

SME0823 - Regressão II

Regressão para os preços de casa do estado da Virginia
(EUA)

Fernando Hitoshi Masumoto (8556609)

Joao Guedes (8724923)

Dezembro 2023

1 Introdução

Este trabalho visa desenvolver um modelo de regressão para prever o valor de uma casa no estado da Virgínia, Estados Unidos, com base em suas características.

1.1 Descrição dos dados

O conjunto de dados possui 3025 linhas por 15 colunas. Segue uma lista com as variáveis do conjunto de dados:

- *yearbuilt* - Inteiro
- *finsqft* - Inteiro
- *cooling* - Inteiro
- *bedroom* - Inteiro
- *fullbath* - Inteiro
- *halfbath* - Inteiro
- *lotsize* - Ponto Flutuante
- *totalvalue* - Ponto Flutuante (variável resposta)
- *esdistrict* - Caracter
- *msdistrict* - Caracter
- *hsdistrict* - Caracter
- *censustract* - Ponto Flutuante
- *age* - Inteiro
- *condition* - Caracter
- *fp* - Booleano

A variável censustract apesar de ser numérica não se trata de nenhuma medida, se trata de um código referente àquele município dentro dos padrões norte americanos.

O modelo que será proposto tem como objetivo predizer os valores da variável *totalvalue*.

1.2 Tratamento dos dados

Uma verificação preliminar dos dados mostra que há 5 linhas que contém o valor “NA” (*Not Available*), 1 (um) na coluna *yearbuilt* e outras 4 (quatro) na coluna *halfbath*. Como tratativa inicial pretendia-se remover todas as linhas por conta da baixa representatividade que elas apresentavam.

Tabela 1: Linhas que contém NA

yearbuilt	finsqft	cooling	bedroom	fullbath	halfbath	lotsize	totalvalue	...	fp
2001.00	1860.00	No Central Air	2.00	2.00	NA	21.84	369800.00	...	1.00
NA	1728.00	Central Air	3.00	3.00	1.00	0.04	284300.00	...	0.00
2017.00	1872.00	Central Air	2.00	2.00	NA	112.04	561800.00	...	0.00
1930.00	984.00	No Central Air	2.00	1.00	NA	1.99	115300.00	...	0.00
1986.00	1508.00	Central Air	3.00	2.00	NA	2.68	298800.00	...	1.00

A variável *yearbuilt* carrega a mesma informação que *age*, inclusive, nota-se que o valor da correlação linear de Pearson para ambas é -1, ou seja, são variáveis colineares, portanto, esta coluna é removida junto com o seu valor “NA”. Como também são retiradas as 4 linhas da coluna *halfbath* que também contêm “NA”.

Outro ponto que foi observado é a presença de 6 linhas com valores duplicados dentro da base de dado e que também foram removidas, deixando apenas uma de cada. A seguir estão as linhas duplicadas:

Tabela 2: Linhas duplicadas

finsqft	...	totalvalue	esdistrict	msdistrict	hsdistrict	censustract	age	condition	fp
1230.00	...	167700.00	Agnor-Hurt	Jouett	Albemarle	107.00	38.00	Average	0.00
1230.00	...	167700.00	Agnor-Hurt	Jouett	Albemarle	107.00	38.00	Average	0.00
1428.00	...	177100.00	Agnor-Hurt	Burley	Albemarle	107.00	47.00	Average	0.00
1428.00	...	177100.00	Agnor-Hurt	Burley	Albemarle	107.00	47.00	Average	0.00
1220.00	...	183400.00	Baker-Butler	Sutherland	Albemarle	103.00	19.00	Average	1.00
1220.00	...	183400.00	Baker-Butler	Sutherland	Albemarle	103.00	19.00	Average	1.00
1402.00	...	240700.00	Baker-Butler	Sutherland	Albemarle	103.00	12.00	Average	0.00
1402.00	...	240700.00	Baker-Butler	Sutherland	Albemarle	103.00	12.00	Average	0.00
1402.00	...	242600.00	Baker-Butler	Sutherland	Albemarle	103.00	12.00	Average	0.00
1402.00	...	242600.00	Baker-Butler	Sutherland	Albemarle	103.00	12.00	Average	0.00
1606.00	...	252500.00	Greer	Jouett	Albemarle	108.00	8.00	Average	0.00
1606.00	...	252500.00	Greer	Jouett	Albemarle	108.00	8.00	Average	0.00

Após todo tratamento mencionado, o conjunto de dados final possui um total de 3015 linhas com 13 covariáveis e uma variável resposta (14 colunas). E por fim, foram escolhidas as colunas: ”cooling”, ”bedroom”, ”fullbath”, ”halfbath”, ”esdistrict”, ”msdistrict”, ”hsdistrict”, ”censustract”, ”condition”, ”fp” como fatores (”as.factor”).

