UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CURSO DE CIÊNCIAS DA COMPUTAÇÃO
DISCIPLINA: Probabilidade e Estatística - INE5405
PROFESSOR: Pedro Alberto Barbetta

REGRESSÃO LINEAR SIMPLES – CADIDATOS POR VAGA E NOTA DO ÚLTIMO COLOCADO NO VESTIBULAR DA UFSC 2010

Lucas Pereira, Paulo Jair, William Rodrigues

Florianópolis
Dezembro de 2010

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 DESCRIÇÃO DO PROBLEMA

Nosso trabalho consiste em analisar um problema de regressão linear simples, descrevendo suas variáveis independente e dependente. Precisávamos escolher essas variáveis e decidimos utilizar o vestibular da UFSC como fonte de dados. Optamos por utilizar o vestibular da UFSC pois já fizemos o primeiro trabalho desta matéria utilizando-o e dessa forma teríamos um prévio conhecimento das abordagens a serem feitas, facilitando o desenvolvimento deste trabalho.

Tendo o tema do nosso estudo definido precisávamos escolher quais variáveis utilizar e foi decidido fazer a regressão linear simples utilizando o número de candidatos por vaga de cada curso do vestibular da UFSC como variável independente e a nota do último classificado no curso como variável dependente.

Intuitivamente acredita-se que em geral se o número de candidatos por vaga é grande, então a nota do último classificado no curso também será. Buscaremos mostrar se essa noção intuitiva está certa e se podemos ou não, dizer que o número de candidatos por vaga realmente influencia em algum nível a nota do último candidato.

2 APRESENTAÇÃO DO ARQUIVO DE DADOS

Foram coletadas **25** observações referentes a alguns dos cursos oferecidos no **vestibular da UFSC 2010**. Para a manipulação dos dados utilizamos o software **Open Office - Planilha Eletrônica**. Abaixo se encontra a matriz com as **25** observações. Na Tabela abaixo **CV** representa o número de candidatos por vaga e **NUC** é a nota do último classificado no respectivo curso.

Adotaremos ao longo deste trabalho a seguinte convenção:

X = Número de candidatos por vaga;

Y = Nota do último classificado no curso.

Curso	CV	NUC
Letras - Língua Alemã	0,88	40,95
Ciências Biológicas – Licenciatura Noturno	1,24	36,64
Engenharia De Aqüicultura	1,50	38,03
Fonoaudiologia - Noturno	1,54	39,9
Serviço Social - Diurno	1,61	36,99
Ciências Sociais - Diurno	1,90	39,5
Letras - Língua Francesa	1,98	40,19
Letras - Língua Portuguesa – Noturno	2,18	38,42
Filosofia - Noturno	2,51	39,74
Engenharia De Produção Elétrica	2,80	61,74
Serviço Social - Noturno	2,98	39,91
Ciências Econômicas - Noturno	3,30	51
História - Noturno	3,64	49,13
Farmácia	4,32	55,84
Engenharia De Alimentos	4,72	61,9
Enfermagem	5,16	48,25
Administração - Diurno	5,45	57,36
Educação Física - Bacharelado	5,87	48,85
Engenharia Sanitária E Ambiental	6,22	65,79
Ciências Biológicas	7,33	61,75
Psicologia	8,86	59,33
Odontologia	9,55	63,6
Oceanografia	10,97	63,38
Direito - Noturno	12,29	70,45
Arquitetura e Urbanismo	14,78	68,15

Tabela 2: Arquivo de dados do vestibular da UFSC 2010

3 COEFICIENTES E EQUAÇÃO DE REGRESSÃO

Para determinar os coeficientes e a equação de regressão é necessário estimar o nosso conjunto de observações. Para isso iremos utilizar o método aprendido em aula: método de mínimos quadrados. Sendo assim a equação de regressão é dada por:

$$\hat{y} = a + bx$$

Além disso temos que:

$$b = \frac{n\sum(x_i y_i) - (\sum x_i)(\sum x_i)}{n\sum x_i^2 - (\sum x_i)^2} \qquad a = \frac{\sum y_i - b\sum x_i}{n}$$

Realizando as contas através do **Open Office - Planilha Eletrônica** chegamos aos seguintes resultados:

$$b = \frac{(25)(7181,56) - (123,58)(1276,79)}{(25)(952,65) - (15272,02)} = \frac{179539,01 - 157785,71}{23816,18 - 157785,71} = \frac{21753,30}{8544,16} = 2,55$$

$$a = \frac{1276,79 - 314,63}{25} = \frac{962,16}{25} = 38,49$$

Finalmente temos a equação de regressão:

$$\hat{y} = 38,49 + (2,55)x$$

4 ANÁLISE GRÁFICA

Após ter encontrado a equação de regressão passamos para a parte de plotagem do gráfico de dispersão juntamente com a reta de regressão que se encontra abaixo.

