Universidade Federal de Santa Catarina
Centro de Florianópolis
Departamento de Informática
e Estatística

Jessica Xafranski Rodrigues (18250264)
Isabelle Pinheiro (13103605)
Lucas Pagotto Tonussi (18250265)

T2 Técnicas Estatísticas de Predição

Florianópolis
23 de novembro de 2019

Jessica Xafranski Rodrigues (18250264)
Isabelle Pinheiro (13103605)
Lucas Pagotto Tonussi (18250265)

T2 Técnicas Estatísticas de Predição

Trabalho de casa. Sistemas de Informação. UFSC.

Orientador: Prof. Dr. Luiz Ricardo Nakamura

Universidade Federal de Santa Catarina
Centro de Florianópolis
Departamento de Informática
e Estatística

Florianópolis
23 de novembro de 2019

Jessica Xafranski Rodrigues (18250264)
Isabelle Pinheiro (13103605)
Lucas Pagotto Tonussi (18250265)

T2 Técnicas Estatísticas de Predição

Trabalho de casa. Sistemas de Informação. UFSC.

Comissão Examinadora

Prof. Dr. Luiz Ricardo Nakamura
Universidade Federal de Santa Catarina
Orientador

Prof. Dr. Luiz Ricardo Nakamura
Universidade Federal de Santa Catarina

Prof. Dr. Luiz Ricardo Nakamura
Universidade Federal de Santa Catarina

Florianópolis, 23 de novembro de 2019

Dedico este trabalho a todos aqueles que, de alguma forma,
auxiliaram para a concretização desta etapa.

Agradecimentos

Agradecemos aos nossos pais, mães e a você.

"Porquanto a vida é mais preciosa do que o alimento,
e o corpo, mais importante do que as roupas."
(Lucas 12.23)

Resumo

Resumo em português.

Palavras-Chave: 1. regressão linear múltipla. 2. predição

Abstract

Abstract in english.
Keywords: 1. multiple linear regression. 2. prediction

Lista de figuras

Lista de tabelas

Lista de Siglas e Abreviaturas

RLM	*Regressão Linear Múltipla*
UFSC	*Universidade Federal de Santa Catarina*
COEFAP	*Coeficiente de aproveitamento*
AREA	*Área do lote*
ACLDECL	*Sentido predominante da topografia por lote*
FRENTE	*Frente do lote*

Sumário

1 Introdução

Nesse presente trabalho, vamos apresentar a regressão linear múltipla de um modelo de amostras de área, coeficiente de aproveitamento do lote na área geo-referenciada, sentido predominante da topografia por lote, e frente do lote, para explicar o aumento ou diminuição do preço do imóvel. Essas amostras foram retiradas do trabalho de Rocha (2016) [1].

O trabalho citado acima, trata o assunto de Geoprocessamento, sendo uma ampla disciplina que visa fornecer técnicas, ferramentas e métodos para compreender melhor o globo terrestre.

O globo terrestre pode ser mapeado por valores contínuos de latitude e longitude fornecendo assim uma malha de localização ponto a ponto, conhecido como Geo-referenciamento. O Geoprocessamento em si, vem das imagens de satélite, que fornecem imagens multi-banda geo-referenciadas, chamadas *rasters*.

Essas imagens especiais contém informações variadas, dependendo da necessidade, do propósito do satélite, e de projetos de mapeamento feitos pelos donos do satélite em órbita.

O trabalho que fora citado também trata de Regressão Linear Múltipla, que é um tópico importante de Estatística. E que serve para criar modelos sofisticados que combinam múltiplas variáveis independentes para explicar uma variável dependente. Resumidamente: $\hat{Y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \varepsilon \qquad (i = 1, \ldots, n)$, onde ε se trata do erro residual.

As próximas seções tratam do modelo de regressão linear múltipla, lembrando que estamos nos baseando no trabalho de monografia de Rocha (2016) [1].

2 Implementações

2.0.1 Apresentar um problema real

O problema do nosso trabalho envolve predizer como a área total, a topografia, a distância em metros da frente do lote e a área construída no lote afetam o preço de um lote.

2.0.2 Variável dependente e as variáveis independentes

As variáveis independentes são:

1. Coeficiente de aproveitamento (*sigla.* COEF_AP)
   Relação entre a área total construída de uma edificação e a área total do lote [1].

2. Área do lote (*sigla.* AREA)
   Área em metros quadrados do lote [1].

3. Sentido predominante da topografia por lote (*sigla* ACL_DECL)
   Situação da topografia do lote em relação à rua [1].