1.3 Exploração dos dados

1.3.1 Variável Resposta - TOTALVALUE

Uma das primeiras características a serem notadas enquanto era feita a exploração dos dados, foi a distribuição bastante irregular da variável *totalvalue* conforme pode ser visto no histograma a esquerda na Figura 1. Por conta disso, decidiu-se por tomar seu logaritmo com o intuito de homogeneizar sua distribuição. Além disso, sabe-se que essa transformação não traz dificuldades tremendas no que tange a interpretabilidade do modelo.

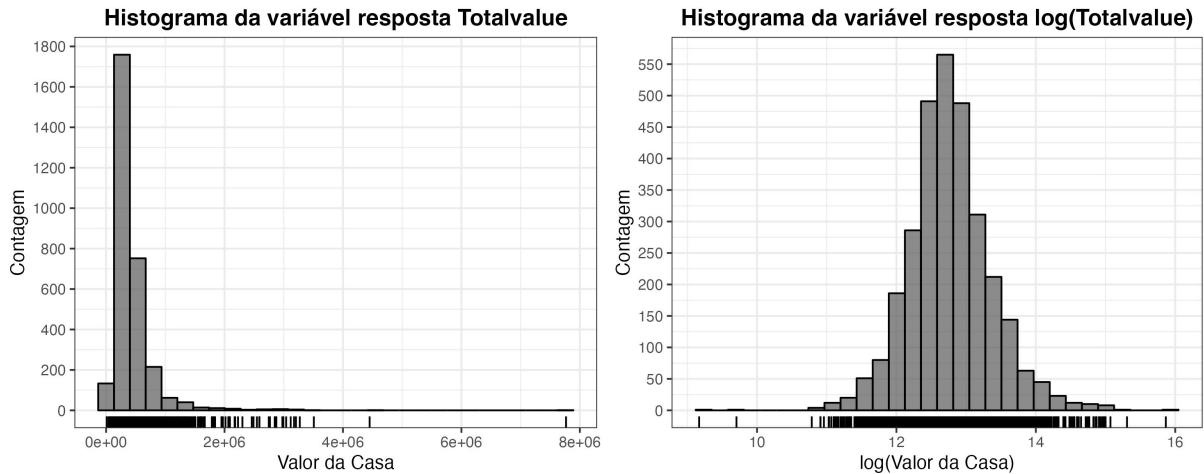


Figura 1: Histogramas para a variável resposta *totalvalue* e da sua transformação por Logaritmo natural

Como pode ser observado, a variável transformada $\ln(\text{totalvalue})$ se tornou uma variável com distribuição aparentemente simétrica.

1.3.2 Variáveis Categóricas

Com o intuito de se debruçar sobre a ligação da variável resposta com algumas variáveis categóricas, foram feitos gráficos de Boxplot. As variáveis foram agrupadas de acordo com o seu significado e ordenadas pela sua média para que fosse facilitada a interpretação.

O primeiro grupo (Figura 2) comprehende dados referentes ao número de banheiros e quartos dentro do imóvel

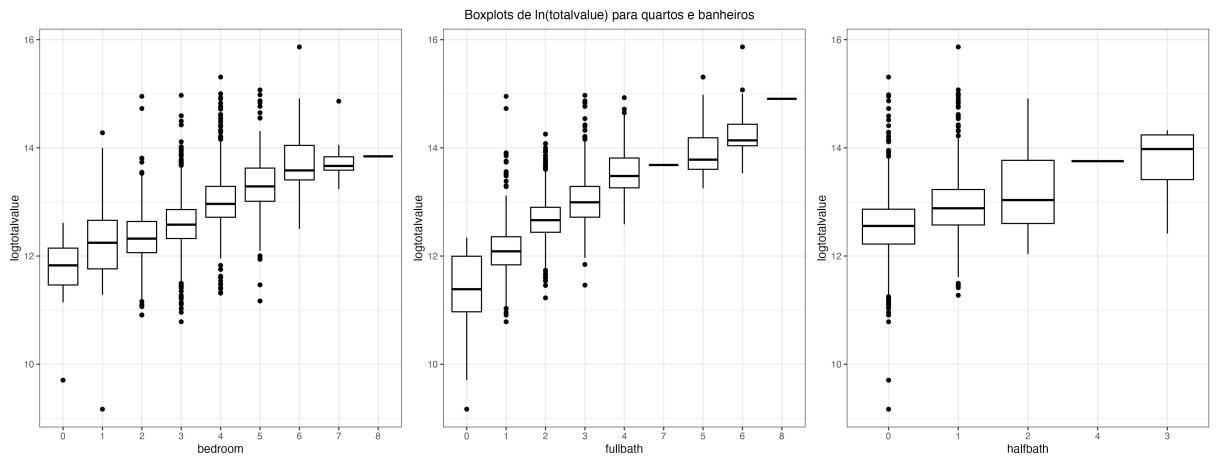


Figura 2: Histogramas para a variável resposta *totalvalue* e da sua transformação por Logaritmo natural

O segundo grupo comprehende dados (Figura 3) referentes a separação feita para os diferentes distritos registrados na base de dados.