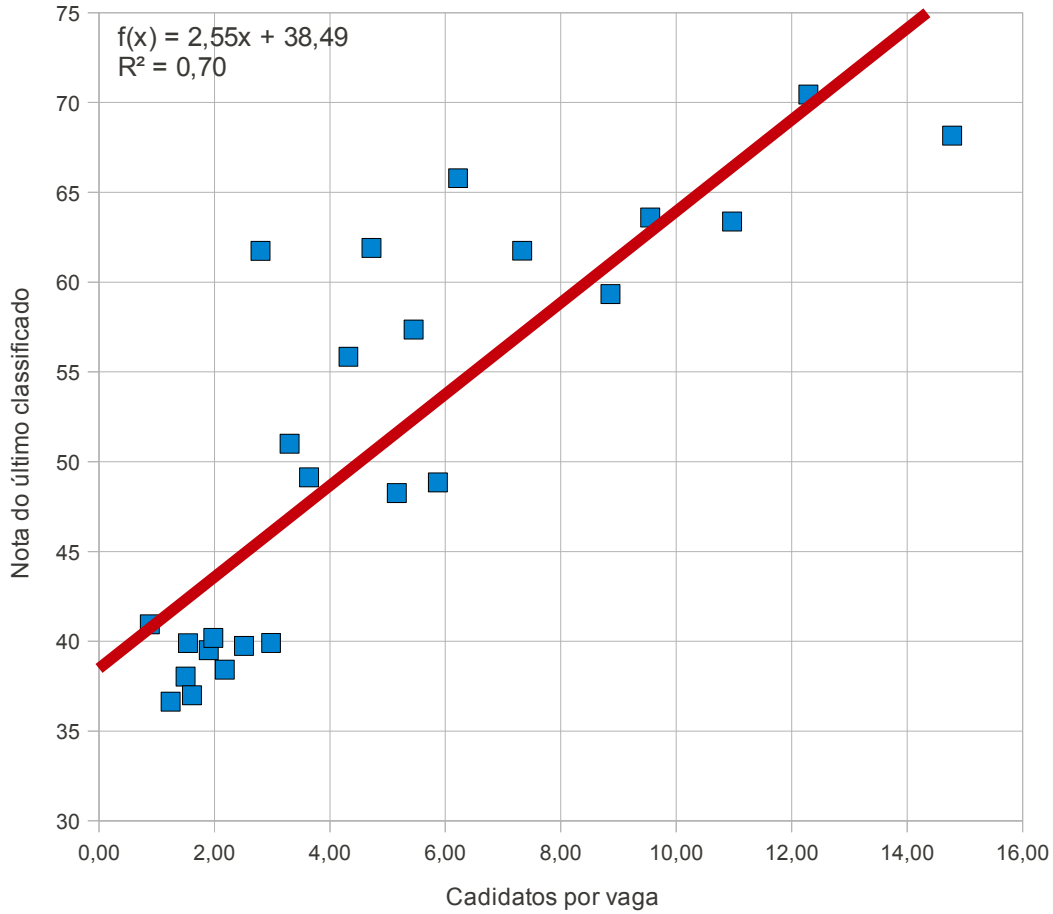

Figura 1: Gráfico de dispersão com reta de regressão.

Vamos agora analisar o coeficiente angular (**b**) da equação da reta de regressão. Para isso iremos fazer um teste de hipóteses e a partir desse teste observar se existe uma associação linear entre o número de candidatos por vaga e a nota do último classificado. As hipóteses do teste são:

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

As hipóteses são tomadas dessa forma pois como na equação de

regressão, **β** está multiplicando **X** e sendo **β** zero, então a equação não é afetada por **X**. Isso mostraria não existir nenhuma associação linear entre as variáveis. Dessa forma se o teste rejeitar $\mathbf{H_0}$ mostrará a existência de uma relação linear.

