ISSN 2448-3230

O QUE INFLUI NOS ACIDENTES DE TRÂNSITO? UMA ANÁLISE DE REGRESSÃO

Jeniffer Johana Duarte Sanchez

Raquel Costa da Silva

Resumo

Atualmente são 3 mil vidas perdidas por dia nas estradas em todo o mundo, esses acidentes já representam um custo de US518 bilhões por ano, ou um percentual entre 1% e 3% do produto interno bruto de cada país. “Traumas e lesões decorrentes da violência no trânsito são umas das principais causas de deficiência – e de mortes – no Brasil” informa o IBDD. Neste trabalho utilizamos o modelo de regressão linear para tentar explicar a taxa de acidentes de trânsito no estado de Minnesota, EUA no ano de 1973. Esse estudo teve como base onze variáveis explicativas das quais ao final apenas quatro delas mostraram ser significativas para o modelo de regressão.

Palavras-chave: Acidentes de trânsito; Análise de regressão linear; Mínimos quadrados; HC4

Introdução

Em linhas gerais podemos dizer que a análise de regressão se preocupa largamente em prever e/ou estimar a média da população ou o valor médio das variáveis dependentes, a partir dos valores conhecidos ou fixados de uma ou mais variáveis explicativas.

Segundo a OMS, em março de 2010, uma resolução definindo o período de 2011 a 2020 como a “Década de ações para a segurança no trânsito”. O documento foi elaborado com base em estudo da Organização Mundial da Saúde (OMS) que contabilizou, em 2009, cerca de 1,3 milhão de mortes por acidente de trânsito em 178 países. Aproximadamente 50

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

milhões de pessoas sobreviveram com sequelas. Se nada for feito, a OMS estima que 1,9 milhão de pessoas devem morrer no trânsito em 2020 (passando para a quinta maior causa) e 2,4 milhões, em 2030. Nesse período, entre 20 milhões e 50 milhões de pessoas sobreviverão aos acidentes a cada ano com traumatismos e ferimentos. Daí a importância desse tipo de modelagem.

Neste trabalho utilizamos a análise de regressão para estudar as taxas de acidentes de trânsito em estradas Norte-Americanas especificamente do estado de Minnesota (1973). Os dados apresentavam 39 observações; a variável resposta é a taxa de acidentes por milhões de quilômetros de veículos, e dentre as variáveis explicativas merecem destaque: o comprimento do segmento em milhas; imite de velocidade e o tipo de estrada. Quando observamos tais variáveis podemos perceber a influência delas nos acidentes ainda nos dias atuais.

Nossos resultados mostraram que os dados apresentam problemas de multicolinearidade, para poder escolher um modelo de regressão adequado usamos os estimadores HC4, no modelo considerado para análise só quatro variáveis foram significativas para a explicação da taxa de acidentes de trânsito, nesse modelo se verificaram todas as suposições e realizou-se a análise de pontos de alavanca, no final o modelo considerado consegue explicar o 78.49% da variabilidade total da variável resposta.

Metodologia

A análise dos dados da taxa de acidentes de trânsito no estado de Minnesota, EUA no ano de 1973 será feita usando a análise de regressão, inicialmente consideraremos todas as variáveis incluídas na base de dados, no total onze variáveis explicativas. Na análise de regressão serão verificados todos os supostos para poder obter conclusões válidas, as quais são: ausência de multicolinearidade, homocedasticidade e normalidade.

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

A estimação da reta de regressão será feita considerando o método de mínimos quadrados ponderados. Como mencionado anteriormente incialmente será considerado o modelo com todas as variáveis explicativas, também serão considerados modelos com transformações das variáveis e subconjuntos do modelo inicial, para a escolha do modelo final será considerado o coeficiente de determinação R^2 e o coeficiente de determinação ajustável $\bar{R}^2$. Como observa Theil: "É aconselhável utilizar $\bar{R}^2$ em vez de R^2, pois R^2 tende a fornecer um quadro demasiadamente otimista do ajuste da regressão, particularmente quando o número de variáveis explicativas não for muito pequeno comparado com o número de observações". Além será considerado o estimador $HC4$ (Cribari-Neto, 2004) para fazer testes quase-t, que permitem escolher o modelo sem ter a necessidade de testar em cada um deles as suposições de normalidade e homocedasticidade.

