Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

RELAÇÃO ENTRE EXPECTATIVA DE VIDA E CARACTERÍSTICAS DOS ESTADOS NORTE-AMERICANOS ENTRE OS ANOS DE 1969 E 1970

CE225 - Modelos Lineares Generalizados

Eduardo Elias Ribeiro Junior

Curitiba, 05 de setembro de 2014

Contents

1 Introdução

O presente estudo tem por objetivo explicar e quantificar, em estados norte-americanos, a relação entre a expectativa de vida nos anos 1969 e 1970 e algumas características destes estados. As características mencionadas são: população estimada em julho de 1975, renda per capita em 1974 em USD (United States dollar), proporção de analfabetos em 1970, taxa de criminalidade por 100 mil habitantes em 1976, porcentagem de estudantes que concluem o segundo grau em 1970, número de dias no ano com temperatura abaixo de 0°C na cidade mais importante do estado e área do estado em milhas quadradas. Substituiremos duas características, população estimada e área do estado pela característica densidade demográfica que será expressa pelo quociente $\frac{\text{população}}{\text{área}}$.

As características, denominadas como variáveis independentes, escolhidas para compor estudos desta natureza são, em geral, escolhidas subjetivamente levando em consideração o conhecimento prévio do pesquisador em relação ao fenômeno estudado. Neste caso é razoável a escolha destas sete variáveis para explicar a expectativa de vida, pois a priori parece haver correlação entre elas. Abaixo temos as dez primeiras observações do respectivo conjunto de dados.

Table 1: Conjunto de dados

Estado	Renda	Analfabetos	Expec. Vida	Crime	Estudos	Dias Frios	Densidade
Alabama	69.05	3624	2.10	15.10	41.30	20	0.07
Alaska	69.31	6315	1.50	11.30	66.70	152	0.00
Arizona	70.55	4530	1.80	7.80	58.10	15	0.02
Arkansas	70.66	3378	1.90	10.10	39.90	65	0.04
California	71.71	5114	1.10	10.30	62.60	20	0.14
Colorado	72.06	4884	0.70	6.80	63.90	166	0.02
Connecticut	72.48	5348	1.10	3.10	56.00	139	0.64
Delaware	70.06	4809	0.90	6.20	54.60	103	0.29
Florida	70.66	4815	1.30	10.70	52.60	11	0.15
Georgia	68.54	4091	2.00	13.90	40.60	60	0.08

2 Metodologia

Relações entre variável resposta e variáveis explicativas, como a descrita neste problema, podem ser analisadas em estatística com o auxílio da teoria de modelos lineares, mais especificamente regressão linear. Em nosso caso trabalharemos com regressão linear múltipla, cuja especificação é descrita abaixo:

$$\begin{aligned} Y|\underset{\sim}{X} &\sim Normal(\mu_{y|\underset{\sim}{x}}, \sigma^2) \\ E(Y|\underset{\sim}{X}) = \mu_{y|\underset{\sim}{x}} &= \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \cdots + \beta_p x_p \end{aligned} \tag{1}$$

Uma abordagem de regressão linear múltipla nos permitirá avaliar a variável de interesse por meio de variáveis explicativas, comumente chamadas de variáveis regressoras. Esta análise levará em conta um conjunto de variáveis regressoras com a finalidade de explicar a variação da variável independente.

3 Modelagem

Nesta seção faremos todas as etapas de análise dos dados desde descrição até predição pelo modelo proposto. As subseções presentes nesta seção carregarão a notação abaixo a fim de simplificar a escrita das variáveis:

- **expvi**: expectativa de vida nos anos 1969 e 1970 .
- **renda**: renda per capita em 1974 em USD (United States dollar).
- **analf**: proporção de analfabetos em 1970.
- **crime**: taxa de criminalidade por 100 mil habitantes em 1976.
- **estud**: porcentagem de estudantes que concluem o segundo grau em 1970.
- **ndias**: número de dias no ano com temperatura abaixo de 0°C na cidade mais importante do estado.
- **densi**: densidade demográfica em habitantes por milhas quadradas.

3.1 Análise descritiva e exploratória

Toda análise de dados incia-se por uma análise descritiva. A análise descritiva ilustra como serão as posteriores análises, indicando possíveis problemas e sugestões de modelagem.

Inicialmente exploraremos as medidas resumo das variáveis.

Table 2: Medidas resumo

	expvi	renda	analf	crime	estud	ndias	densi
Min.	67.96	3098.00	0.50	1.40	37.80	0.00	0.00
1st Qu.	70.12	3993.00	0.62	4.35	48.05	66.25	0.03
Median	70.68	4519.00	0.95	6.85	53.25	114.50	0.07
Mean	70.88	4436.00	1.17	7.38	53.11	104.50	0.19
3rd Qu.	71.89	4814.00	1.57	10.68	59.15	139.80	0.14
Max.	73.60	6315.00	2.80	15.10	67.30	188.00	2.68

De acordo com a 2 temos indícios de assimetria na disposição dos valores de proporção de analfabetos, taxa de criminalidade e densidade demográfica devido as medidas de posição presentes na tabela. A assimetria dos dados, ou ainda, observações muito dispersas nas variáveis explicativas podem acarretar em pontos influentes no estudo.