Adotaremos como nível de significância 5%. Como o teste é bilateral temos uma área de 2,5% na cauda superior onde $\mathbf{t_\alpha}$ é igual a 2,069 (utilizando gl igual a 23). Calculemos agora **t**:

$$t = \frac{b - \beta_0}{S_b}$$

$$onde\ S_b\ foi\ préviamente\ calculado\ com\ valor\ de\ 0{,}3455 \qquad t = \frac{2{,}55 - 0}{0{,}3455} = 7{,}37$$

Observamos que ($\mathbf{t_\alpha}$ = 2,069) < (**t** = 7,37) e dessa forma o teste rejeita $\mathbf{H_0}$ e mostra que existe sim uma relação linear entre o número de candidatos por vaga e nota do último classificado. É interessante analisar que o teste aceita como hipótese verdadeira o $\mathbf{H_1}$ com bastante folga, ou seja poderíamos ter escolhido um nível de significância bem menor que ainda assim o teste rejeitaria $\mathbf{H_0}$.

5 CÁLCULO DE R²

R² é chamado de coeficiente de determinação. Ele representa a proporção da variação de **Y** que pode ser explicada por variações em **X**. Anteriormente fizemos o teste de hipóteses para verificar se **X** possui uma relação linear com **Y**. Agora com **R²** iremos estimar o quão relacionadas estão estas variáveis. A formula para o cálculo de **R²** se encontra abaixo:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$$

A formula acima é a razão entre a variância explicada pela equação de regressão, que são os desvios preditos em relação à média aritmética, e a variância total que é a soma da variância explicada com a variância não explicada. Porém o comentário importante a se fazer neste ponto é que a variância não explicada representa fatores aleatórios não controláveis que podem influenciar em **Y**. Esses fatores são chamados de resíduos, sendo considerados como uma estimativa do erro aleatório. Esses resíduos são a diferença entre o valor observado e o valor predito .

Já calculamos na planilha eletrônica a variância total (3153,79) e a variância explicada (2215,34). Agora vamos ao calculo de **R²**:

$$R^2 = \frac{2215,34}{3153,79} = 0,70$$

Isso mostra que existe uma relação linear razoavelmente boa entre o número de candidatos por vaga e a nota do último classificado. Podemos dizer que cerca de 70% da variação da nota do último classificado pode ser explicada pelo número de candidatos por vaga, enquanto que os outros 30% significam a variação provocada por outros fatores não considerados nesse modelo de regressão simples. Esses outros fatores podem ser ter diversas causas e serem tanto controláveis quanto aleatórios.

6 RESÍDUOS E VALORES PREDITOS

Os resíduos e valores preditos já foram usados no presente trabalho para vários cálculos, como por exemplo no cálculo de R^2 e do teste para β. Agora iremos apresentar a tabela de dados contendo estes valores:

Curso	CV	NUC	Predito NUC	Resíduo NUC
Letras - Língua Alemã	0,88	40,95	40,73	0,2232
Ciências Biológicas – Licenciatura Noturno	1,24	36,64	41,64	-5,0033
Engenharia De Aqüicultura	1,50	38,03	42,31	-4,2753
Fonoaudiologia - Noturno	1,54	39,9	42,41	-2,5071
Serviço Social - Diurno	1,61	36,99	42,59	-5,5953
Ciências Sociais - Diurno	1,90	39,5	43,32	-3,8237
Letras - Língua Francesa	1,98	40,19	43,53	-3,3373
Letras - Língua Portuguesa – Noturno	2,18	38,42	44,04	-5,6165
Filosofia - Noturno	2,51	39,74	44,88	-5,1367
Engenharia De Produção Elétrica	2,80	61,74	45,62	16,1250
Serviço Social - Noturno	2,98	39,91	46,07	-6,1633
Ciências Econômicas - Noturno	3,30	51	46,89	4,1120
História - Noturno	3,64	49,13	47,75	1,3763
Farmácia	4,32	55,84	49,48	6,3551
Engenharia De Alimentos	4,72	61,9	50,50	11,3967
Enfermagem	5,16	48,25	51,62	-3,3736
Administração - Diurno	5,45	57,36	52,36	4,9981
Educação Física - Bacharelado	5,87	48,85	53,43	-4,5812
Engenharia Sanitária E Ambiental	6,22	65,79	54,32	11,4677
Ciências Biológicas	7,33	61,75	57,15	4,6016
Psicologia	8,86	59,33	61,04	-1,7137
Odontologia	9,55	63,6	62,80	0,7996
Oceanografia	10,97	63,38	66,42	-3,0357
Direito - Noturno	12,29	70,45	69,78	0,6736
Arquitetura e Urbanismo	14,78	68,15	76,12	-7,9659

Tabela 2: Arquivo de dados do vestibular da UFSC 2010 contendo valores preditos e resíduos.