No modelo final escolhido serão verificados os supostos, para testar a normalidade e homocedasticidade neste trabalho foram realizados os seguintes testes: Teste de Breusch-Pagan-Godfrey, teste de Koenker e o teste de Jarque-Bera. O tamanho da amostra é 39 pelo qual não serão usados os resultados assintóticos para as distribuições dos testes, no seu lugar será usado o bootstrap não-paramétrico para obter a distribuição das estatísticas de teste, considerou-se 1000 réplicas de bootstrap. No modelo final também será feita a análise de observações influentes, para tal se considera a medida de atipicidade ou alavancagem do padrão de regressores para a observação t. As estatísticas aqui utilizadas para identificar os pontos influentes entre as variáveis independentes foram: resíduos estudentizados, distância de Cook e outras medidas.

Resultados e discussões

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

Os dados tentam explicar a relação entre taxas de acidentes nas estradas norte-americanas (estado de Minnesota, 1973) e várias características da rodovia. Cada caso é um trecho específico da rodovia. As variáveis consideradas são:

- rate: taxa de acidentes por milhões de quilômetros de veículos em 1973 (a resposta);
- len: comprimento do segmento em milhas;
- ADT: contagem média de tráfego diário em mil;
- trks: volume de caminhões como porcentagem do volume total do tráfego;
- sigs1: número de sinais por quilômetro de estrada (ajustado para não conter o zero), em que esse ajuste se dá pela expressão $\frac{sigs \times len + 1}{len}$;
- slim: limite de velocidade em 1973;
- shld: largura (em pés) do "ombro" exterior da estrada;
- lane: número total de tráfego na pista;
- acpt: número de pontos de acesso por milha;
- itg: número de intercâmbios de tipo de auto-estrada por milha;
- lwid: largura da pista em pés;
- hwy: indicador do tipo de estrada ou tipo de financiamento da estrada (MC, FAI, PA ou MA).

O interesse reside na modelagem da taxa de acidentes (rate) usando o modelo linear de regressão, a variável hwy foi incluída no modelo de regressão tendo em conta três variáveis dummy, pois vamos considerar um modelo com intercepto, a amostra tem tamanho 39.

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

Na Tabela 1 apresentamos algumas estatísticas de resumo das variáveis consideradas no modelo final. Podemos ver que a taxa de acidentes de trânsito é relativamente alta, alcançando valores de 9.23, a variável slim mostra aparentemente um comportamento simétrico.

Tabela 1: Algumas estatísticas de resumo para as variáveis, Rate, len, Sigs1 e slim.

	Rate	len	Sigs1	Slim
Mínimo	1,610	2,960	0,045	40
1ro Quartil	2,630	7,995	0,087	50
Mediana	3,050	11,390	0,176	55
Media	3,933	12,884	0,510	55
3ro Quartil	4,595	17,800	0,715	60
Máximo	9,230	40,090	2,789	70

Foram feitos histogramas e diagramas de caixa, considerou-se a variável em sua escala original e o logaritmo dela (nas Figuras 1, 2, 3 e 4 do lado esquerdo se tem a variável em sua escala original e do lado direito a variável em escala logarítmica). Serão apresentados os gráficos para as variáveis rate, len Sigs1 e slim como exemplo. Em alguns casos o logaritmo da variável apresentou um melhor comportamento, tirando a assimetria dos dados e alguns dados atípicos, como no caso das variáveis rate, len, ADT, sigs1, acpt e itg. Quanto a variável trks, ela apresenta maior simetria quando trabalhamos com o logaritmo, já as variáveis slim e shld têm melhor comportamento na sua escala original e finalmente para as variáveis lane e lwid não faz diferença trabalhar com ou sem o logaritmo delas; por este motivo vamos levar em conta esta transformação no momento de considerar os diferentes modelos