4. Frente do lote (*sigla.* FRENTE).
   Distância em metros da frente do lote [1].

Logo mais, explicaremos como a autora fez para mapear os valores utilizados na tabela 1.

A variável dependente é:

1. Valor do imóvel (PRECO) [1].

Justificação

Segundo o trabalho, de Rocha (2016) [1], essas variáveis influenciam no valor do imóvel. Ainda segundo a autora, as variáveis: “área”, “coeficiente de aproveitamento” e “frente do lote” influenciam positivamente no valor do imóvel e possuem correlação forte $r \geq 0,7$, enquanto a variável ACL_DECL influencia negativamente no valor do imóvel $r \geq 0.3$.

[1] XV - coeficiente de aproveitamento é a relação entre a área total construída de uma edificação e a área total da gleba ou lote; Fonte: <https://www.legisweb.com.br/legislacao/?id=316888>.

2.0.3 Apresente o conjunto de dados

nam	preco	area	dist	coef_ap	acl_decl	frente
1	25.000,00	359,87	200,00	1,50	0,75	19,00
2	45.000,00	353,12	500,00	3,00	1,00	12,18
3	45.000,00	341,89	40,00	1,50	1,00	14,98
4	20.000,00	435,99	200,00	1,50	0,75	12,18
5	25.000,00	341,77	500,00	1,50	0,75	12,93
6	20.000,00	366,22	500,00	1,50	0,75	15,00
7	30.000,00	300,69	200,00	1,50	1,00	17,00
8	25.000,00	470,98	200,00	1,50	0,80	11,89
9	40.000,00	373,13	500,00	1,50	1,00	12,82
10	50.000,00	343,12	500,00	3,00	1,00	11,87
11	50.000,00	361,56	500,00	1,50	1,00	39,27
12	55.000,00	538,32	500,00	1,50	1,00	28,96
13	40.000,00	470,69	500,00	1,50	1,00	44,15
14	60.000,00	970,72	40,00	3,00	0,95	40,03
15	50.000,00	351,31	200,00	1,50	1,00	40,82
16	70.000,00	2.057,57	40,00	3,00	1,00	40,10
17	40.000,00	531,65	40,00	3,00	1,00	14,48
18	40.000,00	366,49	500,00	1,50	0,90	15,24
19	40.000,00	401,00	500,00	1,50	1,00	13,09
20	25.000,00	356,88	500,00	1,50	0,90	12,09
21	35.000,00	360,45	200,00	1,50	1,00	12,01
22	35.000,00	468,28	200,00	3,00	0,90	42,57
23	120.000,00	1.004,52	40,00	3,00	0,90	75,40
24	20.000,00	315,51	200,00	1,50	0,75	17,16
25	30.000,00	382,16	500,00	1,50	0,85	54,21
26	30.000,00	293,22	500,00	1,50	0,90	36,12
27	30.000,00	577,84	500,00	1,50	1,00	11,49
28	40.000,00	380,42	200,00	1,50	1,00	12,19
29	30.000,00	504,92	40,00	1,50	0,85	61,49
30	25.000,00	371,89	500,00	1,50	0,80	39,51
31	30.000,00	397,63	200,00	1,50	0,85	31,27

Tabela 1 – Dados Parte 1. Fonte: Rocha (2016) [1].

nam	**preco**	**area**	**dist**	**coef_ap**	**acl_decl**	**frente**
32	30.000,00	354,89	200,00	1,50	1,00	12,52
33	35.000,00	402,76	500,00	1,50	0,90	25,58
34	50.000,00	405,76	40,00	3,00	1,00	42,33
35	50.000,00	382,08	200,00	3,00	1,00	42,78
36	50.000,00	333,64	200,00	3,00	1,00	36,09
37	40.000,00	547,41	200,00	1,50	1,00	13,00
38	50.000,00	561,44	40,00	3,00	1,00	73,11
39	30.000,00	361,96	40,00	1,50	1,00	13,91
40	50.000,00	385,11	200,00	3,00	1,00	12,00
41	30.000,00	355,68	200,00	1,50	1,00	12,08
42	50.000,00	351,25	200,00	3,00	1,00	12,32
43	60.000,00	578,64	200,00	3,00	0,85	48,32
44	60.000,00	354,46	200,00	3,00	0,90	14,50
45	50.000,00	400,30	500,00	3,00	0,90	39,81
46	50.000,00	503,48	200,00	3,00	1,00	48,36
47	30.000,00	357,55	200,00	1,50	0,90	35,85
48	35.000,00	328,33	200,00	3,00	0,85	10,01
49	50.000,00	385,96	40,00	1,50	1,00	12,07
50	60.000,00	369,51	200,00	3,00	0,95	10,28
51	30.000,00	350,02	650,00	3,00	0,90	13,00
52	50.000,00	380,30	500,00	3,00	1,00	12,74
53	20.000,00	388,65	500,00	1,50	0,75	37,00
54	50.000,00	715,12	200,00	3,00	1,00	55,88