Na tabela acima **Predito NUC** representa o valor de $\hat{y}$ que é o resultado da equação de regressão onde **x** vale **CV**. Já **Resíduo NUC** representa a diferença entre o valor observado (**NUC**) e o valor predito (**Predito NUC**).

6.1 Gráfico De Resíduos

O gráfico de resíduos consiste em um gráfico de dispersão com os pares (xi, ei) onde nesse caso $\mathbf{x_i}$ é o **CV** do i-ésimo termo e $\mathbf{e_i}$ é o resíduo do i-ésimo termo.

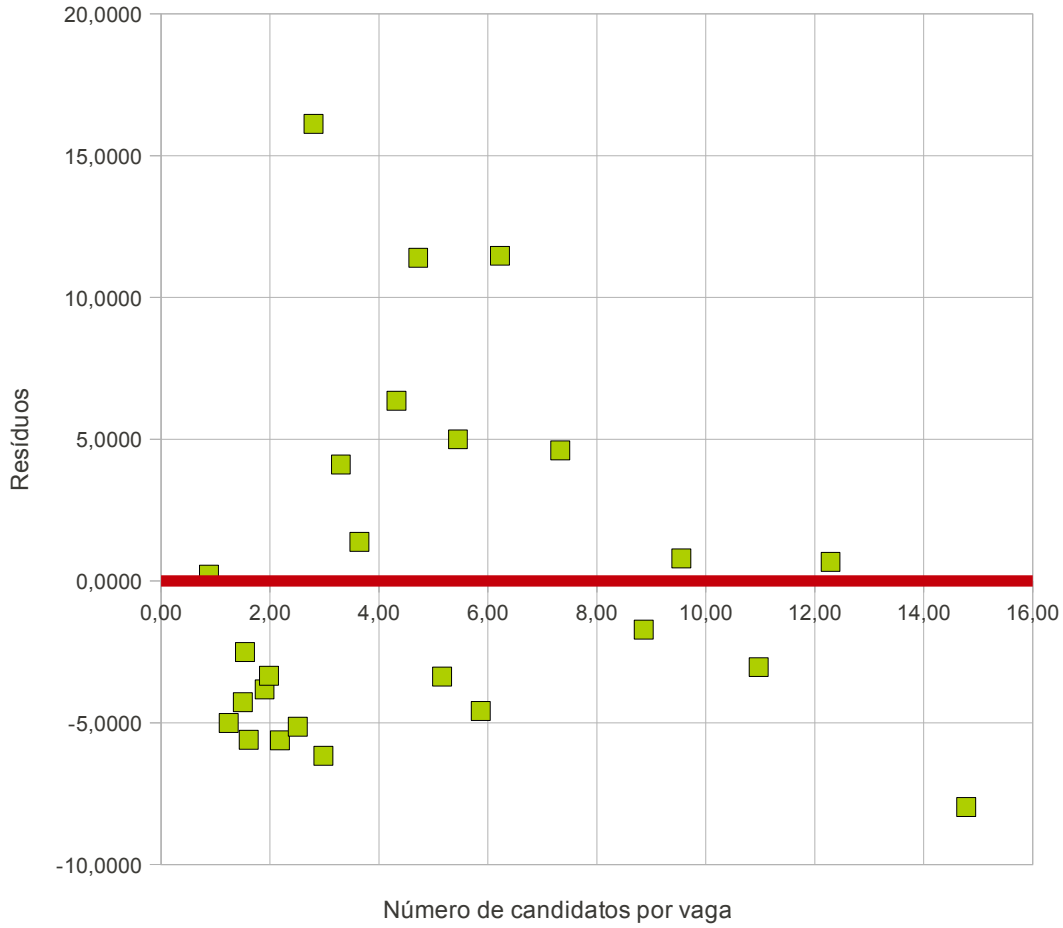

Figura 2: Gráfico de resíduos.

Analisando o gráfico acima percebemos que a adequação do modelo está razoavelmente satisfeita uma vez que os resíduos se apresentam com relativa distribuição aleatória em torno da reta de regressão.

