resíduo, não existe um padrão uniforme da núvem de pontos ao redor de da linha onde o resíduo é zero. Isso indica que existe uma variação não controlada pelos ajustes. Essa fenômento parece ser o mesmo entre ambos os modelos.

Todavia a inspeção do gráfico de probabilidades normal dos resíduos indica um ajuste melhor do modelo binomial negativo. No modelo Poisson deteca-se que os pontos cruzam a bissetriz, de maneira insatisfatória para o

ajuste desse modelo. Com o modelo binomial negativo existe uma massa ligeiramente maior no lado direito da distribuição, porém não muito preocupante em comparação com o ajuste poisson.

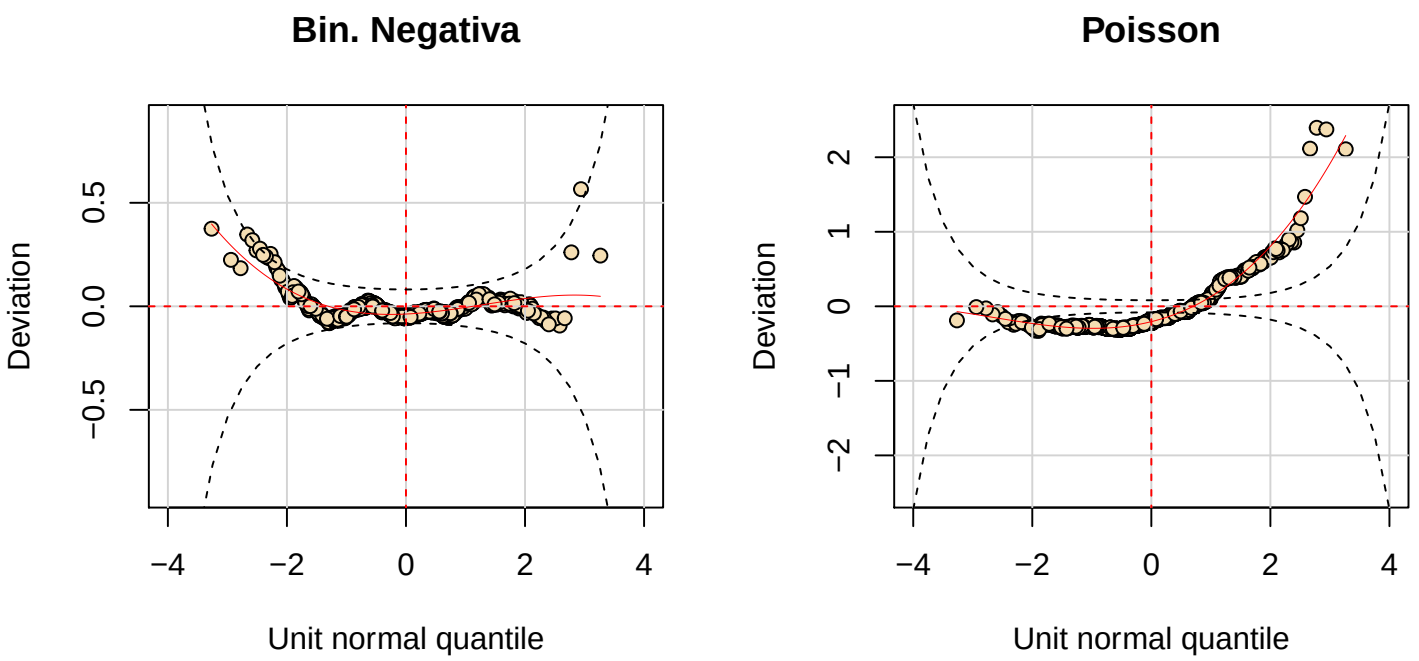




A inspeção dos wormplots evidencia mais ainda a qualidade de ajuste inferior do modelo Poisson. O conjunto de pontos ideal deve se dispor no interior das bandas de confiança em forma de ampulheta na horizontal. No caso da binomial negativa, existe uma pequena sobreposição do resíduo à direita. No caso Poisson a sobreposição é bem maior, indicando que esse o ajuste desse modelo não é satisfatório para o conjunto de dados que dispomos.

```
wp_nbi <- ~{wp(fit_nbi, ylim.all = .9)
title('Bin. Negativa')
}
wp_po <- ~{wp(fit_po, ylim.all = 2.5)
title('Poisson')
}
```

```
plot_grid(wp_nbi, wp_po)
```



1.4 Modelo ajustado em zeros

Vamos seguir com o ajuste via binomial negativa, porém seria interessante saber se a qualidade do ajuste da Poisson poderia ser melhorado levando em conta o excesso de zeros. Partiremos do conjunto de variáveis sem a variável `phd` para a média da distribuição, mas iniciaremos a análise com todas as variáveis para o parâmetro de probabilidade de zeros. Chamaremos esse modelo de *modelo completo* em nossa tabela de coeficientes.

```
fit_zanbi_0 <- gamlss(art ~ . - phd, nu.formula = art ~ ., data=bio, family = ZANBI())
```

Como podemos ver na tabela a seguir, o ajuste do *modelo completo* possui algumas variáveis não significativas, tanto para o parâmetro de localização quando o da probabilidade de zeros. Nomeadamente, no parâmetro de localização precisamos checar a remoção da variável `mar`; para o parâmetro de probabilidade de zero, remoção das variáveis `fem` e `phd` devem ser estudadas. Para checar a remoção conjunta dessas variáveis, realizaremos um teste de razão de verossimihanças do modelo mais simples, sem as variáveis ofensoras, contra o *modelo completo*. Chamaremos esse modelo menor de modelo *reduzido*. A função que utilizaremos para testa a razão de verossimilhanças entre o modelo *completo* e *reduzido* foi a `LR.test` e seus resultados estão na parte inferior da tabela de coeficientes.

```
fit_zanbi_1 <- gamlss(
  formula =  art ~ . -phd -mar,
  nu.formula = art ~ . -fem -phd, data=bio,
  family = ZANBI())
lr_test = LR.test(fit_zanbi_1, fit_zanbi_0, print = F)
```

```
zanbi_list = list(
  'Mod Sem ajust. p/ zeros' = fit_nbi,
  'Mod. completo' = fit_zanbi_0,
  'Mod. reduzido'= fit_zanbi_1
)
```

O teste de razão de verossimilhanças não é significante à um nível de 10%, indicando que não podemos rejeitar a hipótese nula e o modelo *reduzido*, mais simples, é preferível ao modelo *completo*. Notemos que as estimativas dos parâmetros que compõe o parâmetro de localização μ não variaram bruscamente com a adição do ajuste aos zeros. Adicionalmente, todas as estimativas do modelo *reduzido* são significaivas à 10% de confiança, tanto dos parâmetros associados ao componente de localização quanto os associados à probabilidade de zeros, com a

	Mod Sem ajust. p/ zeros	Mod. completo	Mod. reduzido
μ (Intercept)	0.30^{***}	0.34^{***}	0.40^{***}
	(0.08)	(0.12)	(0.10)
μ femWomen	-0.22^{***}	-0.24^{**}	-0.26^{***}
	(0.07)	(0.10)	(0.10)
μ marMarried	0.15^{*}	0.10	
	(0.08)	(0.11)	
μ kid5	-0.18^{***}	-0.15^{**}	-0.13^{*}
	(0.05)	(0.07)	(0.07)
μ ment	0.03^{***}	0.02^{***}	0.02^{***}
	(0.00)	(0.00)	(0.00)
σ (Intercept)	-0.82^{***}	-0.60^{***}	-0.59^{***}
	(0.12)	(0.23)	(0.23)
ν (Intercept)		-0.24	-0.13
		(0.30)	(0.15)
ν femWomen		0.25	
		(0.16)	
ν marMarried		-0.33^{*}	-0.36^{**}
		(0.18)	(0.18)
ν kid5		0.29^{**}	0.25^{**}
		(0.11)	(0.11)
ν phd		-0.02	
		(0.08)	
ν ment		-0.08^{***}	-0.08^{***}
		(0.01)	(0.01)
Desvio	3122.1	3105.2	3108.71
AIC	3134.1	3129.2	3126.71
Estat. RV p/ compl. e reduz.			3.52
GL. RV			3
P valor. RV			0.32
Num. obs.	915	915	915