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

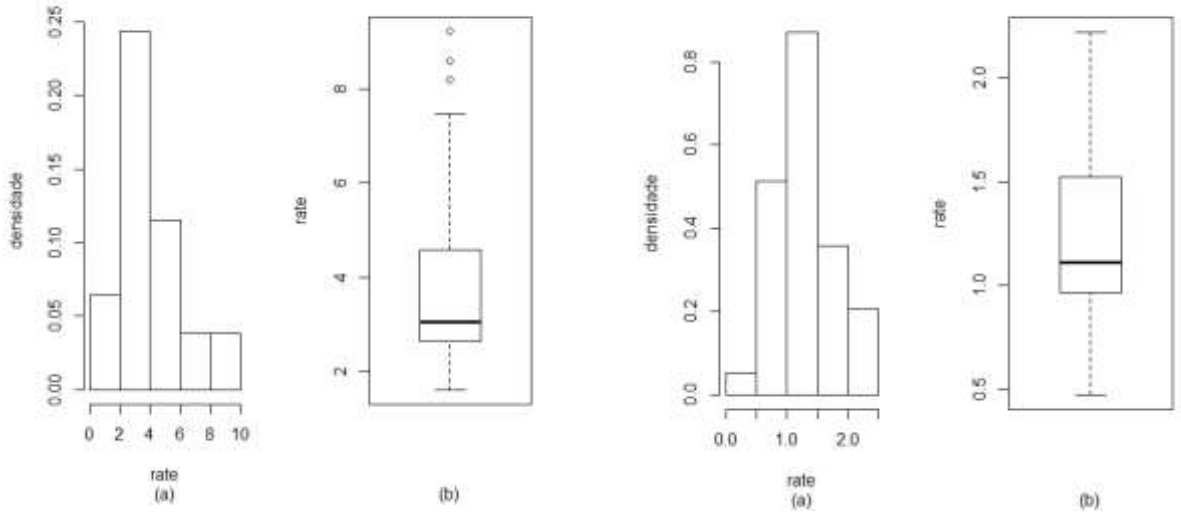
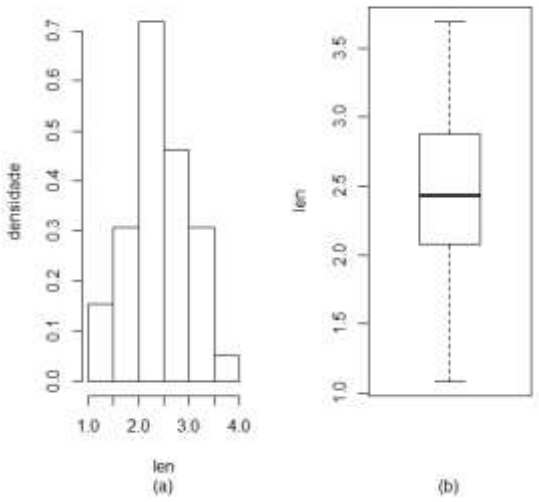


Figura 1: Histograma e box-plot da variável rate, lado esquerdo escala original, lado direito escala logarítmica.



USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

Figura 2: Histograma e box-plot da variável len, lado esquerdo escala original, lado direito escala logarítmica.

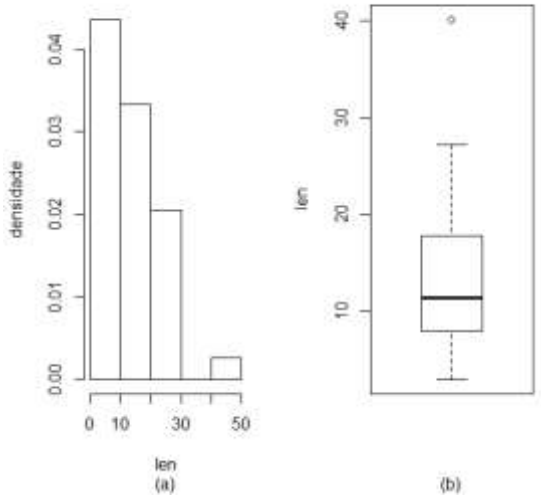


Figura 3: Histograma e box-plot da variável sigs1, lado esquerdo escala original, lado direito escala logarítmica.