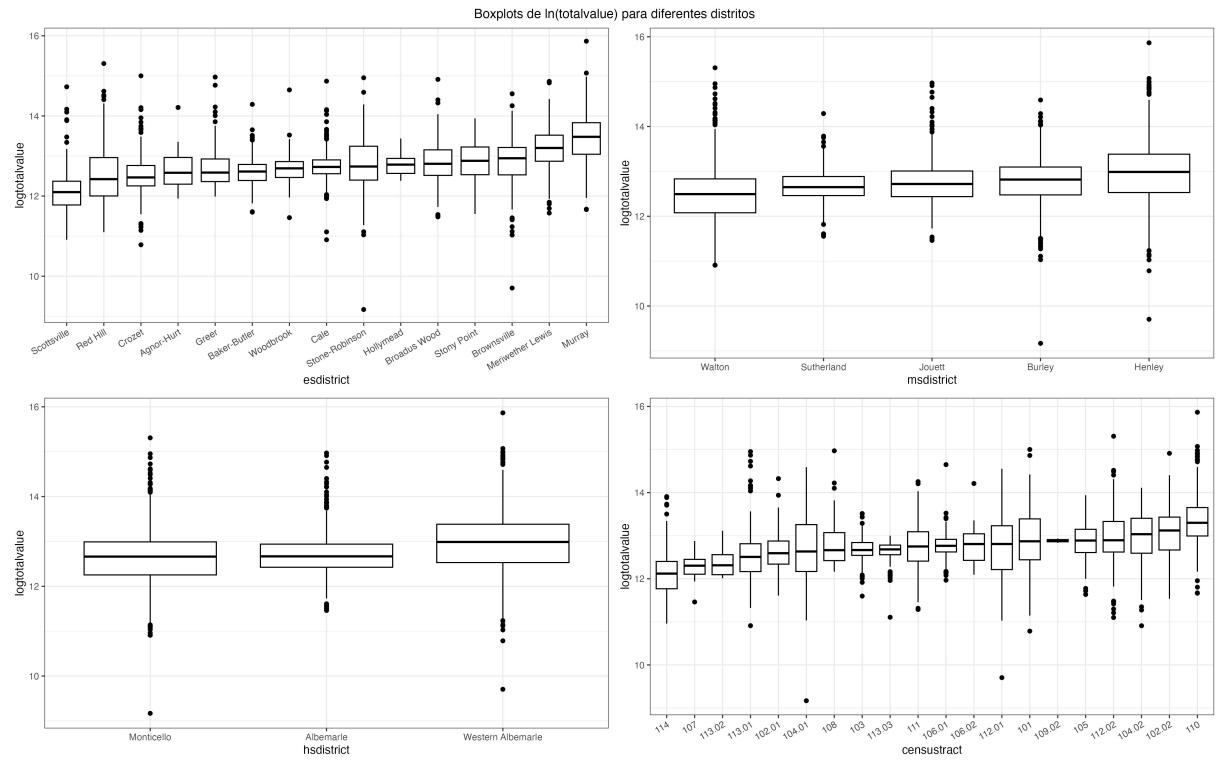


Figura 3: Histogramas para a variável resposta *totalvalue* e da sua transformação por Logaritmo natural

Por fim, foram agrupadas as variáveis que não se encaixaram em nenhum dos dois grupos anteriores (Figura 4) trata-se das variáveis, “cooling”, “condition” e “fp”.