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

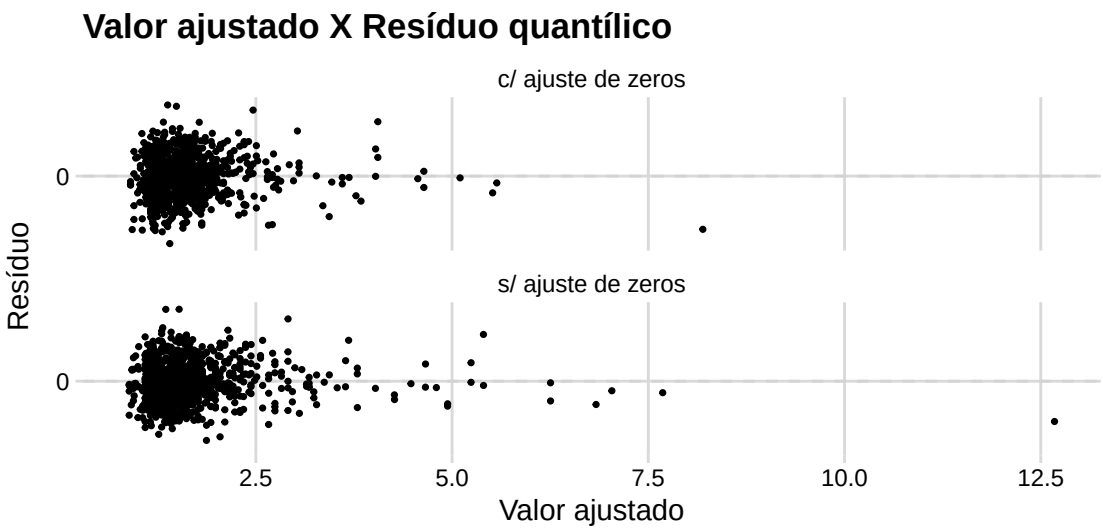
Statistical models

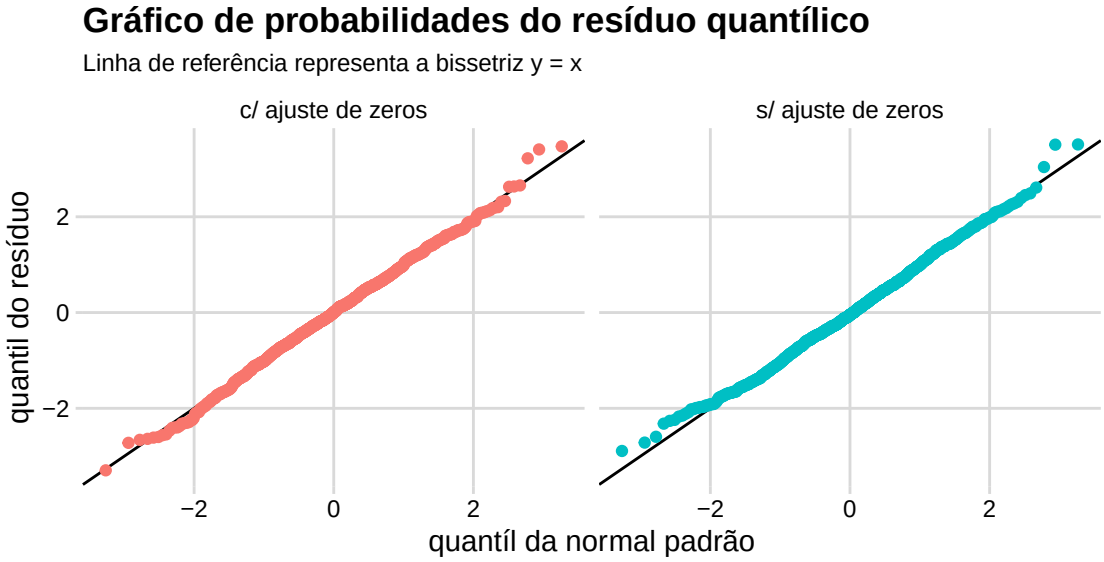
exceção do intercepto deste último.

Outro ponto importante são as medidas de qualidade de ajuste que apontam para o modelo *reduzido* ao candidato de melhor modelo. O desvio indica que a distribuição ZANBI é mais aderente à amostra (apesar da comparação entre desvios de distruibuições diferentes não é válida por terem verossimilhanças diferentes, nesse caso temos uma componente similar entre eles). Além disso, o AIC mostra que, de maneira parcimoniosa, o modelo com menos parâmetros compensa a pequena diminuição do desvio quando comparado com o modelo *completo*.

1.4.1 Análise de resíduos para o modelo ajustado em zeros

Com relação aos gráficos usuais de análise de resíduos, podemos notar que o modelo com a ajuste de zeros de fato apresenta melhoras. Apesar de ainda existir uma dispersão não muito uniforme dos resíduos contra os valores preditos, a dispersão é menor no modelo ZAMBI que no mobelo binomial negativo. Isso indica que parte da variabilidade dos dados não foi controlada pelos modelos propostos. A porção do gráfico de probabilidades que foge da bissetriz também é menor no modelo ajustado para zeros.

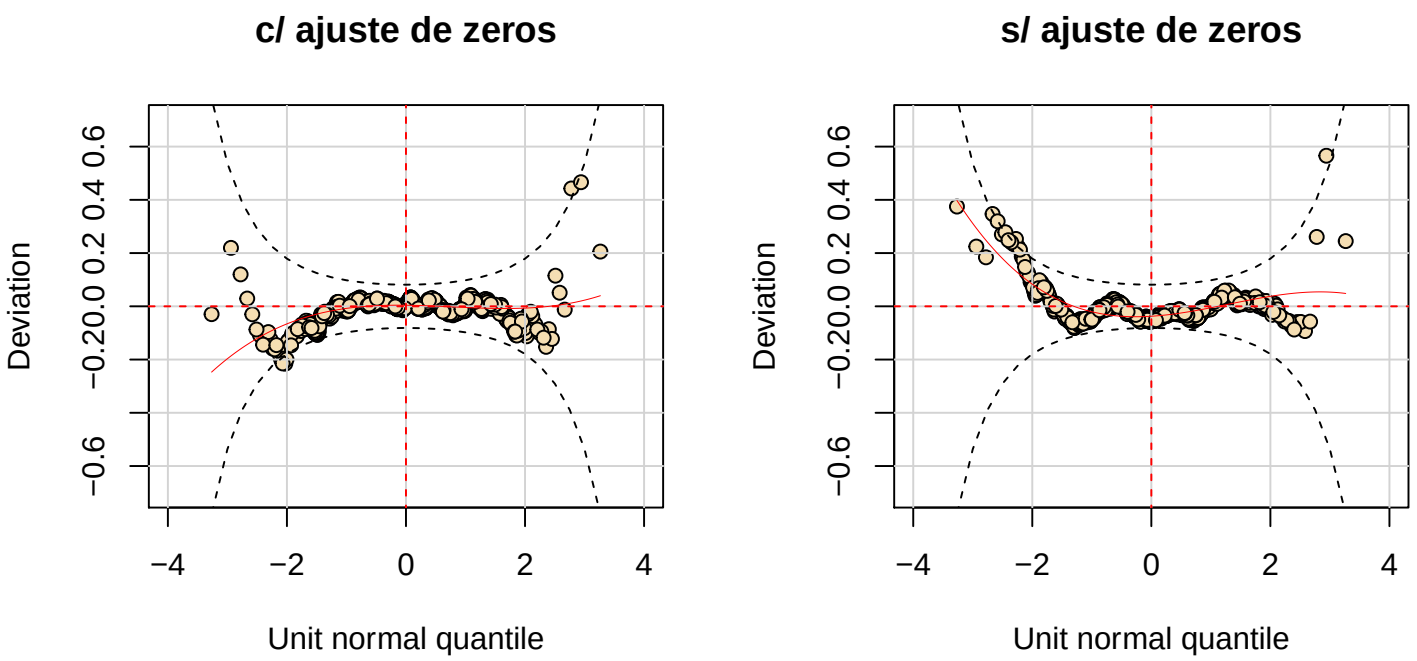




O gráfico de wormplot a seguir confirma as conclusões da análise gráfica anterior. Como mencionamos anteriormente, o modelo sem ajuste de zeros não é ruim. Todavia, o modelo ZANBI não possui pontos que sobrepõe as bandas de confiança indicando uma melhor aderência do modelo à amostra.

```
wp_1 <- ~{wp(fit_zanbi_1, ylim.all = .7)
title('c/ ajuste de zeros')
}
wp_2 <- ~{wp(fit_nbi, ylim.all = .7)
title('s/ ajuste de zeros')
}
```

```
plot_grid(wp_1, wp_2)
```