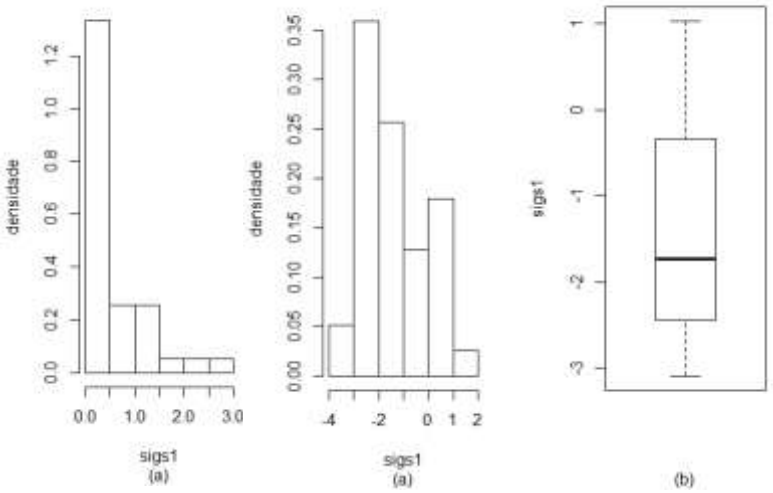
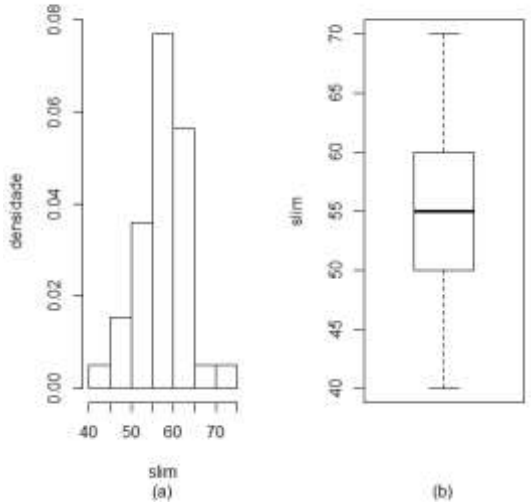
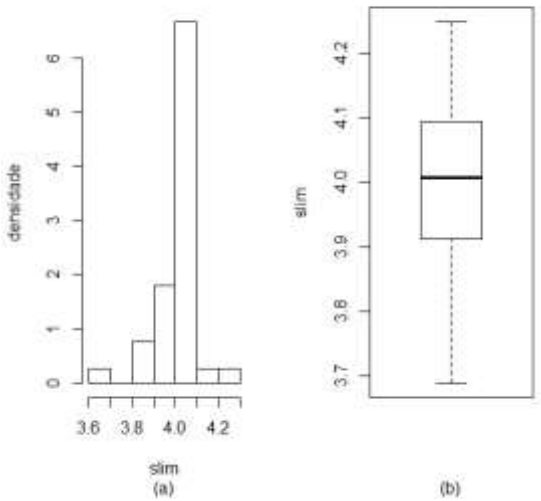


Figura 4: Histograma e box-plot da variável slim, lado esquerdo escala original, lado direito escala logarítmica.



Por último vamos considerar a matriz de correlação das variáveis regressoras excluindo a variável categórica hwy (apresentada a seguir), a qual nos



USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

permite concluir que as variáveis ADT, lane, itg e lwid apresentam baixa correlação com a variável resposta (rate), sendo a correlação de -0.03, -0.03, -0.02 e -0.01 respectivamente, o qual permite pensar que estas variáveis não ajudam explicar a variável resposta; por outro lado temos que a correlação entre as variáveis lane e ADT é de 0.82 e entre as variáveis itg e ADT é de 0.90, as quais são maiores do que 0.8, isso permite concluir que existem problemas de multicolin earidade entre as variáveis.