Tabela 2 – Dados Parte 2. Fonte: Rocha (2016) [1]

Tomando as tabelas acima como referência, é importante mostrar como a autora fez para mapear esses dados:

1. Coeficiente de aproveitamento (COEF_AP).
2. Área do lote (AREA).
3. Sentido predominante da topografia por lote (ACL_DECL).
4. Frente do lote (FRENTE).

2.0.4 Principais hipóteses

As hipóteses nulas são as seguintes:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

Estamos considerando que β_0 não precisa ser analisado nesse modelo. Dito isso, se algum dos $\beta_i (i = 1 \dots 4)$ for nulo, significa que uma ou mais variáveis independentes estão falhando e não tem correlação com Y (valor do imóvel).

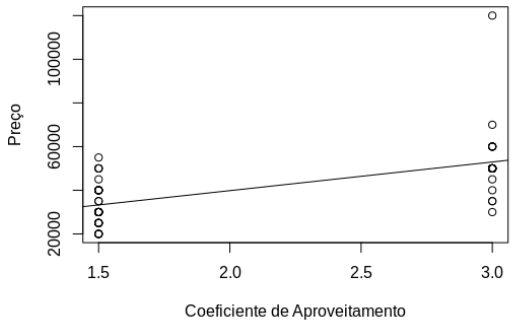
Por outro lado, se todas as variáveis β forem diferentes de 0 e diferentes entre si, significa que elas tem alguma correlação com o valor do imóvel. Ou seja, a hipótese alternativa ficaria assim:

$$H_1 : \beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq 0$$

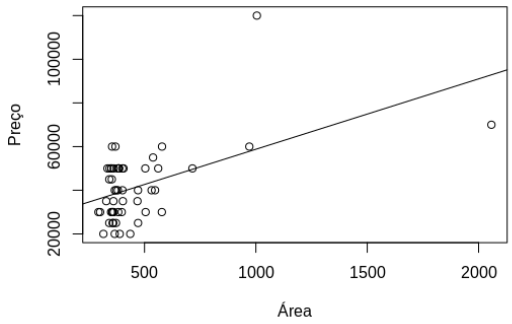
Que nos diz que todos os coeficientes angulares tem influência sobre Y.

2.0.5 Diagramas de Dispersão (2 à 2)

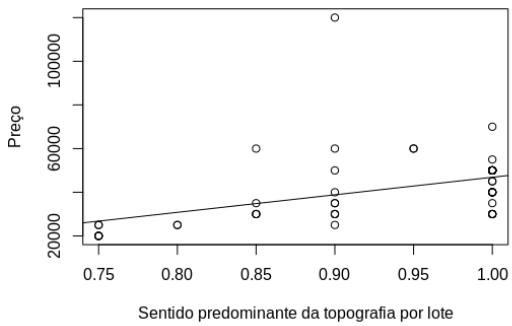
A seguir será exposto as 5 variáveis em estudo (1 dependente, 4 independentes), ou seja: $plot(Y_{\text{preço}_i}, X_{\text{área}_i})$, $plot(Y_{\text{preço}_i}, X_{\text{coefap}_i})$, $plot(Y_{\text{área}_i}, X_{\text{acldecl}_i})$, $plot(Y_{\text{preço}_i}, X_{\text{frente}_i})$ $(i = 0, \ldots, 54)$. A variável **dis**t não será utilizada na RLM deste trabalho, não utilizamos **dist** pois a autora não utilizou também e 4 variáveis já é um número bom de variáveis independentes.



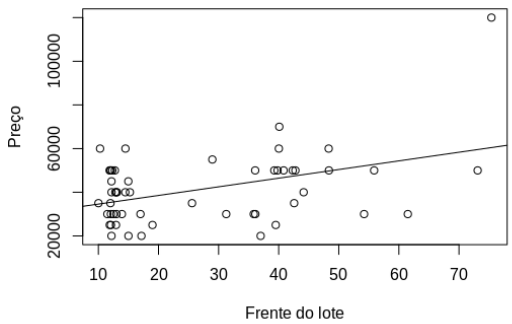
$R^2 = 0,2615, r = 0,5113$



$R^2 = 0,3502, r = 0,5917$



$R^2 = 0,1846, r = 0,4296$



$R^2 = 0,178, r = 0,4219$

Figura 1 – Diagramas de Dispersão (Sem Variáveis Transformadas)

Mas o que os gráficos acima querem nos dizer?