1.5 Interpretação do modelo ZANBI

Com base nas análises anteriores, o modelo que melhor ajustou os dados foi a versão reduzida do modelo binomial negativo ajustado para zeros. Reproduzimos as estimativas desse modelo para facilitar à consulta aos parâmetros. Assumindo o modelo como correto, podemos interpretar cada um dos parâmetros estimados conforme os pontos a seguir:

- parâmetros lineares associados ao parâmetro de localização μ:
  - Intercept: Aqui o intercepto corresponde à média de artigos esperada para um doutor do sexo masculino, sem filhos até 5 anos de idade que possuiu um orientador que não publicou nenhum artigo. Sabendo que de número de artigos é não nulo, é esperado que esse doutor tenha publicado 1.49 artigos nos últimos 3 anos.
  - **fem**: exponenciando a estimativa desta variável ($e^{-0.26}$) nos indica que a média esperada de artigos publicados nos últimos três anos de uma doutura é 78% quando comparada à média dos doutores, considerando todo o resto constante.
  - *kid5*: para cada criança adicional com até 5 anos que um doutor tenha, é esperado que sua média de artigos publicados seja $100 \cdot (1 - e^{-0.13})\% = 12\%$ menor, considerando todas as outra variáveis constantes.
  - **ment**: cada artigo adicional publicado pelo orientador do doutor está associado com um aumento de 2% na quantidade de artigos publicados nos últimos 3 anos, considerando as outras variáveis sem alterações.
- parâmetros lineares associados à probabilidade de zero μ:
  - Intercepto: note que a noção de sucesso para esta parcela do modelo é que o doutor não tenha tido nenhum artigo publicado nos ultimos 3 anos. Nesse sentido o intercepto representa a chance de um doutor (ou doutora) solteiro, sem filhos com menos de 5 anos em que o orientador tenha publicado zero artigos. Para esse hipotético doutor a chance de que ele não tenha publicado artigos é 0.8780954, refletindo que a probabilidade que esse doutor tenha publicado artigos nos últimos anos é maior que não tenha.
  - **mar**: o coeficiente associado com o estado civil de casado indica que as chances de um doutor(a) casado não publicar nenhum artigo são $100 \cdot (1 - e^{-0.36})\% = 30\%$ menores menores quando comparados com os solteiros, todo o resto constante.
  - **kid5**: cada filho com menos de cinco anos adicional que um doutor tenha está associado com uma

chance $100 \cdot (e^{.25} - 1)\% = 28\%$ maior de que ele não tenha publicado artigos nos últimos 3 anos.
- **ment**: cada artigo adicional que o seu orientador tenha publicado indica que as chances de que o doutor em questão não tenha publicado artigos nos últimos 3 anos é $100 \cdot (1 - e^{-0.08})\% = 8.08\%$ menor, dado o nível das demais variáveis.

	Mod. reduzido ZANBI
μ (Intercept)	0.40 (0.10)***
μ femWomen	−0.26 (0.10)***
μ kid5	−0.13 (0.07)*
μ ment	0.02 (0.00)***
σ (Intercept)	−0.59 (0.23)***
ν (Intercept)	−0.13 (0.15)
ν marMarried	−0.36 (0.18)**
ν kid5	0.25 (0.11)**
ν ment	−0.08 (0.01)***
Desvio	3108.71
AIC	3126.71
Num. obs.	915

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

Statistical models

2 Exercício 7

```
df <- rent %>%
  select(R, Fl, A, H, loc)
```

Iremos realizar a análise da base de dados consistente de uma amostra de 1967 unidades habitacionais em Munich em 1993. O objetivo é explicar o valor mensal líquido do aluguel líquido da unidade por meio das seguintes variáveis:

- *Fl*: área útil em metros quadrados
- *A*: ano da construção
- *H*: presença ou não de aquecimento central (0 sim, 1 não)
- *loc*: qualidade da localização do imóvel

Para ilustração, abaixo mostro as 5 primeiras linhas da base de dados:

```
df %>% head(5) %>% kable(format = 'latex') %>% kable_styling(position = 'center', full_width = T)
```

R	Fl	A	H	loc
693.3	50	1972	0	2
422.0	54	1972	0	2
736.6	70	1972	0	2
732.2	50	1972	0	2
1295.1	55	1893	0	2

2.1 Análise descritiva

2.1.1 Variáveis explicativas contínuas

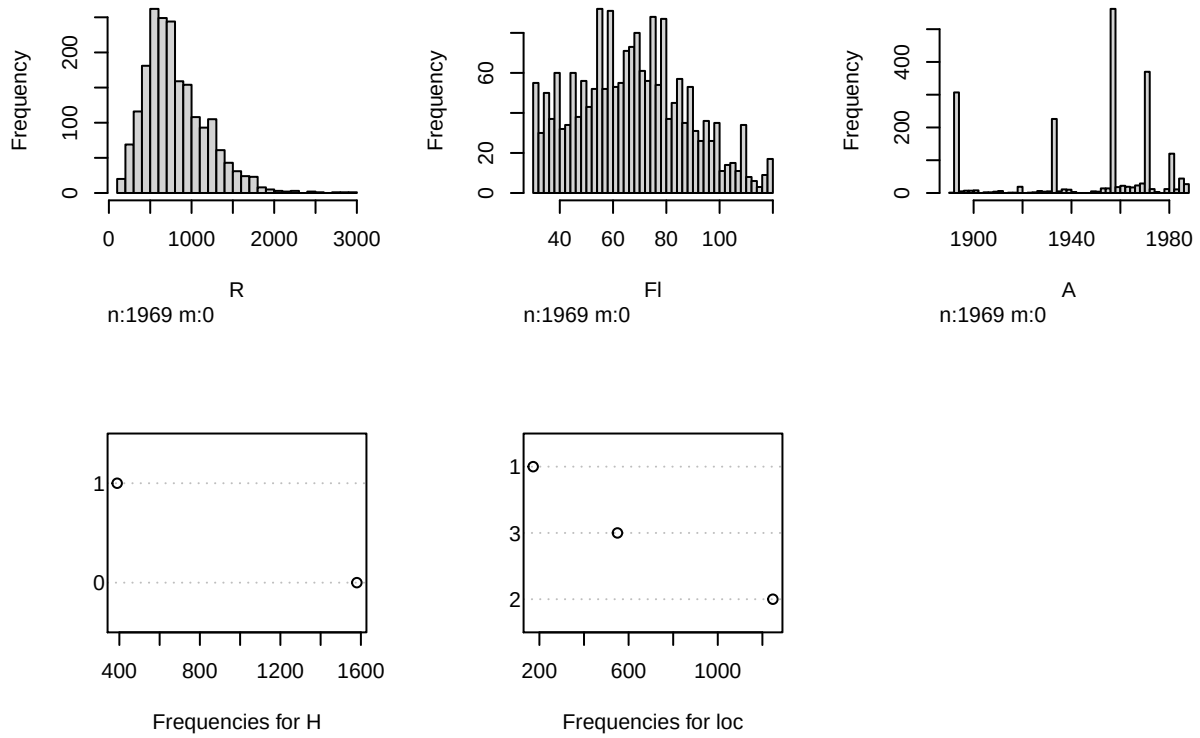
Abaixo mostro os histogramas para as variáveis contínuas e frequências para as variáveis nominais. Alguns comentários são pertinentes:

Matriz de correlação

	R	Fl	A
R	1.00	0.48	0.14
Fl	0.48	1.00	-0.14
A	0.14	-0.14	1.00

- com relação à variável resposta, R, é possível ver que ela tem uma distribuição assimétrica à direita e é estritamente positiva, indicando que a modelagem via a distribuição gamma é apropriada nessa aplicação.
- a variável de metragem parece preencher bem o seu intervalo máximo e mínimo, sem sinal de regiões pouco representadas.
- o ano de construção tem um intervalo de ocorrência bem amplo, com quase 100 anos, com algumas modas que distoam do resto.
- a variável H que representa a presença de aquecimento no imóvel é extremamente concentrada em imóveis que têm aquecimento. 80% dos imóveis da amostra possuem aquecimento.
- a qualidade da localização 1 (abaixo da média) é pouco representada, com apenas 9% dos imóves tendo esse nível. 63% estão na média (loc=2) e 28% com localidades acima da média.

```
hist(df)
```