	rate	len	ADT	trks	sigs1	slim	shld	lane	acpt	itg	lwid
rate	1.00	−0.47	−0.03	−0.51	0.60	−0.68	−0.39	−0.03	0.75	−0.02	−0.01
len		1.00	−0.27	0.50	−0.39	0.19	−0.10	−0.20	−0.24	−0.25	−0.31
ADT			1.00	−0.10	0.16	0.24	0.46	0.82	−0.22	0.90	0.13
trks				1.00	−0.48	0.30	0.01	−0.15	−0.36	−0.07	−0.16
sigs1					1.00	−0.42	−0.12	0.26	0.51	0.09	0.06
slim						1.00	0.69	0.26	−0.68	0.24	0.10
shld							1.00	0.48	−0.42	0.38	−0.04
lane								1.00	−0.21	0.70	0.10
acpt									1.00	−0.20	−0.04
itg										1.00	0.10
lwid											1.00

Modelo de regressão: O modelo inicial considera todas as variáveis, devemos lembrar que a variável hwy foi dividida em três variáveis dummy, tomando como categoria de base FAI, temos então que as variáveis dummy ficam definidas como segue:

$$PA = \begin{cases} 1 \text{ se } hwy = PA, \\ 0 \text{ caso contrário,} \end{cases} \quad MA = \begin{cases} 1 \text{ se } hwy = MA, \\ 0 \text{ caso contrário,} \end{cases} \quad MC = \begin{cases} 1 \text{ se } hwy = MC, \\ 0 \text{ caso contrário,} \end{cases}$$

Assim definimos o modelo inicial como segue

$$E(rate) = \beta_1 + \beta_2 len + \beta_3 ADT + \beta_4 trks + \beta_5 sigs1 + \beta_6 slim + \beta_7 shld + \beta_8 lane + \beta_9 acpt + \beta_{10} itg + \beta_{11} lwid + \beta_{12} PA + \beta_{13} MA + \beta_{14} MC.$$

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

Ao realizar o ajuste do modelo temos que a um nível de significância de 5% nenhuma das variáveis explicativas é significativa no modelo de regressão, mas o R^2 é alto (0.7662) e o teste F indica a significância da regressão, isso é um indicativo de multicolinearidade, na análise descritiva vimos que a correlação entre alguns regressores foi maior a 0.8 reafirmando a hipótese de multicolinearidade.

Foram analisadas outras medidas para detectar problemas de multicolinearidade (fatores de inflação da variância, o número de condição, regressar cada regressor sobre os outros) para poder concluir com maior precisão a existência deste problema no conjunto de dados. Com base nas análises de multicolinearidade, as variáveis ADT, lane e itg podem ser consideradas as causadoras deste problema. As suposições de normalidade e homoscedasticidade do modelo de regressão serão testadas usando o teste de normalidade Jarque-Bera e os testes de homoscedasticidade Breusch-Pagan-Godfrey e Koenker, estes últimos baseiam-se em resultados assintóticos e o tamanho amostral é 39, assim vamos fazer uso do bootstrap não-paramétrico para poder obter a distribuição das estatísticas de teste, considerou-se 1000 réplicas de bootstrap, os valores obtidos para as estatísticas de teste são apresentados na Tabela 2, como os valores das estatísticas de teste são menores do que os valores das estatísticas bootstrap, concluímos que o modelo cumpre as suposições de normalidade e homoscedasticidade.

Tabela 2: Estatísticas de teste para normalidade e homocedasticidade no modelo inicial.

	Estatística de teste	Estatística Bootstrap
Jarque-Bera	0,4040	27,8259
Breusch-Pagan-Godfrey	23,2378	45,8001
Koenker	20,8674	28,9067

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

Para poder obter conclusões válidas do modelo de regressão precisamos ajeitar o problema de multicolinearidade existente, para o qual vamos considerar algumas possíveis combinações das variáveis ADT, lane e itg. No total consideramos 126 modelos resultantes das possíveis combinações. Para testar a significância dos parâmetros de cada um dos modelos sem ter que testar em cada um deles as suposições de normalidade e homoscedasticidade vamos considerar o estimador HC4 que apresenta um bom comportamento em testes de hipóteses.