O interesse neste caso será avaliar a variável expectativa de vida em função das demais variáveis explicativas por meio de uma regressão linear, portanto uma análise exploratória visando avaliar preliminarmente se há relação entre essas variáveis é imprescindível. Na figura 1 são exibidos 49 gráficos que ilustram a relação de variáveis combinadas duas a duas, ou seja, temos todas as possíveis combinações de duas variáveis.

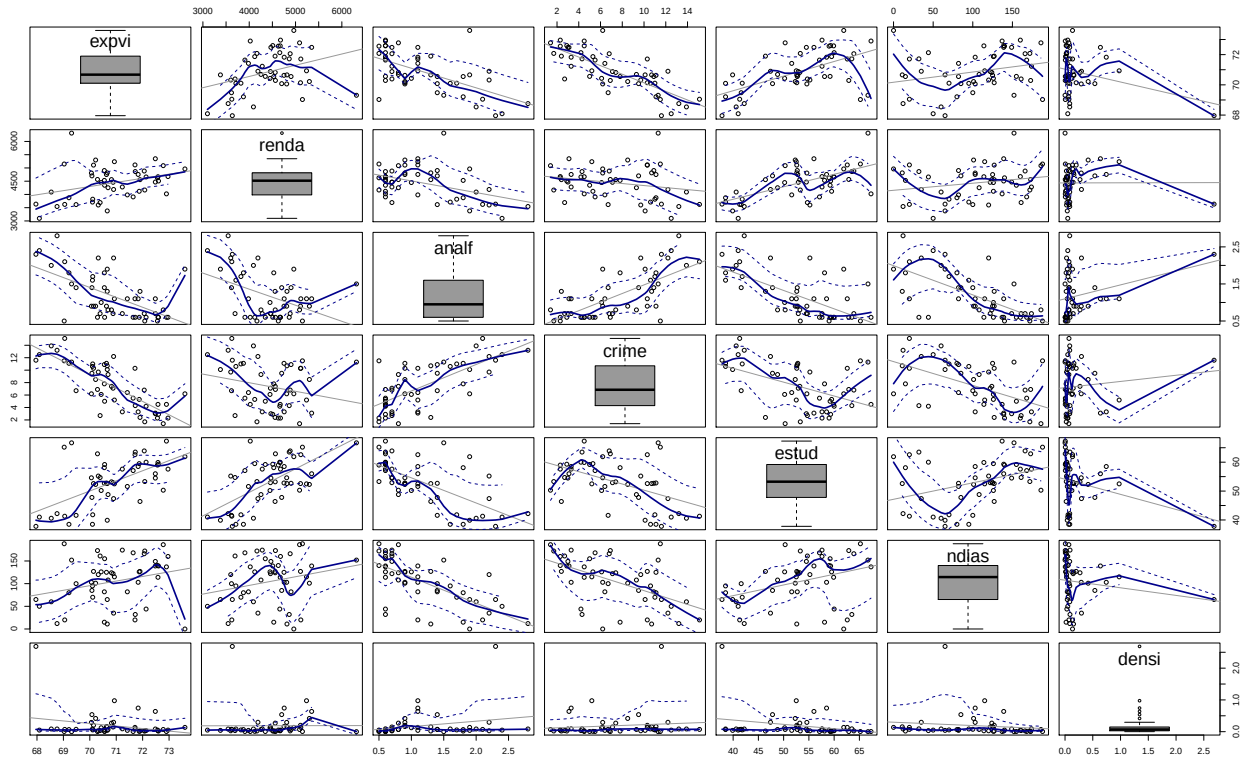

Figure 1: Representação Gráfica das Relações entre as Variáveis

A figura 1 apresenta uma matriz 7x7 de gráficos, sendo que na diagonal principal temos gráficos univariados, mais especificamente boxplots das variáveis em estudo, e nesta diagonal percebe-se a assimetria de algumas variáveis, como também visto na tabela 2. O que mais chama atenção dentre os gráficos univariados é o boxplot referente a densidade demográfica, pois neste percebe-se uma fortíssima assimetria à direita. Para os gráficos fora da diagonal principal, são exibidos gráficos bivariados identidade acima e abaixo da diagonal, que somente invertem seus eixos. Explorando esses gráficos temos na primeira linha a variável de interesse contra as demais variáveis e podemos observar a tendência da variável resposta em relação as variáveis explicativas. Agora excetuando a primeira linha temos os gráficos das variáveis regressoras duas a duas e nestes gráficos é

desejável que não se tenha evidências de relação linear entre elas, evitando a presença de colinearidade na análise. Visualmente a tendência linear mais evidente parece estar entre as variáveis proporção de analfabetos contra taxa de criminalidade e proporção de analfabetos contra porcentagem de estudantes concluintes do segundo grau. A fim de quantificar a relação linear entre as variáveis explicativas, para que não tenhamos problemas de colinearidade, vamos explorar a matriz de correlação entre elas.