1. Fator de Topografia: corrige as diferenças relativas ao sentido predominante da topografia do imóvel no contexto em que está inserido.

2. Fator de Área: imóveis maiores agregam maior valor ao lote.

3. Fator de Localização: ou fator de transposição, corrige as diferenças relativas ao posicionamento do imóvel no contexto em que está inserido.

4. Fator de Frente: frentes mais extensas agregam maior valor ao lote.

1. Todas as relações com a variável dependente parecem lineares? Se sim, os coeficientes de correlação linear indicam que essa relação é forte?

   Observando os gráficos acima na figura 1, pode-se considerar que as variáveis possuem sim correlação linear, umas mais fortes outras mais fracas. Por exemplo, a variável Coeficiente de Aproveitamento causa um crescimento positivo do valor somente em suas extremidades avaliadas. Já a variável área, mostra uma explosão no valor do preço em seus primeiros valores no gráfico com uma relação muito forte, depois os valores crescem muito lentamente, mas também positivamente linear. No gráfico Sentido predominante da topografia por lote vs Preço pode-se observar um crescimento linear positivo com relação fraca. No último gráfico, as variáveis Frente do lote e Preço possuem relação forte no começo do gráfico e mais fraca conforme a variável independente vai aumentando.

2. Há pontos discrepantes? Os gráficos sugerem alguma transformação?

   Sim, houveram pontos 2 discrepantes no gráfico de resíduos vs preditos. Com a retirada deles o gráfico ficou assim:

Antes Depois

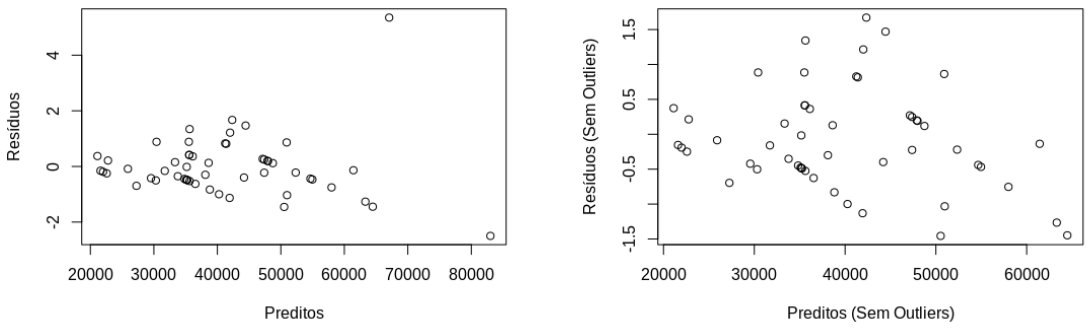


Figura 2 – Resíduos vs Preditos para a equação: $Y_{\text{preço}_i} = -41771,930 + 17,276 \times X_{\text{área}_i} + 8195,227 \times X_{\text{coefap}_i} + 55767.600 \times X_{\text{acldecl}_i} + 221,517 \times X_{\text{frente}_i}$

3. Faça as devidas mudanças sugeridas nos itens (a)–(c) para a sequência do trabalho. Se fizer alguma transformação ou retirada de casos discrepantes, refaça os diagramas.

Olhando para os diagramas de dispersão 2a2 que foram mostrados anteriormente (vide figura 1), observamos alguns pontos que poderiam ser considerados discrepantes, mas não removemos eles por causa do trabalho de Rocha (2016), onde a autora já removeu todos os pontos que foram considerados inválidos para a modelagem. A autora (Rocha, 2016) tinha 60 observações, onde 53 foram utilizadas. E a coluna "distância até a via principal"(em metros) também foi removida pela autora.

Podemos citar, também, que a autora utilizou a seguinte equação em sua modelagem final: $ln(Y_{\text{preço}_i}) = 10.546673 + 0.227991 \times ln(X_{\text{área}_i}) + 0.041921 \times (X_{\text{coefap}_i})^2 - 1.534453 \times (X_{\text{acldecl}_i})^{-1} + 0.003400 \times (X_{\text{frente}_i})^3$. Essas transformações feitas pela autora, não são muito expressivas do ponto de vista numérico. Os índices de determinação não melhoram tanto assim, e inclusive o coeficiente angular de ACL_DECL muda para negativo. Se o coeficiente angular de ACL_DECL muda para negativo significa que a influência de ACL_DECL (transformada) é contrária à sua versão não transformada.