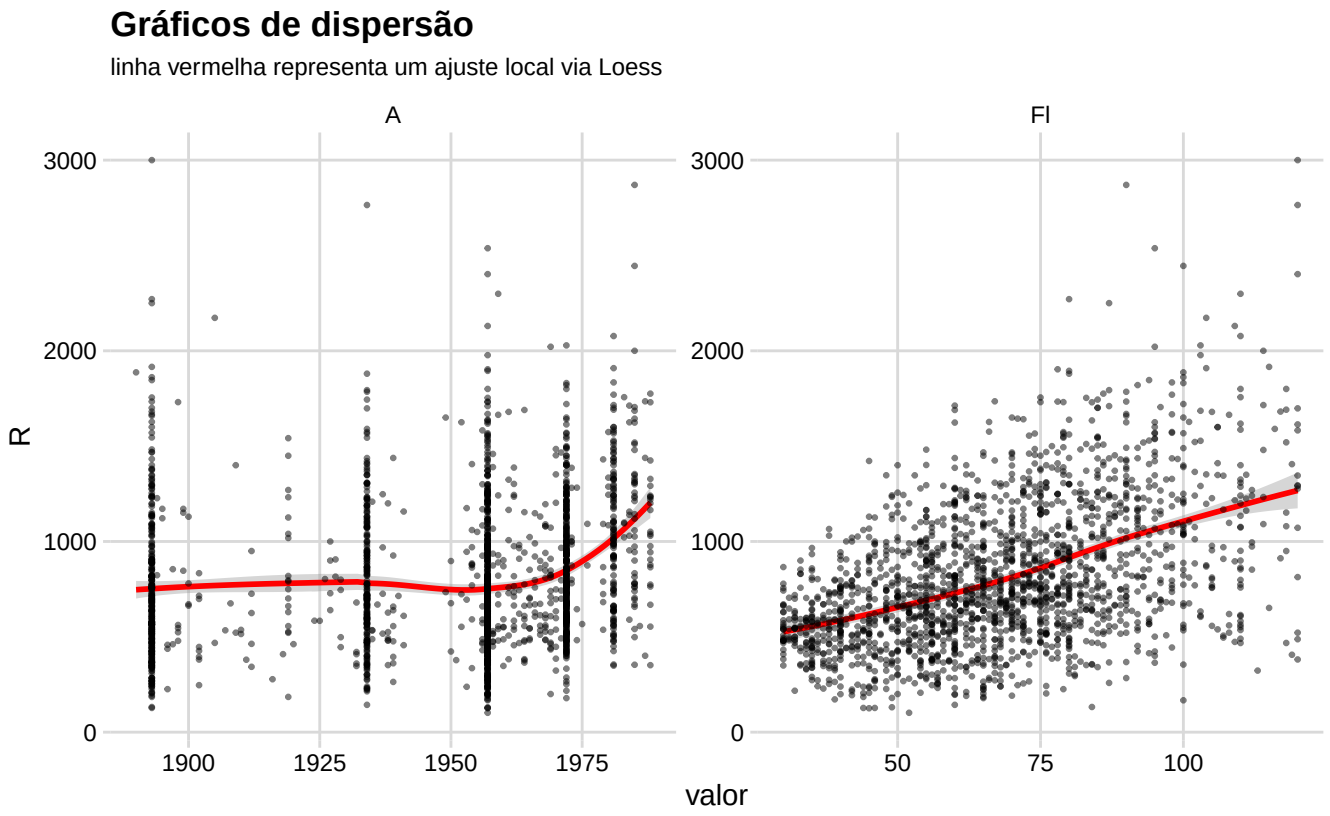
Abaixo temos a matriz de correlação entre as variáveis contínuas. Nota-se que a variável FL tem uma correlação relativamente forte com a resposta R. A variável ano, nem tanto. Outro ponto importante é notar que existe uma fraca correlação negativa entre as duas variáveis explicativas do modelo, importante em termos de multicoliniaridade e identificabilidade do modelo, alfo que não parece ser um problema.

```
cor(df %>% select_if(is.numeric)) %>%
  round(., 2) %>%
  kable( caption = 'Matriz de correlação')  %>%
  kable_styling(position = 'c', full_width = F)
```

A seguir podemos confirmas as tendências apontadas pela correlação liner por meio de um gráfico de dispersão entre as variáveis contínuas. De fato existe uma tendência linear expressiva entre o aluguel e a metragem do

imóvel. Além disso, existe uma tendência crescente entre o aluguel do imóvel após os anos 50.

```
df %>%
  select_if(is.numeric) %>%
  gather(var, value, -R)  %>%
  ggplot(aes(value, R)) +
  geom_smooth(method='loess', formula = y~x, color = 'red' ) +
  geom_point(size=.5, alpha=.5) +
  facet_wrap(~var, scales='free') +
  labs(title = 'Gráficos de dispersão',
       subtitle='linha vermelha representa um ajuste local via Loess',
       x = 'valor')
```



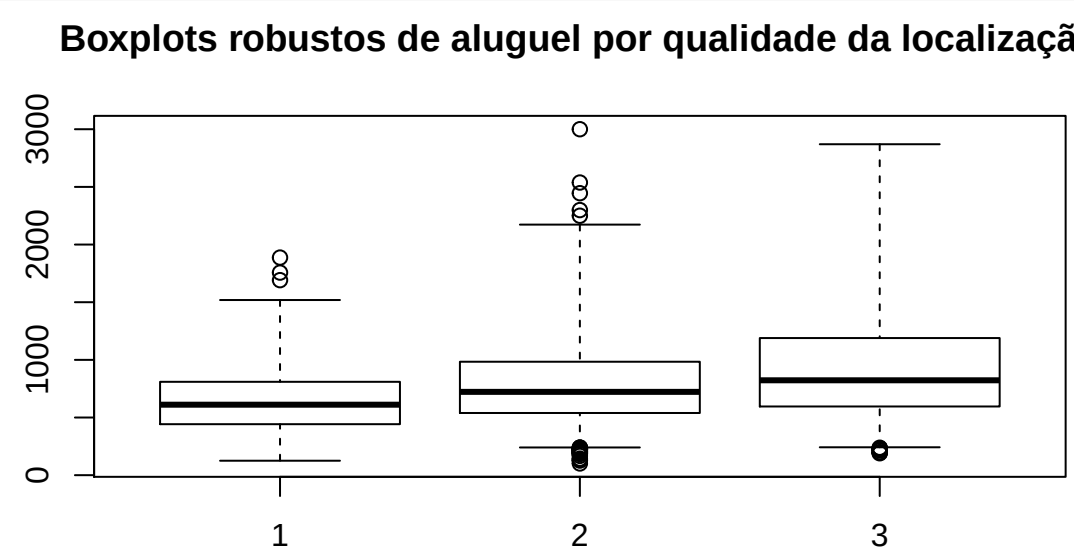
2.2 Variáveis explicativas categóricas

A análise descritiva das variáveis categóricas passará por uma inspeção dos boxplots robustos para avaliar dispersão da variável resposta e um gráfico de perfis médios para entender as tendências de localização.