É interessante notar alguns pontos discrepantes que diminuem a adequação do modelo. Além disso esta parece ser uma situação onde os valores grandes de **X** acabam por determinar uma leve inclinação na reta, onde nota-se que a variância de **Y** aumenta proporcionalmente com **X**. Sendo assim a transformação logarítmica tanto nos valores de **X** quanto nos calores de **Y** poderia ser uma boa

opção para melhorar levemente a adequação do modelo.

6.2 Transformação Logarítmica do Modelo

Como falado na seção anterior a aplicação da transformação logarítmica poderia deixar o modelo mais adequado. Desta forma fizemos os gráficos considerando a transformação logarítmica tanto nos valores de X quanto nos valores de Y. O resultado é encontrado nos gráficos abaixo:

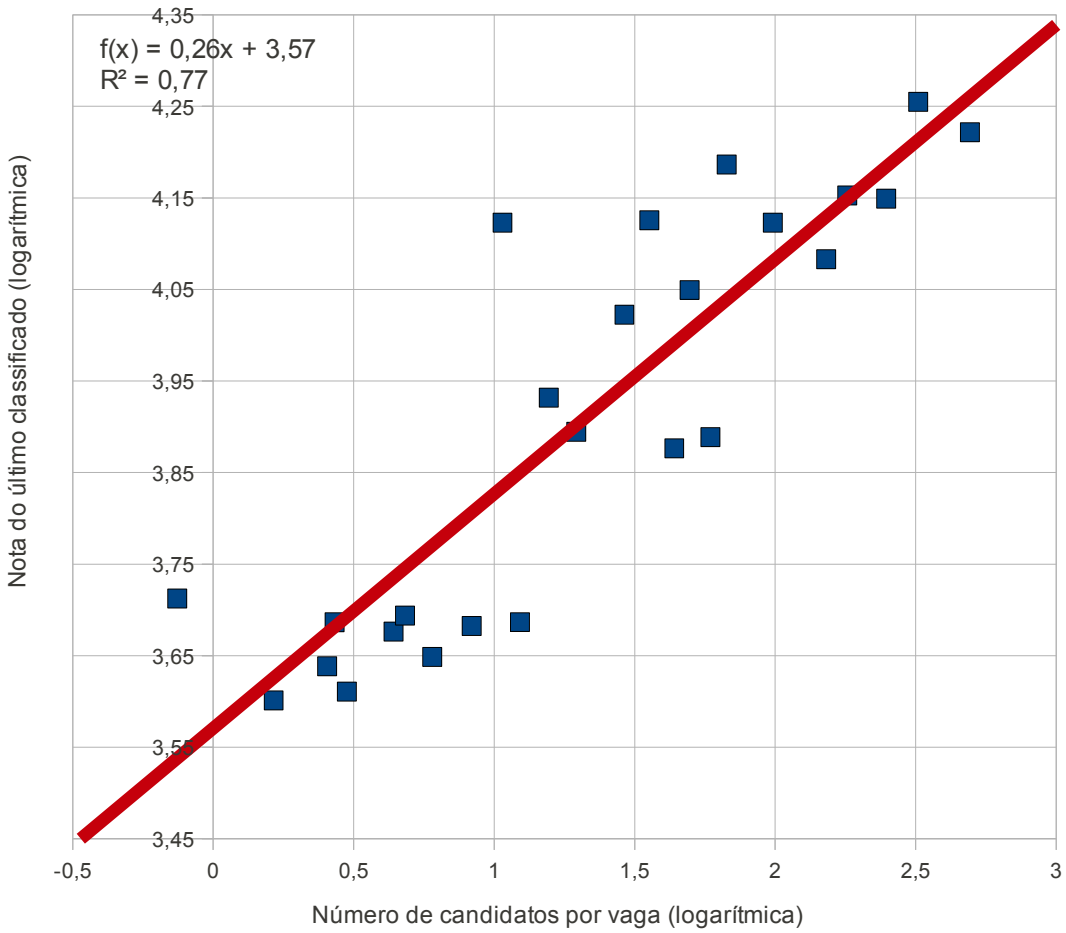

Figura 3: Gráfico de dispersão com reta de regressão com transformação logarítmica.