É possível observar pela figura 4 que existe uma aleatoriedade nos pontos do diagrama, de magnitude -1.5 à $+1.5$, isso, é um bom sinal para mostrar que a equação de RLM vai predizer valores de forma razoavelmente boa.

2.0.6 Sobre possíveis transformações das variáveis

O artigo de Rocha (2016) mostra que a equação com variáveis transformadas final, após ajustes, é:

$ln(Y_{\text{preço}_i}) = 10.546673 + 0.227991 \times ln(X_{\text{área}_i}) + 0.041921 \times (X_{\text{coefap}_i})^2 - 1.534453 \times (X_{\text{acldecl}_i})^{-1} + 0.003400 \times (X_{\text{frente}_i})^3$

Tendo em vista essa equação podemos ver que a autora tentou melhorar a variabilidade dos dados, mas a melhoria não foi substancial. E um dos coeficientes ficou negativo (ACL_DECL).

Veja os resultados a seguir:

```
coefap <- geo$coef_ap^2
frente <- geo$frente^3
acldecl <- 1/geo$acl_decl
m4 <- lm(log(geo$preco) ~ log(geo$area) +
      coefap + acldecl + frente, data=geo)
summary(m4)
```

```
lm(formula = log(geo$preco) ~ log(geo$area) + coefap + acldecl +
    frente, data = geo)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.30260 -0.14413 -0.03326  0.09578  0.64391

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.546673   0.660423  15.970  < 2e-16 ***
log(geo$area)  0.227991   0.098045   2.325   0.0242 *
coefap         0.041921   0.009218   4.547 3.58e-05 ***
acldecl       -1.534453   0.258429  -5.938 2.92e-07 ***
frente         0.003400   0.001822   1.867   0.0679 .
```

Os resultados para os coeficietnes angulares não parecem muito sugestivos, e temos que a variável FRENTE tem uma um p-valor maior que 0.05, evidenciando problemas para predição usando frente ao cubo. Resumidamente, o modelo da autora ficou estranho com as variáveis transformadas.

4. Verifique se não há forte relação entre as variáveis independentes.

Para fazer isso precisamos testar 2a2 todas as 4 variáveis independentes.

y=geo$coefap x=geo$acldecl y=geo$coefap x=geo$frente y=geo$coefap x=geo$area	Não existe correlação
y=geo$frente x=geo$acldecl y=geo$frente x=geo$coefap	Não existe correlação.
y=geo$frente x=geo$area	Existe correlação.
y=geo$acldecl x=geo$area y=geo$acldecl x=geo$frente y=geo$acldecl x=geo$coefap	Não existe correlação.
y=geo$area x=geo$acldecl y=geo$area x=geo$coefap	Não existe correlação.
y=geo$area x=geo$frente	Existe correlação.

Tabela 3 – Tabela mostrando as quais testes de correlação foram feitos

Para averiguar se existe correlação entre essas combinações de variáveis independentes, plotamos cada gráfico de dispersão para analisar se existe uma correlação visível, e apenas AREA e FRENTE, ou FRENTE e AREA mostraram uma correlação aparente. As demais não tem correlação alguma (pontos aleatórios).

2.0.7 Equação de regressão e outras estatísticas

Sem transformar variáveis

Dada a equação genérica do modelo de RLM $\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \hat{\epsilon}$, modifica-se essa equação para se igualar às análises que foram realizadas, ficando assim:

$$Y_{\text{preço}_i} = -41771,930 + 17,276 \times X_{\text{área}_i} + 8195,227 \times X_{\text{coefap}_i} + 55767,600 \times X_{\text{acldecl}_i} + 221,517 \times X_{\text{frente}_i}$$

Isso significa:

Que a cada 1 unidade acrescida de área em metros quadrados, aumenta o preço em 17 reais, aproximadamente.

Que a cada 1 unidade acrescida no coeficiente de aproveitamento em metros, aumenta o preço em 8195 reais, aproximadamente.

Que a cada 1 unidade acrescida de "sentido predominante da topografia por lote"em metros, aumenta o preço em 55767 reais, aproximadamente.

Que a cada 1 unidade acrescida de "frente do lote"em metros quadrados, aumenta o preço em 221 reais, aproximadamente.