Depois de considerar vários modelos, o modelo final é apresentado a seguir:

$$E(\log(rate)) = \beta_1 + \beta_2 \log(len) + \beta_3 \log(sigs1) + \beta_4 slim + \beta_5 PA.$$

Vamos checar no modelo final as suposições de normalidade e homoscedasticidade, considerou-se 1000 réplicas de bootstrap, na Tabela 3 são apresentados os valores das estatísticas aqui consideradas.

Tabela 3: Estatísticas de teste para normalidade e homocedasticidade no modelo final.

	Estatística de teste	Estatística Bootstrap
Jarque-Bera	0,1284	2,7104
Breusch-Pagan-Godfrey	5,0813	6,7635
Koenker	5,7520	9.4871

Como nenhum valor das estatísticas de teste é maior a seu respectivo valor crítico bootstrap o modelo final cumpre as suposições de normalidade e homoscedasticidade, o número de condição é 22.47752 o qual é menor do que 30, que nos permite concluir que nosso modelo não tem problemas de multicolinearidade. O modelo estimado é apresentado a seguir:

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

$$\log(\bar{rate}) = 3.9538 - 0.2313 \log len + 0.1512 \log sigs1 - 0.0321\, slim - 0.3220\, PA + \hat{e}.$$

Pontos de alavanca e influentes: Para identificar se alguma observação tem padrão atípico, vamos considerar a medida de alavancagem para identificar os pontos de alavanca (padrão atípico nos regressores) e os resíduos studentizados, DFFITS e a distância de Cook para identificar observações influentes (erro da observação é grande). Segundo a medida de alavancagem a observação 25 é ponto de alavanca, o valor da medida de alavancagem para ela é 0,2988 a qual é maior do que 0,2564 (2 (5/39)), as observações 27 e 34 são influentes segundo a medida dos resíduos studentizados, o valor do resíduo studentizado para essas observações é 2,0610 e -2,6852 os quais são superiores em modulo a 2, de acordo com a distância de Cook não há nenhuma observação influente e segundo o DFFITS a observação 27 é influente, o valor do DFFITS para essa variável é 0,7831 o qual é maior do que 0,7161 $\left(2\sqrt{5/39}.\right)$ Um ponto de alavanca pode ou não, ser ponto influente, após análises de todas as medidas consideradas a observação 25 mostrou-se não influente. As observações 27 e 34 apresentaram-se influentes.

Para observar o comportamento das estimativas dos parâmetros sem as observações influentes, primeiro vamos considerar o modelo sem a observação 27, em seguida sem a observação 34 e finalmente sem as duas observações, as estimativas dos parâmetros, $\hat{\sigma}^2$, R^2e $\bar{R}^2$ são apresentadas na tabela a seguir.

Tabela 4: estimativas dos parâmetros do modelo de regressão sem as observações influentes.

	β_1	β_2	β_3	β_4	β_5	$\hat{\sigma}^2$	R^2	$\bar{R}^2$
Sem 27	4,0262	-0,2318	0,1261	-0,0347	-0,2661	0,2365	0,7510	0,7208
Sem 34	3,8920	-0,1924	0,1647	-0,0316	-0,3610	0,2276	0,7782	0,7514

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

Sem 27 e 34	3,9619	-0,1947	0,1409	-0,3075	-0,3075	0,2170	0,7849	0,7580

O menor valor de $\hat{\sigma}^2 = 0.217$ é obtido quando tiramos do modelo final as observações 27 e 34, também obtemos o maior valor do coeficiente de determinação ajustado $(\bar{R}^2 = 0.7580)$, por isso removemos as duas observações do modelo final, no entanto a remoção das observações deve ser feita em conjunto com o esperto no tema no qual se está trabalhando. Aqui foi feita a remoção das observações e a estimação do modelo sem elas.

O modelo estimado após a remoção das duas observações influentes é apresentado na equação a seguir, em geral, a variação nas estimativas não foi muito grande, a maior variação apresentou-se nas estimativas dos parâmetros que acompanham os regressores len e PA.

$$\log(\bar{rate}) = 3.9619 - 0.1947 \log len + 0.1409 \log sigs1 - 0.0341 slim - 0.3075 PA .$$

Análise de resultados: O modelo escolhido apresentado na equação acima explica o 78.49% da variabilidade total do logaritmo da taxa de acidentes por milhões de quilômetros de veículos.