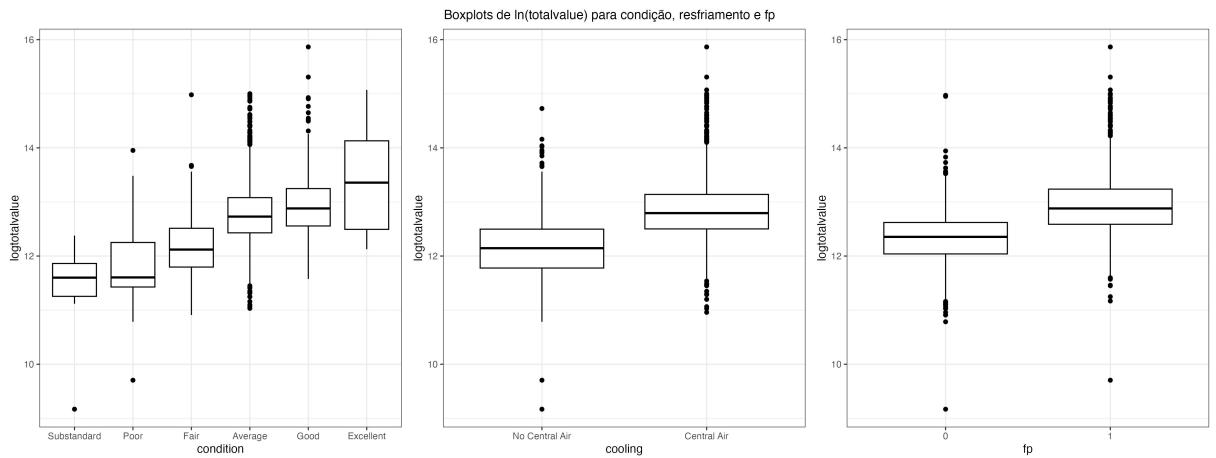


Figura 4: Histogramas para a variável resposta *totalvalue* e da sua transformação por Logaritmo natural

1.3.3 Variáveis Contínuas

A mesma ideia para explorar a relação entre as variáveis contínuas foi utilizada, porém com gráficos de dispersão (Figura 5). Além do gráfico também foram calculados os valores dos seus coeficiente de correlação de Spearman, uma vez que o gráfico de dispersão não mostra nenhuma relação explicitamente linear.

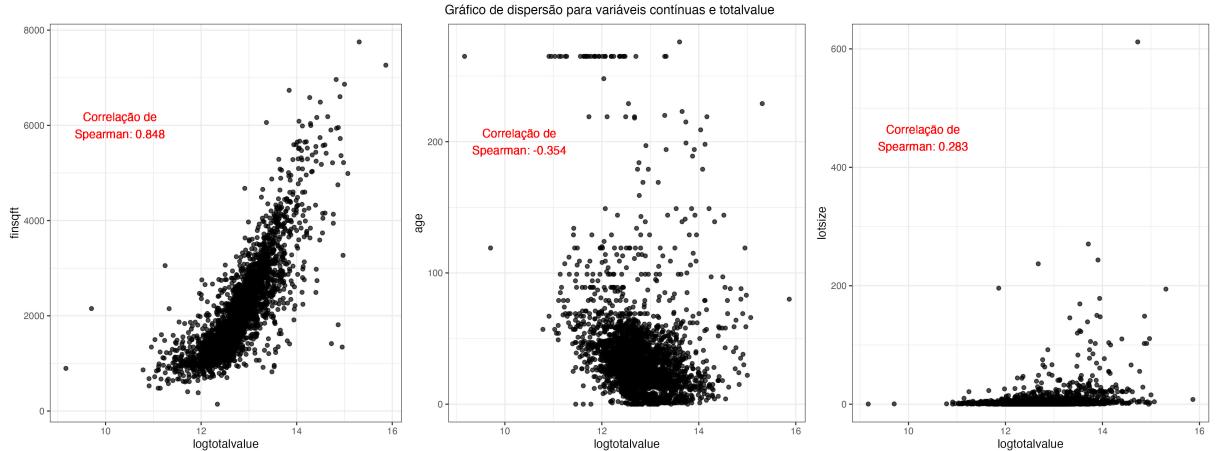


Figura 5: Gráfico de dispersão para a variável *logtotalvalue* e outras variáveis contínuas

1.4 Verificação e seleção do modelo

A seleção de variáveis foi feita, primeiramente, observando os gráficos da seção anterior e depois utilizando a função StepGaic para as variáveis selecionadas.

Com relação as variáveis que denotam localidade, esdistrict, msdistrict, hsdistrict, censustract, como pode ser observado na Figura 3, há uma variação muito menos expressiva para esses parâmetros do que aquelas que são referentes a propriedades físicas do imóvel. Devido a quantidade maior de categorias que

existem e principalmente quando se pensa em utilizar a interação foram escolhidas as categorias a partir dos seguintes critérios:

- esdistrict: 3 categorias com menor e maior mediana ('Scottsville', 'Red Hill', 'Crozet', 'Brownsville', 'Meriwether Lewis', 'Murray')
- msdistrict: A maior e menor mediana ('Walton', 'Henley')
- hsdistrict: Aquele com maior mediana ('Western Albemarle')
- censustract: 4 categorias com menor e 3 com a maior mediana ('114', '107', '113.02', '113.01', '104.02', '102.02', '110')

Com essas categorias selecionadas, cada uma delas foi dicotomizada para transformá-las em covariáveis do modelo. Com isso obtemos um conjunto de dados com um total de 27 colunas para nossas tentativas iniciais para criar o modelo de regressão. Utilizou-se então a função StepGAIC para automatizar a seleção as variáveis de maior importância.