Com variáveis transformadas

2.0.8 Críticas sobre o modelo ajustado

1. O modelo é útil? (Teste F de utilidade do modelo)

```
anova(modelo)
Analysis of Variance Table

Response: geo$preco
             Df     Sum Sq    Mean Sq F value    Pr(>F)
geo$area      1 3768050386 3768050386 31.0008 1.076e-06 ***
geo$coef_ap   1 3041508426 3041508426 25.0233 7.675e-06 ***
geo$acl_decl  1  975813933  975813933  8.0283  0.006669 **
geo$frente    1  668082305  668082305  5.4965  0.023152 *
Residuals    49 5955804210  121547025
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conforme os resultados de "confint"abaixo, listamos os limites inferiores das variáveis nas colunas $\alpha = 5\%$.

Intervalos de confiança tomando $\alpha = 5\%$

```
confint(modelo)
                     2.5 %      97.5 %
(Intercept)  -74524.984097 -9018.87548
geo$area          4.366277    30.18589
geo$coef_ap    3697.400603 12693.05438
geo$acl_decl  19586.756576 91948.44340
geo$frente       31.641688   411.39239
```

O teste F é utilizado para analisar a variância entre dois conjuntos de dados diferentes e compará-los utilizando o teste de hipóteses.

Testes de Hipóteses:

Todo p-valor $< \alpha$

$$H_0 : \beta_{\text{área}} = 0 \quad || \quad \beta_{\text{coefap}} = 0 \quad || \quad \beta_{\text{acldecl}} = 0 \quad || \quad \beta_{\text{frente}} = 0$$

$$H_1 : \beta_{\text{área}} \neq 0 \quad \&\& \quad \beta_{\text{coefap}} \neq 0 \quad \&\& \quad \beta_{\text{acldecl}} \neq 0 \quad \&\& \quad \beta_{\text{frente}} \neq 0$$

Ao analisarmos os dados gerados pelo CONFINT(MODELO), verificamos a ausência de 0 nos intervalos de confiança do modelo, com isso podendo rejeitar a hipótese nula (H_0).

Intervalos de confiança tomando $\alpha = 5\%$

```
                    2.5 %      97.5 %
(Intercept)  -74524.984097 -9018.87548
geo$area          4.366277    30.18589
geo$coef_ap    3697.400603 12693.05438
geo$acl_decl  19586.756576 91948.44340
geo$frente       31.641688   411.39239

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -41771.930  16298.496  -2.563 0.013501 *
geo$area         17.276      6.424   2.689 0.009761 **
geo$coef_ap    8195.227   2238.198   3.662 0.000614 ***
geo$acl_decl  55767.600  18004.224   3.097 0.003227 **
geo$frente      221.517     94.485   2.344 0.023152 *
```

Para o teste de hipótese H_1:

$\beta_{\text{área}} = 0.009761 \quad \&\& \ \beta_{\text{coefap}} = 0.000614 \quad \&\& \ \beta_{\text{acldecl}} = 0.003227 \quad \&\& \ \beta_{\text{frente}} = 0.023152$

Observando os resultados, obtidos pelo R, acima, podemos ver que os p-valores dados pelo Test F deram todos inferiores à 0.05. Isso nos indica, com segurança, que a variabilidade de Y pode ser explicada em função da variabilidade de

$$\beta_{\text{área}} \quad \&\& \quad \beta_{\text{coefap}} \quad \&\& \quad \beta_{\text{acldecl}} \quad \&\& \quad \beta_{\text{frente}}$$

2. Todas as variáveis independentes têm efeitos significativos (utilizar $\alpha = 5\%$) (Teste t sobre os coeficientes).

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -41771.930  16298.496  -2.563 0.013501 *
geo$area         17.276      6.424   2.689 0.009761 **
geo$coef_ap    8195.227   2238.198   3.662 0.000614 ***
geo$acl_decl  55767.600  18004.224   3.097 0.003227 **
geo$frente      221.517     94.485   2.344 0.023152 *
```

Todos os p-valores acima (coluna **Pr(>|t|)** para o teste T de cada um dos coeficientes angulares nos retornam valores menores que o nível de significância de 0.05 ($\alpha = 5\%$).

3. As variáveis significativas têm sinal como esperado?

   Ao observar os diagramas de dispersão (vide figura 1) 2a2 podemos ver que todas as retas estão inclinadas para cima, ou seja, teoricamente os coeficientes serão positivos no modelo RLM.