Como podemos ver ao comprar esse gráfico onde a transformação

logarítmica foi aplicada com o primeiro gráfico de regressão percebemos notoriamente a diferença. Neste modelo os pontos estão nitidamente mais alinhados à reta de regressão e além disso, obtivemos um calculo de **R²** igual a 0,77, mostrado dessa forma a melhoria comparada com o primeiro gráfico onde **R²** valia 0,70. A equação de regressão ficou da seguinte forma:

$$\hat{y} = 3{,}57 + (0{,}26)x$$

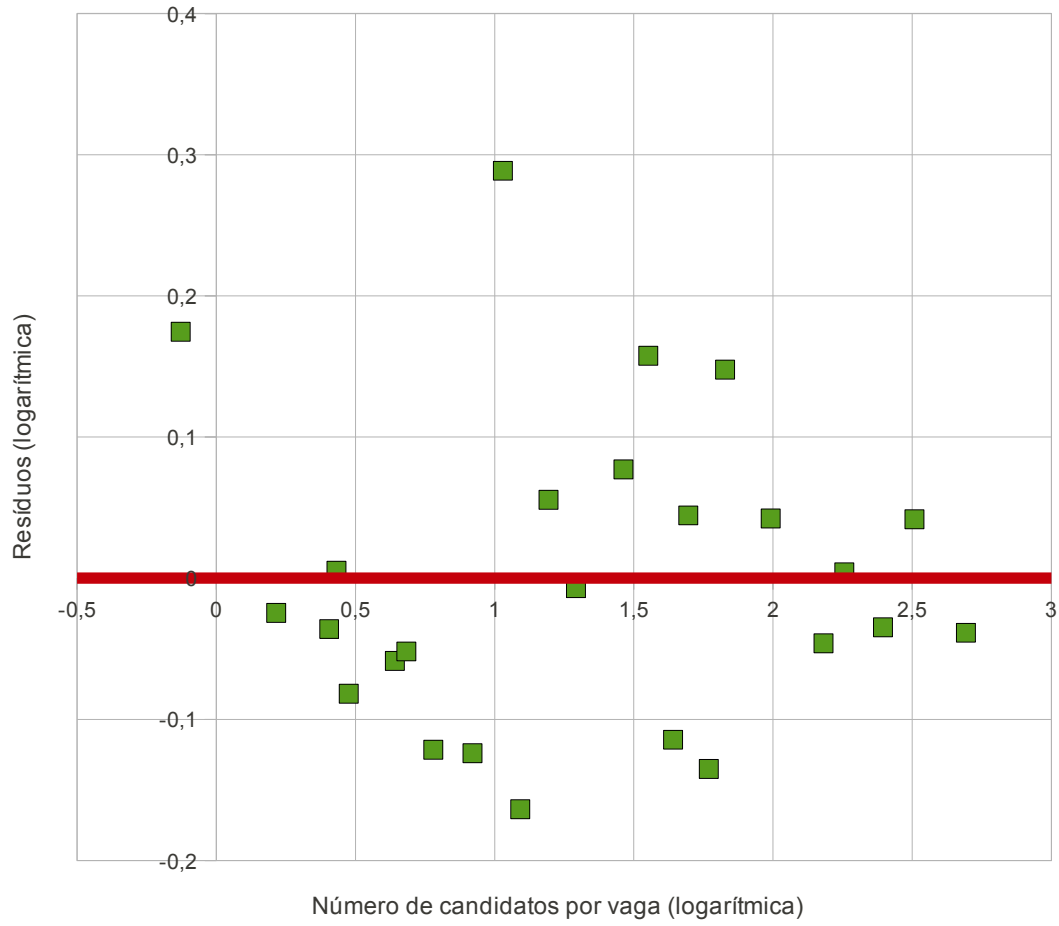

Figura 4: Gráfico de resíduos com transformação logarítmica.

O gráfico de resíduos com a transformação logarítmica apenas evidencia o fato da melhora na adequação do modelo onde os resíduos se distribuem de forma mais aleatória comparado com o primeiro gráfico de resíduos.

REFERÊNCIAS

UNIVERSIDADE FEDERAL DE SANTA CATARINA. **Relatório Oficial do Vestibular da UFSC 2010**. Disponível em: <http://www.vestibular2010.ufsc.br/relatorio/>. Acesso em: 26 de agosto de 2010.

WIKIPÉDIA. **Anexo: Lista de Cursos Superiores**. Disponível em: <http://pt.wikipedia.org/wiki/Anexo:Lista_de_cursos_superiores>. Acesso em: 28 de agosto de 2010.