Com base nos boxplots robustos da variável *loc*, que mostra a qualidade da localização, é possível perceber:

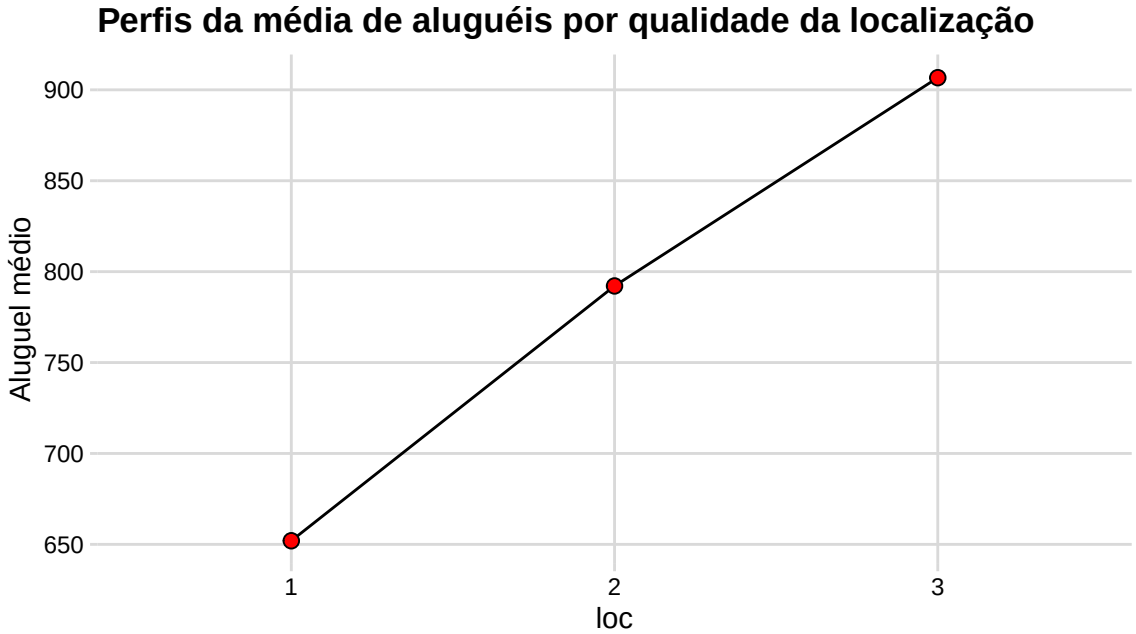
- as medianas dos salários são crescentes, intuitivamente, com a qualildade da região.
- a dispersão também aumenta, como indicado pela amplitude do intervalo inter quartil.
- no grupo de qualidade *loc=2* existem alguns pontos que podem ser caracterizados por terem aluguéis anormalmente baixos para a sua tendência. O mesmo acontece no grupo *loc=3*.
- No grupo *loc=1* e *loc=2* temos ponto que se destacam por serem aluguéis superiores à tendência dos grupos.

```
robustbase::adjbox(R ~ loc, data=df, main='Boxplots robustos de aluguel por qualidade da localizaçã
```



O gráfico de perfis para a qualidade da localização confirma a tendência evidenciada pelas medianas. Porém, devido à assimetria positiva dos aluguéis, as médias por grupo de qualidade mostram uma tedência ascendente mais clara

```
df %>%
  group_by(loc) %>%
  summarise(mean_r = mean(R), .groups='drop') %>%
  ggplot(aes(x=loc, y = mean_r))+
  geom_line(group=1) +
  geom_point(size=2.4, pch=21, fill='red') +
  labs(title='Perfis da média de aluguéis por qualidade da localização', y = 'Aluguel médio')
```

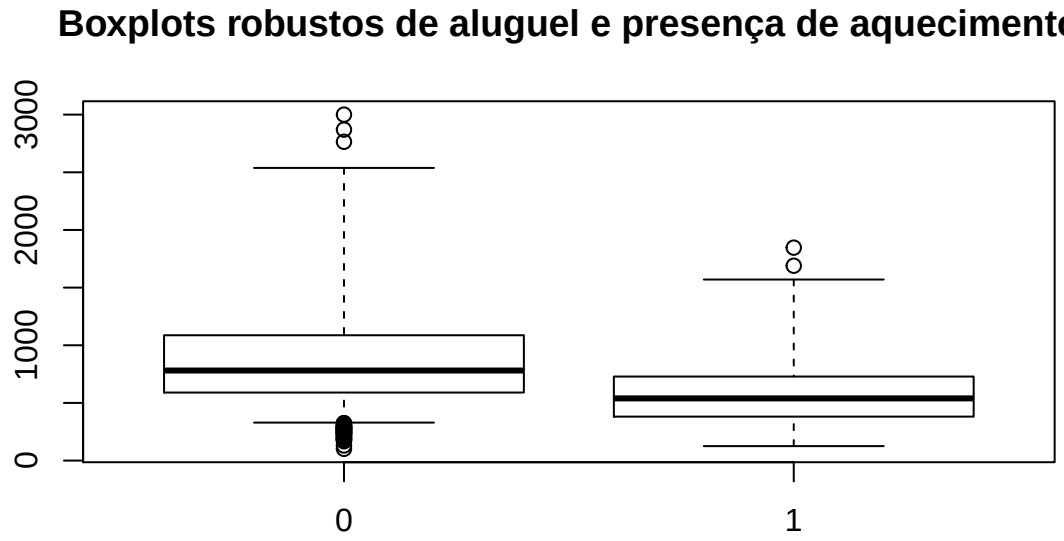


Quando se compara os aluguéis de acordo com a presença de aquecimento central, existem algumas distinções:

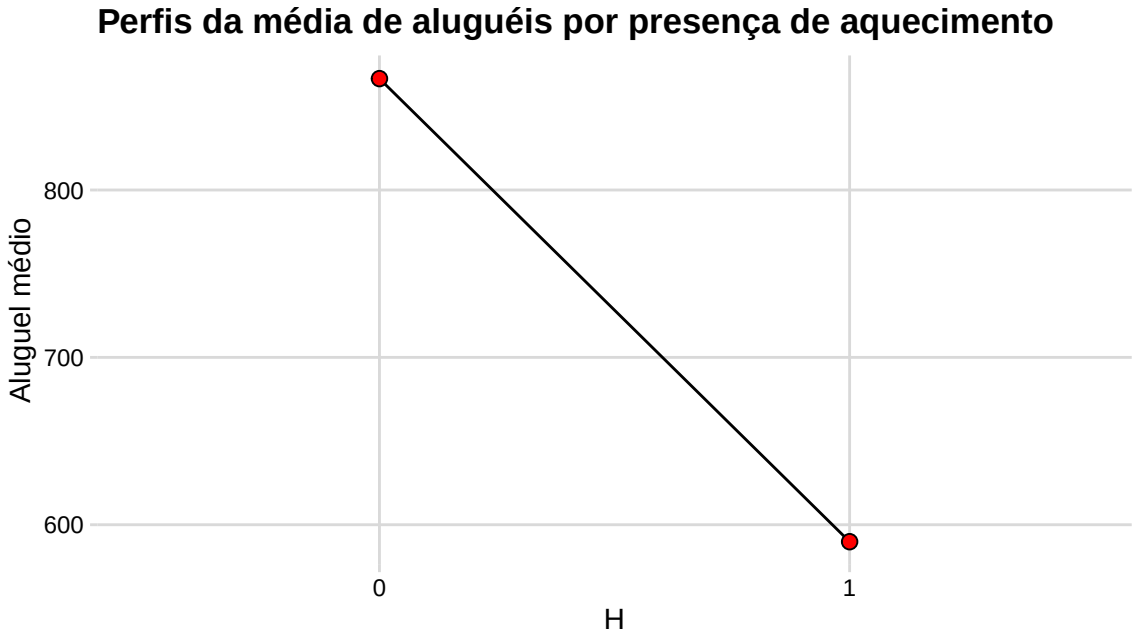
- imóveis sem aquecimento central tem uma media bem próxima ao primeiro quartil do aluguel dos imóveis que têm aquecimento.
- não temos imóveis com aluguéis excessivamente baixos para a categoria de imóveis sem aquecimento. Porém existem aluguéis excessivamente baixos no grupo de imóveis com aquecimento.