4. A análise dos resíduos, os gráficos sugerem que o modelo está adequado?

Antes Depois

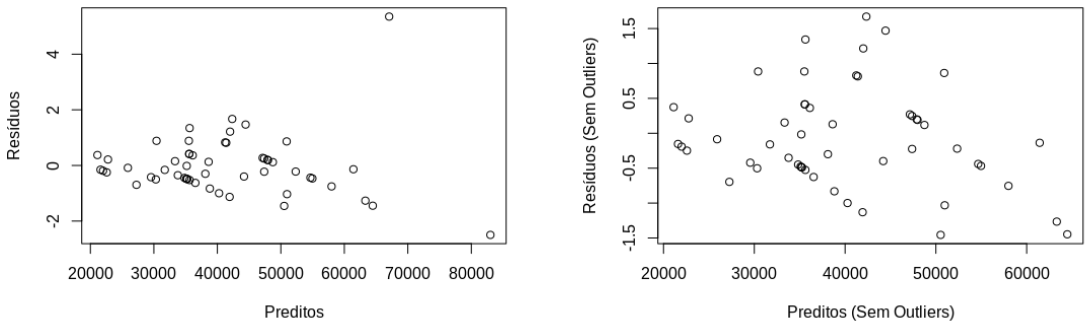


Tabela 4 – Resíduos vs Preditos para a equação: $Y_{\text{preço}_i} = -41771,930 + 17,276 \times X_{\text{área}_i} + 8195,227 \times X_{\text{coefap}_i} + 55767.600 \times X_{\text{acldecl}_i} + 221,517 \times X_{\text{frente}_i}$

É possível observar pela figura 4 que existe aleatoriedade nos pontos do diagrama, de magnitude -1.5 à $+1.5$, isso, é um bom sinal para mostrar que a equação de RLM vai predizer valores de forma razoavelmente boa.

QQPlot e Teste de Shapiro-Wilk

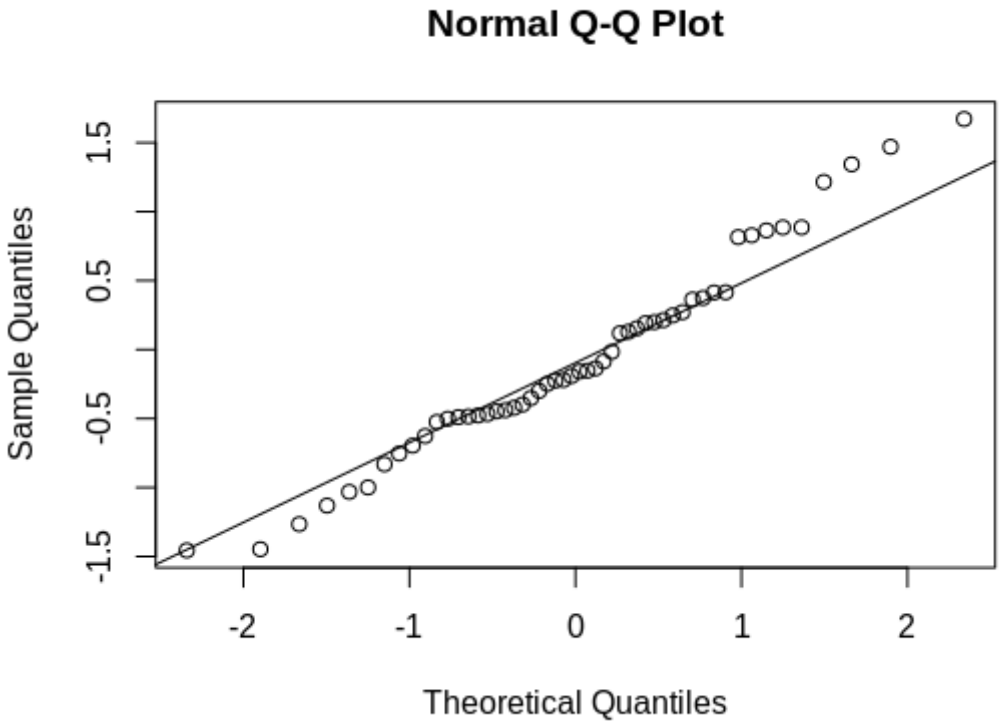


Figura 3 – Gráfico da Normalidade (Sem Variáveis Transformadas)