1.5 Modelos

Como mencionado anteriormente a primeira impressão eram de que a *logtotalvalue* possue distribuição simétrica, após a transformação, e como é sabido pela própria natureza da variável resposta, sabe-se que se trata de uma variável positiva (preço). Com isso em mente, procurou-se por distribuições que pudessem atender a essas restrições e que também tivessem versatilidade o bastante para que fosse feito um bom ajuste aos dados.

Foram então escolhidas as seguintes distribuições, Log-Normal(LOGNO), distribuições da família T(TF2) e distribuição T Box-Cox (BCT).

2 Resultados

2.1 *Generalized Akaike Information Criterion (GAIC)*

Segue a tabela de resultados para o AIC de diferentes funções de ligação.

Tabela 3: Valores de AIC para cada uma das famílias utilizadas

	Log-Normal	Família T	BoxCox T
AIC	77428.9	-195.034	-221.342

Utilizando o GAIC chegou-se as seguintes variáveis para os modelos:

- LOGNO: Sinsqft, lotsize, condition, fullbath, fp, age, esScottsville, esMurray, centralair, esBrownsville, esMeriwetherLewis, esRedHill, halfbath, bedroom;
- TF2: finsqft, fullbath, esScottsville, lotsize, fp, condition, esMurray, age, esMeriwetherLewis, esRedHill, centralair, esBrownsville, msWalton, esCrozet;
- BCT: finsqft, fullbath, esScottsville, lotsize, fp, esMurray, condition, age, esMeriwetherLewis, esRedHill, centralair, esBrownsville, msWalton, esCrozet.

2.2 Regressão

As imagens a seguir são os resumos da regressão dos dados.

```

Family: c("LOGNO", "Log Normal")

Call: gamlss(formula = totalvalue ~ finsqft + lotsize + condition +      fullbath + fp + age + esScottsville + esMurray +
  centralair + esBrownsville + esMeriwetherLewis +      esRedHill + halfbath + bedroom, family = LOGNO(mu.link = "identity",
  sigma.link = "log"), data = dtse, trace = FALSE)

Fitting method: RSC()

-----
Mu link function: identity
Mu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.112e+01 3.941e-02 282.261 < 2e-16 ***
finsqft 3.333e-04 8.083e-06 41.231 < 2e-16 ***
lotsize 6.713e-03 2.622e-04 25.598 < 2e-16 ***
condition 1.308e-01 8.770e-03 14.914 < 2e-16 ***
fullbath 1.116e-01 8.068e-03 13.828 < 2e-16 ***
fp 1.412e-01 1.198e-02 11.783 < 2e-16 ***
age -1.286e-03 1.450e-04 -8.868 < 2e-16 ***
esScottsville -3.207e-01 2.399e-02 -13.371 < 2e-16 ***
esMurray 3.552e-01 2.196e-02 16.174 < 2e-16 ***
centralair 9.470e-02 1.756e-02 5.392 7.49e-08 ***
esBrownsville 6.217e-02 1.668e-02 3.727 0.000197 ***
esMeriwetherLewis 1.921e-01 1.885e-02 10.195 < 2e-16 ***
esRedHill -1.056e-01 2.105e-02 -5.015 5.61e-07 ***
halfbath 1.439e-02 9.410e-03 1.530 0.126223
bedroom -9.866e-03 6.995e-03 -1.410 0.158529
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----
Sigma link function: log
Sigma Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.34949 0.01288 -104.8 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----
No. of observations in the fit: 3015
Degrees of Freedom for the fit: 16
Residual Deg. of Freedom: 2999
at cycle: 2

Global Deviance: 77400.28
AIC: 77432.28
SBC: 77528.46
*****
```

Figura 7: Sumário para o modelo de Log-Normal

```
*****
Family: c("TF2", "t Family 2")

Call: gamlss(formula = logtotalvalue ~ finsqft + fullbath + esScottsville + lotsize + fp + condition + esMurray +
   age + esMeriwetherLewis + esRedHill + centralair + esBrownsville + msWalton + esCrozet, family = TF2(mu.link = "identity",
   sigma.link = "log", nu.link = "logshiftto2"), data = dtse, control = gamlss.control(n.cyc = 200), trace = FALSE)

Fitting method: RSC()

-----
Mu link function: identity
Mu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.126e+01 3.471e-02 324.555 < 2e-16 ***
finsqft 3.416e-04 8.432e-06 40.510 < 2e-16 ***
fullbath 1.023e-01 5.149e-03 19.873 < 2e-16 ***
esScottsville -3.279e-01 2.273e-02 -14.427 < 2e-16 ***
lotsize 1.037e-02 6.167e-04 16.821 < 2e-16 ***
fp 1.310e-01 9.340e-03 14.027 < 2e-16 ***
condition 9.553e-02 7.746e-03 12.333 < 2e-16 ***
esMurray 3.104e-01 2.481e-02 12.513 < 2e-16 ***
age -1.499e-03 1.578e-04 -9.499 < 2e-16 ***
esMeriwetherLewis 1.925e-01 1.490e-02 12.924 < 2e-16 ***
esRedHill -1.478e-01 2.336e-02 -6.327 2.87e-10 ***
centralair 8.647e-02 1.556e-02 5.558 2.96e-08 ***
esBrownsville 5.831e-02 1.288e-02 4.526 6.26e-06 ***
msWalton 3.558e-02 1.385e-02 2.568 0.0103 *
esCrozet -2.414e-02 1.578e-02 -1.529 0.1263
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----
Sigma link function: log
Sigma Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.30254 0.04211 -30.93 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----
Nu link function: logshiftto2
Nu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2040 0.1762 1.158 0.247

No. of observations in the fit: 3015
Degrees of Freedom for the fit: 17
Residual Deg. of Freedom: 2998
at cycle: 21

Global Deviance: -229.0341
AIC: -195.0341
SBC: -92.84107
*****
```