- as medianas são bem diferentes entre os dois grupos.
- o gráfico de perfis mostra a mesma tendência, com imóveis tendo aluguéis mais caros em média quando possuem aquecimento central.

```
robustbase::adjbox(R ~ H, data=df, main='Boxplots robustos de aluguel e presença de aquecimento')
```




```
df %>%
  group_by(H) %>%
  summarise(mean_r = mean(R), .groups='drop') %>%
  ggplot(aes(x=H, y = mean_r))+
  geom_line(group=1) +
  geom_point(size=2.4, pch=21, fill='red') +
  labs(title='Perfis da média de aluguéis por presença de aquecimento', y = 'Aluguel médio')
```



2.3 Ajuste de modelos

Vamos ajustar um modelo gamma duplo para os parâmetros de localização e disperção tentando explica-los utilizando as variáveis explicativas que estamos estudando. Para tanto, vamos recorrer ao método de Akaike para a seleção das variáveis de um sub-modelo, partindo do modelo apenas com o intercepto e adicionando variáveis com o objetivo de minimizar o AIC do modelo. Vamos realizar esse procedimento para o componente μ

da localização, mantendo o modelo do parâmetro de dispersão σ com somente o intercepto. Quando o ajuste de μ ficou satisfatório, realizamos o mesmo processo de seleção para a dispersão. Por fim, somente por completude, vamos realizar o procedimento mais uma vez de seleção de variáveis explicativas de μ, porém usando o modelo com as variáveis selecionadas no lugar de apenas usar o intercepto. A lógica é checar se nada pode vir a mudar bruscamente com as estimativas da localização após a escolha das variáveis de σ. O código a seguir realiza esse procedimento.

```
fit_0 <- gamlss(R ~ 1,
                sigma.formula = ~ 1,
                data=df, family = GA(),
                control = gamlss.control(trace=F))

fit_mu <- stepGAIC(
  fit_0,
  scope =  list(
    'lower'= ~1,
    'upper' = ~ Fl + A + H + loc
    ),
  what = 'mu',
  direction = 'forward', trace = 0
)
```

```
## Start:  AIC= 28615.58
##  R ~ 1
```

```
fit_sigma_mu <- stepGAIC(
  fit_mu,
  scope =
    list(
      'lower'= ~1,
      'upper' = ~ Fl + A + H + loc
      ),
  what = 'sigma',
  direction = 'forward', trace = 0
)
```

```
## Start:  AIC= 27778.59
##  ~1
```

```
fit_sigma <- gamlss(R ~ 1,
                    sigma.formula = ~ Fl + A + H + loc,
                    data=df,
                    family = GA(),
                    control = gamlss.control(trace=F))


fit_mu_sigma_2 <- stepGAIC(
  fit_sigma,
  scope =
    list(
      'lower'= ~1,
      'upper' = ~ Fl + A + H + loc
      ),
  what = 'mu',
```

	Nulo	Localização	Local. e Disper.	Local. e Disper. -FL	Local. e Disper. -FL -H
μ (Intercept)	6.699***	2.865***	1.911***	1.852***	1.796***
	(0.010)	(0.571)	(0.606)	(0.604)	(0.610)
σ (Intercept)	−0.779***	−0.982***	4.935***	5.351***	6.009***
	(0.015)	(0.016)	(0.909)	(0.877)	(0.827)
μ Fl		0.011***	0.011***	0.011***	0.011***
		(0.000)	(0.000)	(0.000)	(0.000)
μ H1		−0.300***	−0.288***	−0.287***	−0.287***
		(0.023)	(0.025)	(0.025)	(0.024)
μ loc2		0.191***	0.201***	0.201***	0.204***
		(0.031)	(0.034)	(0.033)	(0.033)
μ loc3		0.264***	0.270***	0.273***	0.275***
		(0.033)	(0.035)	(0.035)	(0.035)
μ A		0.002***	0.002***	0.002***	0.002***
		(0.000)	(0.000)	(0.000)	(0.000)
σ A			−0.003***	−0.003***	−0.004***
			(0.000)	(0.000)	(0.000)
σ loc2			−0.107*	−0.100*	−0.105*
			(0.056)	(0.056)	(0.056)
σ loc3			−0.158***	−0.153**	−0.162***
			(0.061)	(0.060)	(0.060)
σ H1			0.076*	0.065	
			(0.041)	(0.040)	
σ Fl			0.001		
			(0.001)		
Desvio	28611.58	27764.59	27708.47	27710.84	27713.33
AIC	28615.58	27778.59	27732.47	27732.84	27733.33
Num. obs.	1969	1969	1969	1969	1969

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

Statistical models

```
  direction = 'forward', trace =0
)
```

```
## Start:  AIC= 28555.12
##  R ~ 1
```

A tabela a seguir mostra o resultado do método anterior. Importante mencionar que o modelo selecionado correspondente à parcela μ não se alterou tomando a disperção com o intercepto ou com o submodelo para o sigma. Todavia, o modelo para a dispersão possui uma variável não significativa ao nível de 10%, nomeadamente *FL*, por isso propomos mostramos o modelo sem ela.

Após a remoção da variável *FL*, notamos que *H* também se torna não significativa ao nível de 10% e ajustamos o modeo retirando ela também.

```
md_list = list(
  'Nulo' = fit_0,
  'Localização' = fit_mu,
  'Local. e Disper.' =  fit_sigma_mu,
  'Local. e Disper. -FL ' =  update(fit_sigma_mu, ~  A + H + loc, what='sigma'),
  'Local. e Disper. -FL -H' =  update(fit_sigma_mu, ~  A + loc, what='sigma')
)
```

Por via das dúvidas, decidmos relizar um teste de razão de verossimilhanças para testar a exclusão simultânea de H e FL. O teste é significativo ao nível de 10% (p valor de ~8%), sugerindo que o modelo que contém ambos os efeitos, a hipótese alternativa como preferível.

```
LR.test(alternative = fit_sigma_mu, null= update(fit_sigma_mu, ~  A + loc, what='sigma'))
```

```
##  Likelihood Ratio Test for nested GAMLSS models.
##  (No check whether the models are nested is performed).
##
##        Null model: deviance= 27713.33 with  10 deg. of freedom
##  Altenative model: deviance= 27708.47 with  12 deg. of freedom
##
##  LRT = 4.86787 with 2 deg. of freedom and p-value= 0.08769108
```

Porém, ao realizarmos o mesmo teste apenas para a remoção de Fl os resultados estão em concordância com o teste de Wald na tabela das regressões, conforme podemos ver abaixo.

```
LR.test(alternative = fit_sigma_mu, null= update(fit_sigma_mu, ~ H +  A + loc, what='sigma'))
```

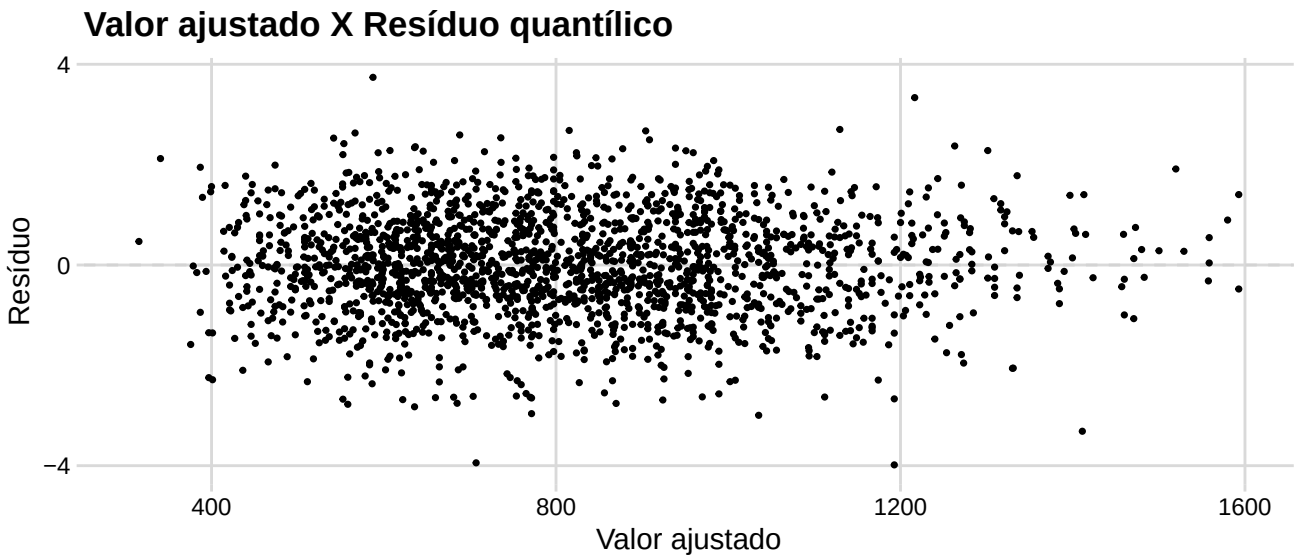
```
##  Likelihood Ratio Test for nested GAMLSS models.
##  (No check whether the models are nested is performed).
##
##        Null model: deviance= 27710.84 with  11 deg. of freedom
##  Altenative model: deviance= 27708.47 with  12 deg. of freedom
##
##  LRT = 2.3748 with 1 deg. of freedom and p-value= 0.1233074
```