Pode-se observar que a equação $Y_{\text{preço}_i} = -41771,930 + 17,276 \times X_{\text{área}_i} + 8195,227 \times X_{\text{coefap}_i} + 55767.600 \times X_{\text{acldecl}_i} + 221,517 \times X_{\text{frente}_i}$ segue uma distribuição normal, aproximadamente falando.

```
qqnorm(residuos_sem_outliers)
qqline(residuos_sem_outliers)
shapiro.test(residuos_sem_outliers)
```

```
Shapiro-Wilk normality test
data:  residuos_sem_outliers
W = 0.97457, p-value = 0.3269
```

Pelo teste acima, vemos que p-valor é 0.327, sendo maior que 0.05. A hipótese nula desse teste nos diz que a população é normalmente distribuída. Então, de um lado, temos que se o p-valor é menor que o nível de significância, então a hipótese nula é rejeitada e existe evidência de que as amostras testadas **não** são normalmente distribuídas. Por outro lado, se o p-valor é maior que o nível de significância escolhido, então a hipótese nula diz que a população veio de dados normalmente distribuídos e **não** pode ser rejeitada.

2.0.9 Faça os devidos ajustes do modelo, se necessário, conforme sugerido no item 6.

Nós optamos por não retirar nenhum dado da amostra, além dos que já foram retirados pela autora oficial das observações de campo. Ela começou com 60 observações e caiu para 54 amostras.

Então o modelo segue da mesma forma que calculamos antes:

$Y_{\text{preço}_i} = -41771,930 + 17,276 \times X_{\text{área}_i} + 8195,227 \times X_{\text{coefap}_i} + 55767.600 \times X_{\text{acldecl}_i} + 221,517 \times X_{\text{frente}_i}$

3 Conclusões

Pela observação dos aspectos analisados, ficou clara a importância da Análise de regressão múltipla para a equipe. Neste trabalho observamos que podemos construir modelos que descrevem de maneira razoável relações entre várias variáveis explicativas de um determinado processo.

Segundo a autora, O Geoprocessamento pode,através de outros tipos de análises, gerar mapas de concentração ao redor de um ponto valorizante, ou relacionar a uma área de influência (questões de entorno), possibilitar a geração de mapas derivados (curvas isotimas) ou ainda simular o comportamento da superfície a partir de mudanças nos registros das variáveis que definem o valor (cronologicamente) [1].

Realizamos testes de hipóteses para conseguir verificar mais a fundo a influência da correlação linear. Em nossa análise, as variáveis independentes que mais influenciaram no valor do imóvel (em reais) foram "sentido predominante da topografia por lote"(escala deduzida pelas classes do histograma) e "coeficiente de aproveitamento"(escala deduzida pelas classes do histograma), "área"(metros quadrados) e "frente"(metros).

Conseguimos concluir que o modelo é útil visto que foi rejeitada a hipótese nula, considerando assim que as variáveis independentes possuíam relação com a variável dependente. Também, nosso modelo de resíduos apresentou um resultado satisfatório, mostrando uma aleatoriedade simétrica entre positivos e negativos.

Portanto, conclui-se que conseguimos prever o valor de um lote utilizando como variáveis a sua área, "sentido predominante da topografia por lote", frente do terreno e aproveitamento da área construída do lote.

3.0.1 Resumidamente

1. Modelo criado, sem transformações, é útil.

2. Dentro do escopo, é possível prever valor de imóveis, razoavelmente bem.

Para finalizar veja alguns valores preditos *vs.* valores observados dos preços dos imóveis:

i	Y_{obsi}	$\hat{Y}_i$
0	25000	22772.58
1	45000	47379.96
2	45000	35513.36
3	20000	22576.89
4	25000	21115.27
5	20000	21996.21
6	30000	35249.05
7	25000	25905.52
8	40000	35574.58

Tabela 5 – Alguns valores Preditos contra os valores de Y da amostra

3.0.2 Trabalhos Futuros

Mesclar esse trabalho com variáveis de valor do imóvel no tempo, inflação, juros imobiliários, juros de mercado em geral.

Referências Bibliográficas

1 ROCHA, R. R. Técnicas de geoprocessamento aplicadas à avaliação de imóveis. estudo de caso: Região central de ibirité. 2005. 44 f. monografia (especialização). *Curso de Geoprocessamento, Cartografia, Instituto de Geociências, Universidade Federal de Minas Gerais, Belo Horizonte, 2005. Disponível em: <http://csr.ufmg.br/geoprocessamento/publicacoes/RaquelResendeRocha.pdf>. Acesso em: 22 out. 2019.*, v. 0, p. 44, 2016. 9, 12, 13, 14, 15, 26