Quando os regressores len e sigs1 são um, slim é zero e o tipo de estrada e diferente de PA, temos que o valor médio do logaritmo da taxa de acidentes é 3.961951, isto é, quando o comprimento do segmento em milhas é um, o número de sinais por quilômetro de estrada é um, o limite de velocidade é zero e o tipo de estrada é diferente de PA, o valor médio do logaritmo da taxa de acidentes é 3.961951, agora quando os regressores len e sigs1 são um, slim é zero e o tipo de estrada é PA, temos que o valor médio do logaritmo da taxa de acidentes é 3.654451 (3.961951-0.3075), podemos notar que a uma relação negativa entre as variáveis logaritmo do comprimento do segmento em milhas, limite de velocidade e tipo de estrada PA com a variável resposta.

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

Por cada unidade percentual que aumente no comprimento do segmento em milhas e mantendo o número de sinais por quilômetro, limite de velocidade e tipo de estrada constante, a taxa de acidentes de trânsito diminui 19.47%. A taxa de acidentes trânsito aumenta em 14.09%, por cada unidade percentual que aumente no número de sinais por quilômetro mantendo o comprimento do segmento em milhas, limite de velocidade e tipo de estrada fixo. A cada adição de uma unidade no limite de velocidade mantendo o número de sinais por quilômetro, o comprimento do segmento em milhas e tipo de estrada constante, a taxa de acidentes de trânsito diminui em 3.41%.

Considerações finais

Para a análise da taxa de acidentes de trânsito no estado de Minnesota nas estradas norte-americanas no ano de 1973 realizou-se o ajuste de um modelo de regressão, como os dados apresentavam problemas de multicolinearidade tentamos conseguir uma combinação das variáveis causadoras deste problema, de tal forma que não perdêssemos informação, para isso consideramos 126 modelos, nos quais observamos algumas possíveis variáveis que aumentavam o nível de significância de alguns parâmetros da regressão, no final, nenhum desses modelos mostrou um bom ajuste, pelo qual decidimos remover as variáveis menos significativas, deste modo o modelo final ficou dependendo das variáveis len, sigs1, slim e PA; como na análise descritiva observamos que as variáveis len, sigs1 e a variável resposta apresentam um melhor comportamento em escala logarítmica, foi considerado um modelo com as variáveis transformadas resultando ser o modelo que apresentou o melhor ajuste.

No modelo final se observou que há uma relação negativa entre as variáveis logaritmo do comprimento do segmento em milhas, limite de velocidade e tipo de estrada PA com a

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br

ISSN 2448-3230

variável resposta. Em geral, no senso comum atual o aumento do limite de velocidade é um fator que tende a aumentar o número de acidentes de trânsito, como é apresentado nas midas, desta forma os resultados obtidos parecem contrariar o senso comum, embora as características deste tipo de estudos possam ter mudado com o decorrer dos anos, além disso, não se consideraram algumas variáveis como: sono, álcool, medicamentos e drogas ilícitas que são importantes segundo Marin, 2000. Em um novo estudo se poderia considerar essas variáveis como variáveis explicativas.

Referências

1. Henri Theil, *Introduction Econometrics.* Prentice-Haull, Englewood Cliffs, N.J., 1978, p. 135.
2. Marin, Letícia e Queiroz, Marcos S. *A atualidade dos acidentes de trânsito na era da velocidade: uma visão geral.* Cad. Saúde Pública [online]. 2000, vol.16, n.1, pp. 7-21. ISSN 0102-311X.
3. Cribari-Neto, F. *Asymptotic inference under heteroskedasticity of unknown form*. Computational Statistics & Data Analysis 45, (2004), 215-233.
4. Autor desconhecido. *Estudo da Organização Mundial da Saúde (OMS) sobre mortes por acidentes de trânsito em 178 países é base par década de ações para segurança.* Em discussão, 2013. Disponível em: http://www.senado.gov.br Acesso em: 01 novembro. 2013.

USP, jjduartes@unal.edu.com
IF Setão Pernambucano, raquelmendiola@yahoo.com.br