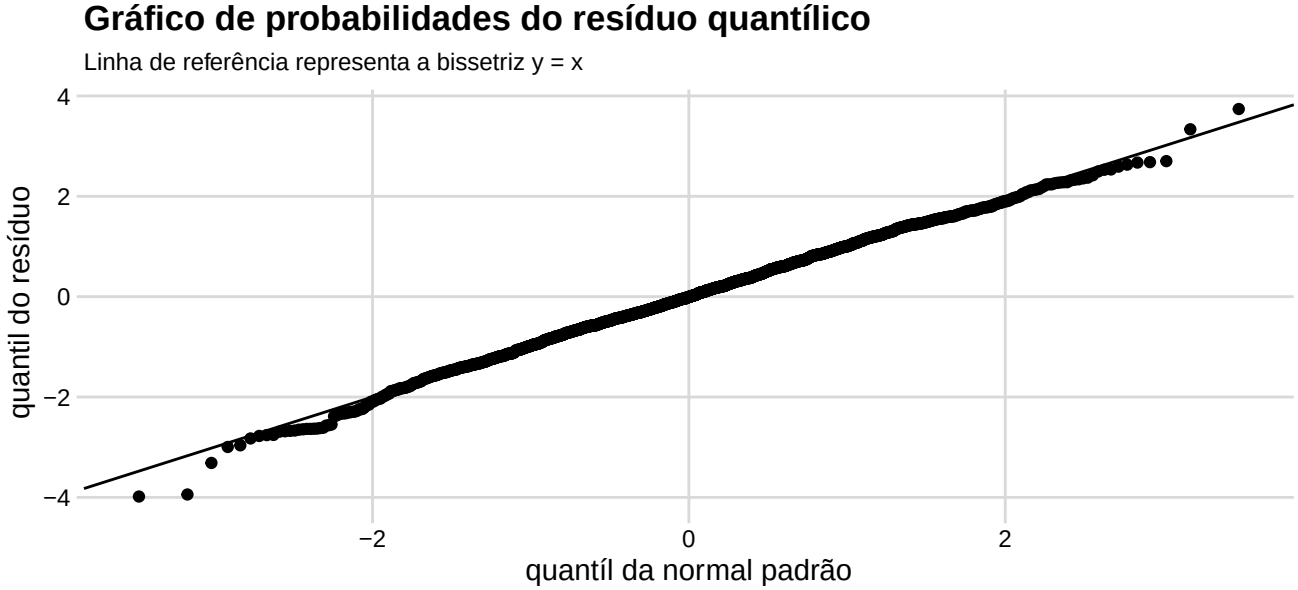
Para resolver esse impasse, vamos nos basear em 2 argumentos. O primeiro mais prático, onde no enunciado do exercício nos diz para não manter variáveis que não sejam significantes à um nível de 10% de significância. Vamos adotar aqui está se referindo ao teste de Wald. O segunto argumento seria o apelo ao princípio da parcimônia, onde preferimos um modelo mais simple, ou seja com menos parâmetros. Como não temos nenhum motivo maior para manter H, vamos retirá-la do componente de dispersão. Assim o nosso modelo final nessa sessão é o mais a direita na tabela apresentada.

2.4 Análise de resíduos

```
md <- md_list[[5]]
```

Com os gráficos de resíduos abaixo, podemos ver que o ajuste do modelo gamma duplo que concluímos na sessão anterior consegue controlar a varibilidade de maneira satisfatória. Apenas parece existir um pouco de heterocedasticidade na região direita do gráfico do valor ajustado contra o resíduo, onde a padrão da nuvem de pontos é mais estreito que o resto.

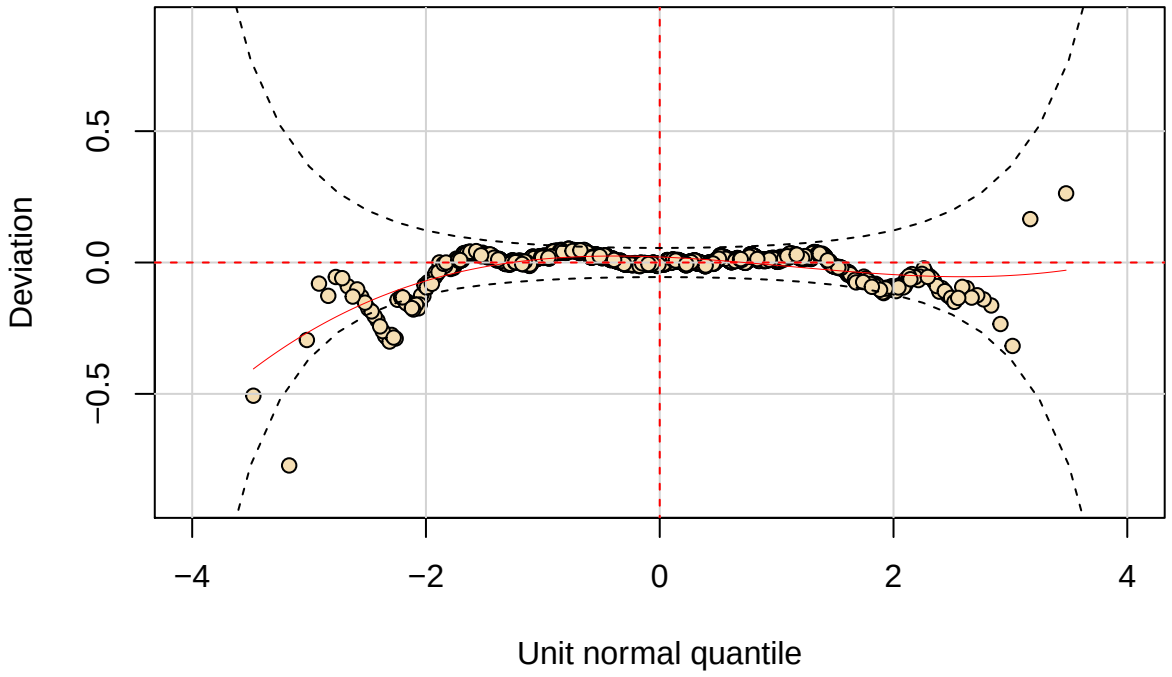




Com relação ao gráfico normal de probabilidades, podemos ver alguns leves desvios da distribuição normal antes do quantíl -2 e mais adiante do quantil 2. Mesmo assim o ajuste nos parece adequado.

A inspeção do gráfico wormplot também levanta os mesmos pontos que a inspeção do gráfico qq no tocante à região antes do quantíl -2. Todavia não há a sobreposição nos quantis mais altos, diminuindo um pouco a preocupação levantada pela análise do gráfico de probabilidades normal.

```
gamlss::wp(resid=resid(md), ylim.all = .9)
```



2.5 Interpretação de parâmetros

Vamos focar a interpretação de parâmtros do componente da média, uma vez que esses são os mais interessantes e interpretáveis para o problema de explicar os aluguéis em Monique em 1993. Reproduzimos aqui a tabela do modelo tentativo até então.

	modelo interpretado
μ (Intercept)	1.796 (0.610)***
μ Fl	0.011 (0.000)***
μ H1	−0.287 (0.024)***
μ loc2	0.204 (0.033)***
μ loc3	0.275 (0.035)***
μ A	0.002 (0.000)***
σ (Intercept)	6.009 (0.827)***
σ A	−0.004 (0.000)***
σ loc2	−0.105 (0.056)*
σ loc3	−0.162 (0.060)***
Desvio	27713.33
AIC	27733.33
Num. obs.	1969

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Statistical models

- intercepto: aqui o intercepto não tem uma interpretação direta, uma vez que um apartamento com metragem nula ou construído no ano zero não faz muito sentido.
- *Fl*: quando o imóvel têm um metro quadrado a mais, é esperado que o seu alguel seja, todo o resto constante, $(100 \cdot (e^{0.011} - 1))\% = 1.01\%$ mais caro.
- *H1*: se um imóvel não possui sistema de aquecimento central, é esperado que seu aluguel médio seja $100 \cdot (1 - e^{-0.287})\% = 25\%$ menor que um apartamento idêntico em todas as outras variáveis.
- *A*: a cada ano que se passa é esperado que o aluguel de um apartamento seja $(100 \cdot (e^{0.002} - 1))\% = 0.20\%$ mais caro em média, todo o resto constante.
- *loc2*: quando um imóvel está localizado em uma região de qualidade próxima à média, é esperado que seu

aluguél seja $(100 \cdot (e^{0.204} - 1))\% = 22.63\%$ mais caro quando comparado com um imóvel numa região de qualidade baixa.