Figura 9: Sumário para o modelo de família T

```
*****
Family: c("BCT", "Box-Cox t")

Call: gamlss(formula = logtotalvalue ~ finsqft + fullbath + esScottsville + lotsize + fp + esMurray + condition +
   age + esMeriwetherLewis + esRedHill + centralair + esBrownsville + msWalton + esCrozet, family = BCT(mu.link = "identity",
   sigma.link = "log", nu.link = "identity", tau.link = "log"),      data = dtse, control = gamlss.control(n.cyc = 200),
   trace = FALSE)

Fitting method: RS()

-----
Mu link function: identity
Mu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.126e+01 3.392e-02 332.038 < 2e-16 ***
finsqft 3.423e-04 6.456e-06 53.023 < 2e-16 ***
fullbath 1.040e-01 6.103e-03 17.040 < 2e-16 ***
esScottsville -3.292e-01 2.195e-02 -14.996 < 2e-16 ***
lotsize 1.207e-02 8.162e-04 14.785 < 2e-16 ***
fp 1.274e-01 9.534e-03 13.358 < 2e-16 ***
esMurray 3.376e-01 2.620e-02 12.886 < 2e-16 ***
condition 9.457e-02 7.684e-03 12.308 < 2e-16 ***
age -1.388e-03 1.370e-04 -10.134 < 2e-16 ***
esMeriwetherLewis 1.902e-01 1.505e-02 12.634 < 2e-16 ***
esRedHill -1.512e-01 2.310e-02 -6.545 6.97e-11 ***
centralair 8.121e-02 1.512e-02 5.372 8.40e-08 ***
esBrownsville 5.597e-02 1.310e-02 4.272 1.99e-05 ***
msWalton 3.461e-02 1.376e-02 2.516 0.0119 *
esCrozet -2.481e-02 1.558e-02 -1.592 0.1115
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----
Sigma link function: log
Sigma Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.3236 0.0247 -175.1 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----
Sigma link function: log
Sigma Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.3236 0.0247 -175.1 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----
Nu link function: identity
Nu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.388 1.528 4.181 2.99e-05 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----
Tau link function: log
Tau Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.22245 0.07356 16.62 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----
No. of observations in the fit: 3015
Degrees of Freedom for the fit: 18
   Residual Deg. of Freedom: 2997
      at cycle: 11

Global Deviance: -257.342
      AIC: -221.342
      SBC: -113.1376
*****
```

Figura 11: Sumário para o modelo de Box Cox T

Além dos sumários também são apresentados gráficos de resumo para os modelos ajustados:

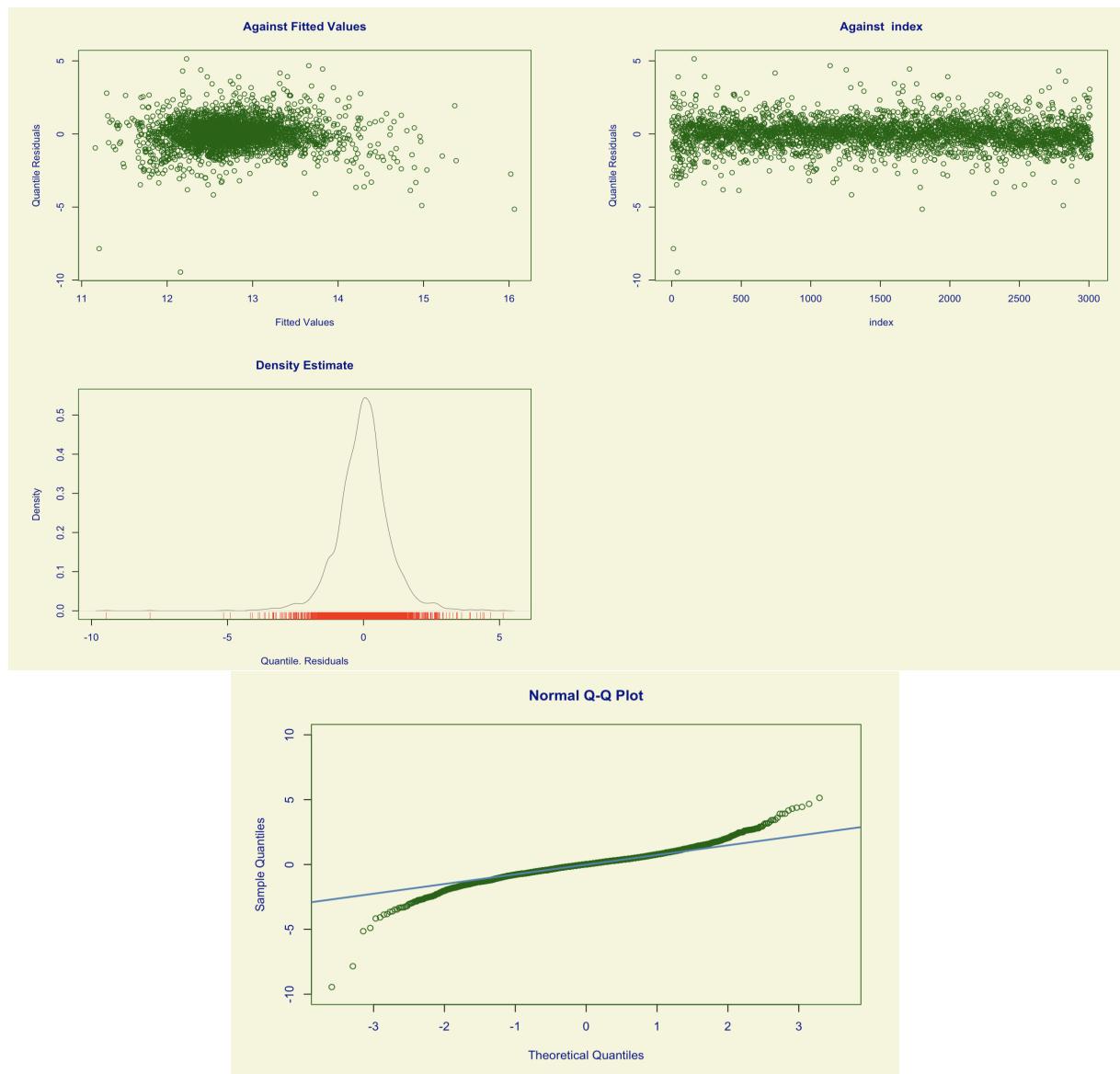


Figura 13: Resumo do ajuste para o modelo de Log-Normal

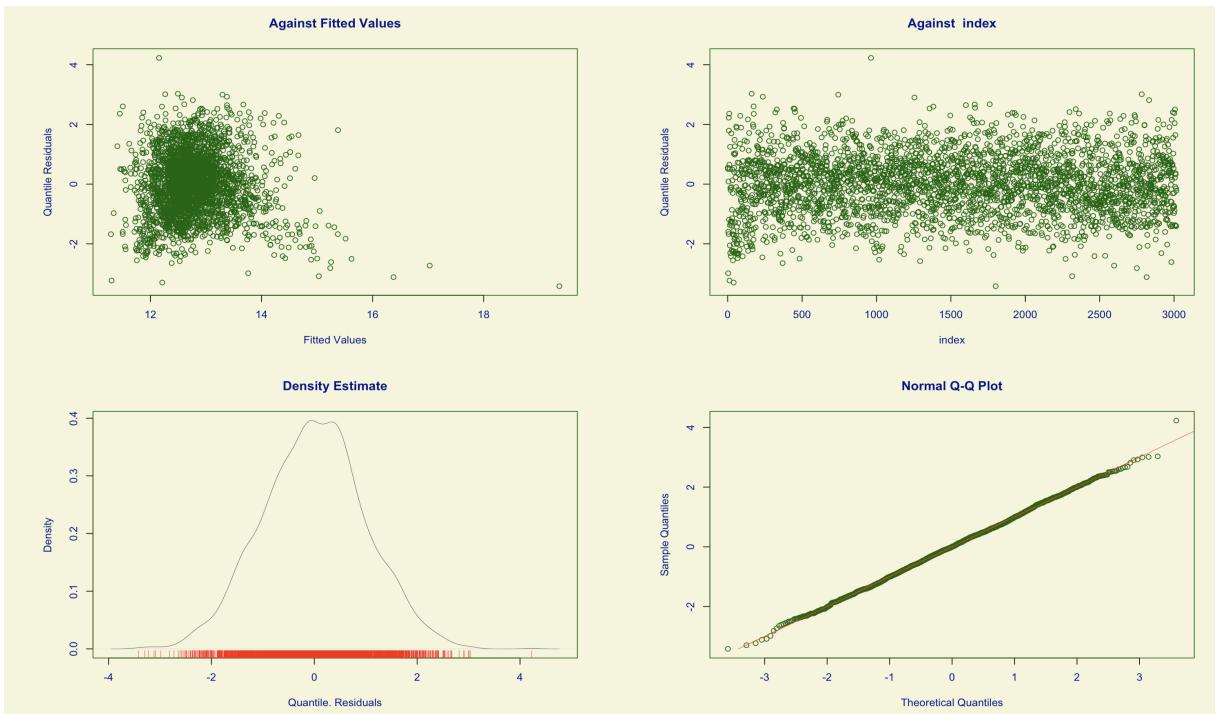


Figura 14: Resumo do ajuste para o modelo de Log-Normal

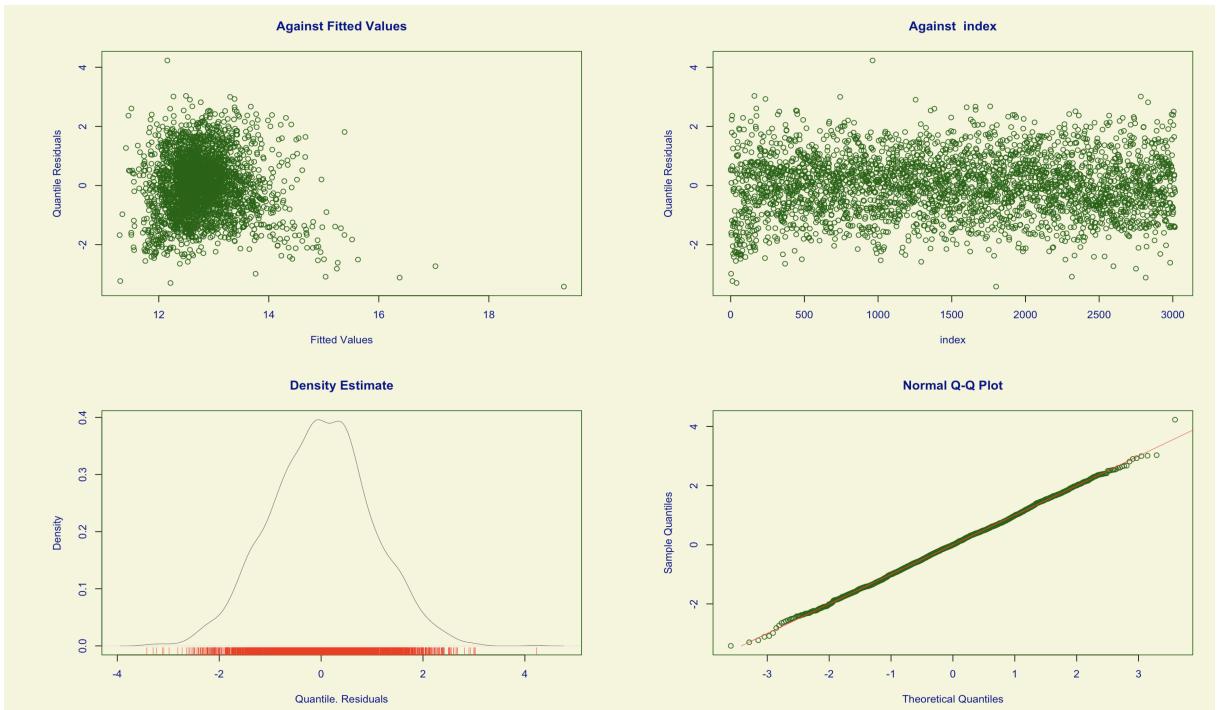


Figura 15: Resumo do ajuste para o modelo de Log-Normal

3 Discussão

Dentro dos modelos ajustados para o problema proposta, os modelos TF e BCT trouxeram os melhores ajustes, baseando-se nos gráficos apresentados nas Figuras 14 e 15. Contudo, seus poderes preditivos não foram bons, como pode ser observado no gráfico de quantis por valores ajustados.

Já o modelo Log-Normal, cuja simplicidade era a primeira ideia de ajuste que se teve, devido a semelhança visual com a distribuição normal, após a transformação, não teve o melhor ajuste conforme observado em seu qqplot, porém aos avaliarmos os valores dos quantis por ajustados, observa-se uma distribuição mais homogênea dos dados, mesmo que ainda bastante concentrada num elipsóide. Isso também foi motivador para que se procusse modelos de caudas mais pesadas.

Apesar do bom ajuste observado, faltou a este trabalho, debruçar-se sobre os significados para os coeficientes dos parâmetros de cada modelo, mesmo que não se tivesse interesse inicial em interpretar um modelo como BCT devido a sua complexidade.