- *loc3*: quando um imóvel está localizado em uma região de qualidade acima da média, é esperado que seu aluguél seja $(100 \cdot (e^{0.275} - 1))\% = 31.65\%$ mais caro quando comparado com um imóvel numa região de qualidade baixa.

Interpretação de parâmetros associados ao componente de dispersão:

- *A*: devido ao sinal negativo, podemos ver que, com o passar dos anos, é esperado que a dispersão dos aluguéis seja ligeiramente menor.
- *loc2* e *loc3*: nesses dois grupos, é esperado que a dispersão dos aluguéis seja menor quando comparado à dispersão dos imóveis em regiões de qualidade piores.

2.6 Ajuste via splines

Vamos ajustar uma o modelo colocando a variável A, que representa o ano de construção do imóvel, como um spline cúbico natural. Isso permitirá controlar para o efeito do tempo nos aluguéis atráves de uma relação mais flexível que a linear. Abaixo mostramos o resultado das estimativas sob esse novo modelo e sob o modeo anterior que estávamos estudando. Como a variáve A está presente em ambos os componentes do modelo gamma duplo, vamos ajustar o modelo via spline em ambos.

```
md_spline <-  gamlss(R ~ Fl + cs(A, df=3) + H + loc,
                     sigma.formula = ~ cs(A, df=3) + loc,
                     data=df,
                     family = GA(),
                     control = gamlss.control(trace = F)
                     )
```

	linear	spline cúbico nat.
μ (Intercept)	1.796 (0.610)***	2.763 (0.590)***
μ Fl	0.011 (0.000)***	0.010 (0.000)***
μ H1	−0.287 (0.024)***	−0.284 (0.024)***
μ loc2	0.204 (0.033)***	0.208 (0.032)***
μ loc3	0.275 (0.035)***	0.291 (0.034)***
μ A	0.002 (0.000)***	
σ (Intercept)	6.009 (0.827)***	7.125 (0.784)***
σ A	−0.004 (0.000)***	
σ loc2	−0.105 (0.056)*	−0.106 (0.056)*
σ loc3	−0.162 (0.060)***	−0.170 (0.060)***
μ cs(A, df = 3)		0.002 (0.000)***
σ cs(A, df = 3)		−0.004 (0.000)***
Desvio	27713.33	27600.14
AIC	27733.33	27632.14
Num. obs.	1969	1969

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

Statistical models

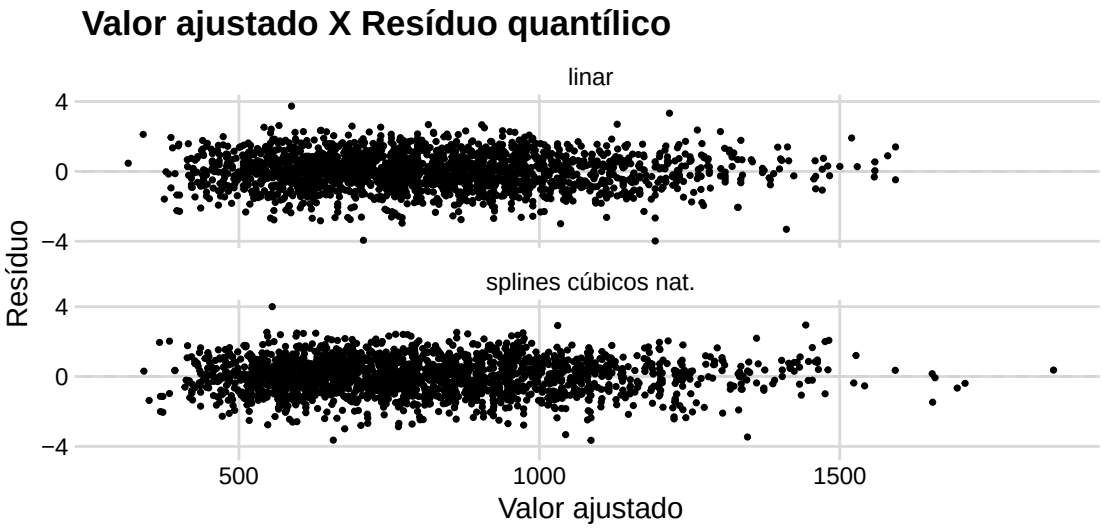
As estimativas dos demais coeficientes ficaram estáveis, com excessão do coeficiente associado aos imóveis de boa qualidade, que ficou 5.82% maior. Isso mostra que, controlando melhor para a variável tempo, o efeito nos aluguéis quando o imóvel está em uma localização melhor é ainda mais evidente. Outro ponto importante a melhora no AIC, tendo um número quase 100 pontos menor quando a variável A é modelada pelo spline cúbico natural. Outra maneira de se ver essa melhora é via um teste de razão de verossimilhanças, que tem um p-valor bem pequeno, sugerindo a adoção do modelo mais geral, i.e. o modelo com splines.

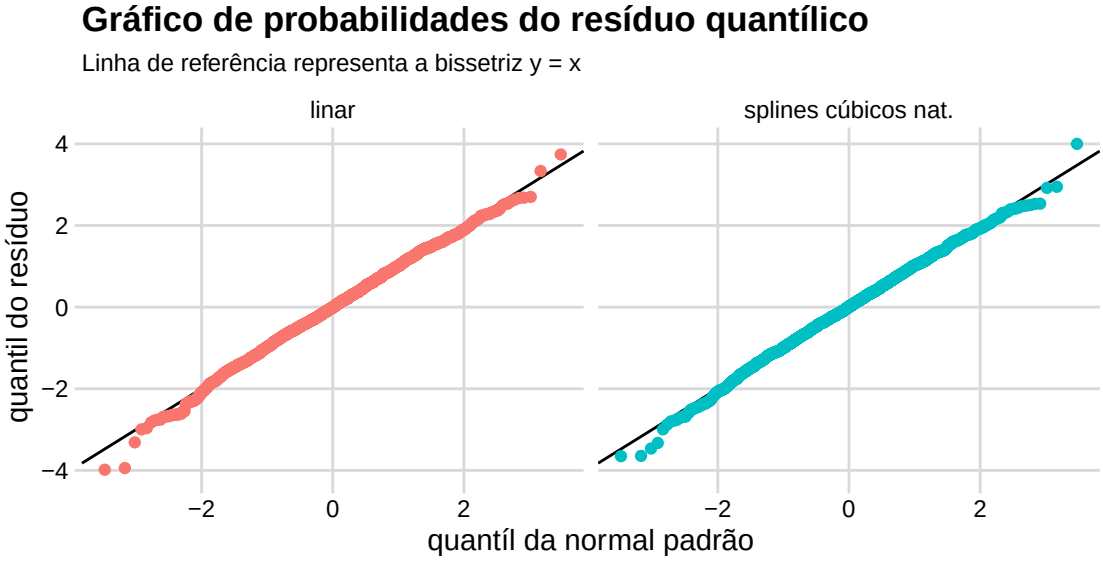
```
LR.test(md, md_spline)
```

```
##  Likelihood Ratio Test for nested GAMLSS models.
##  (No check whether the models are nested is performed).
##
##        Null model: deviance= 27713.33 with  10 deg. of freedom
##  Altenative model: deviance= 27600.14 with  16.00131 deg. of freedom
##
##  LRT = 113.1954 with 6.001308 deg. of freedom and p-value= 0
```

2.6.1 Análise de resíduos

A análise de resíduos parece apresentar também uma melhora no ajuste. O gráfico qq teve os problemas associados às pontas atenuados pelo ajuste com spline cúbico natural. O gráfico de resíduo contra a resposta ajustada parece evidenciar um controle melhor pelo modelo não paramétrico da região que afunilava a variação à direita do gráfico, deixando a dispersão dos pontos mais uniforme.





Por fim o gráfico de wormplot também mostra uma melhora, uma vez que não temos mais sobreposições da banda de confiança tão grande na região antes do quantil -2 da normal padrão.

```
wp_1 <- ~{wp(resid = resid(md), ylim.all = 1)
title('linear')
```

```
}
wp_2 <- ~{wp(resid = resid(md_spline), ylim.all = 1)
title('spline cúbico nat.')
}
```

```
plot_grid(wp_1, wp_2)
```