Table 3: Matrix de Correlação entre as Variáveis

	expvi	renda	analf	crime	estud	ndias	densi
expvi	1.00	0.34	-0.59	-0.78	0.58	0.26	-0.26
renda	0.34	1.00	-0.44	-0.23	0.62	0.23	0.00
analf	-0.59	-0.44	1.00	0.70	-0.66	-0.67	0.25
crime	-0.78	-0.23	0.70	1.00	-0.49	-0.54	0.11
estud	0.58	0.62	-0.66	-0.49	1.00	0.37	-0.27
ndias	0.26	0.23	-0.67	-0.54	0.37	1.00	-0.13
densi	-0.26	0.00	0.25	0.11	-0.27	-0.13	1.00

Na tabela 3 a diagonal principal é preenchida com todos os elementos iguais a 1, pois a correlação de uma variável com ela mesma é perfeita. Observando os elementos fora da diagonal principal temos o maior valor absoluto igual a 0.70, proveniente da correlação entre as variáveis proporção de analfabetos e taxa de criminalidade, também observado nos gráficos da figura 1, porém não assumiremos como uma correlação forte para abandonarmos uma das variáveis antes de partirmos para os modelos de regressão.

3.2 Modelo com todos os efeitos aditivos

Após análise descritiva continuamos com todas as variáveis como candidatas a compor ao modelo. Como primeira opções ajustaremos um modelo aditivo saturado, ou seja, incluindo todas as variáveis sem considerar interação. Interações entre as variáveis não serão consideradas neste estudo, pois todas as variáveis explicativas são numéricas. Interações neste caso dificultam a interpretação dos parâmetros do modelo e não auxiliam na identificação das relações marginais, interesse de nosso estudo.

A forma do modelo ajustado será conforme descrito na seção 2, abaixo temos sua representação com os respectivos nomes das variáveis:

$$\begin{gathered} Y|\underset{\sim}{X}\ Normal(\mu_{\hat{y|\underset{\sim}{x}}}, \sigma^2) \\ \mu_{\hat{y|\underset{\sim}{x}}} = \hat{\beta}_0 + \hat{\beta}_1 renda + \hat{\beta}_2 analf + \hat{\beta}_3 crime + \hat{\beta}_4 estud + \hat{\beta}_5 densi + \hat{\beta}_6 ndias \end{gathered} \tag{2}$$

Após modelo ajustado faremos a seleção das variáveis que permanecerão no modelo final, porém vamos realizar um breve diagnóstico do modelo saturado para averiguação da adequação do modelo.

Na figura 2 não notamos nenhuma fuga dos pressupostos para o modelo. No primeiro dos gráficos exibidos os pontos parecem se comportar aleatoriamente com resíduos positivos e negativos de magnitudes aleatórias considerando os valores preditos. No segundo gráfico temos os resíduos padronizados se dispondo sobre a linha teórica da distribuição Normal, com isso não rejeitaremos a hipótese dos resíduos se distribuírem normalmente. No terceiro gráfico não temos evidência de tendência e novamente percebemos uma disposição aleatória dos pontos não caracterizando uma relação média variância. O último gráfico nos indica suspeitos a outliers, ou seja, pontos fora das bandas em vermelho seccionado podem ser classificados como observações influentes, neste caso não temos indícios de observações influentes. Com isso podendo seguir com as demais análises a partir deste modelo.

3.3 Seleção de variáveis

Um ponto importante no processo de modelagem é a seleção de variáveis para compor o melhor modelo. A nomenclatura "melhor modelo" não seria a mais adequada ao se tratar de modelos estatísticos, a nomenclatura correta para a análise nesta seção seria "escolha do melhor modelo segundo algum critério". O critério adotado é subjetivo, sendo função do estatístico justificar o critério adotado.

A seleção de variáveis para este trabalho será considerando o algoritmo **stepwise**, que fará a inclusão e exclusão de variáveis no modelo simultaneamente. O algoritmo considerará como medida de seleção o AIC (Critério de Informação de Akaike), cujo a fórmula está descrita abaixo:

$$AIC_{model} = -2log(L) + 2p, \begin{cases} L : \text{Verossimilhança} \\ p : \text{número de parâmetros} \end{cases}$$

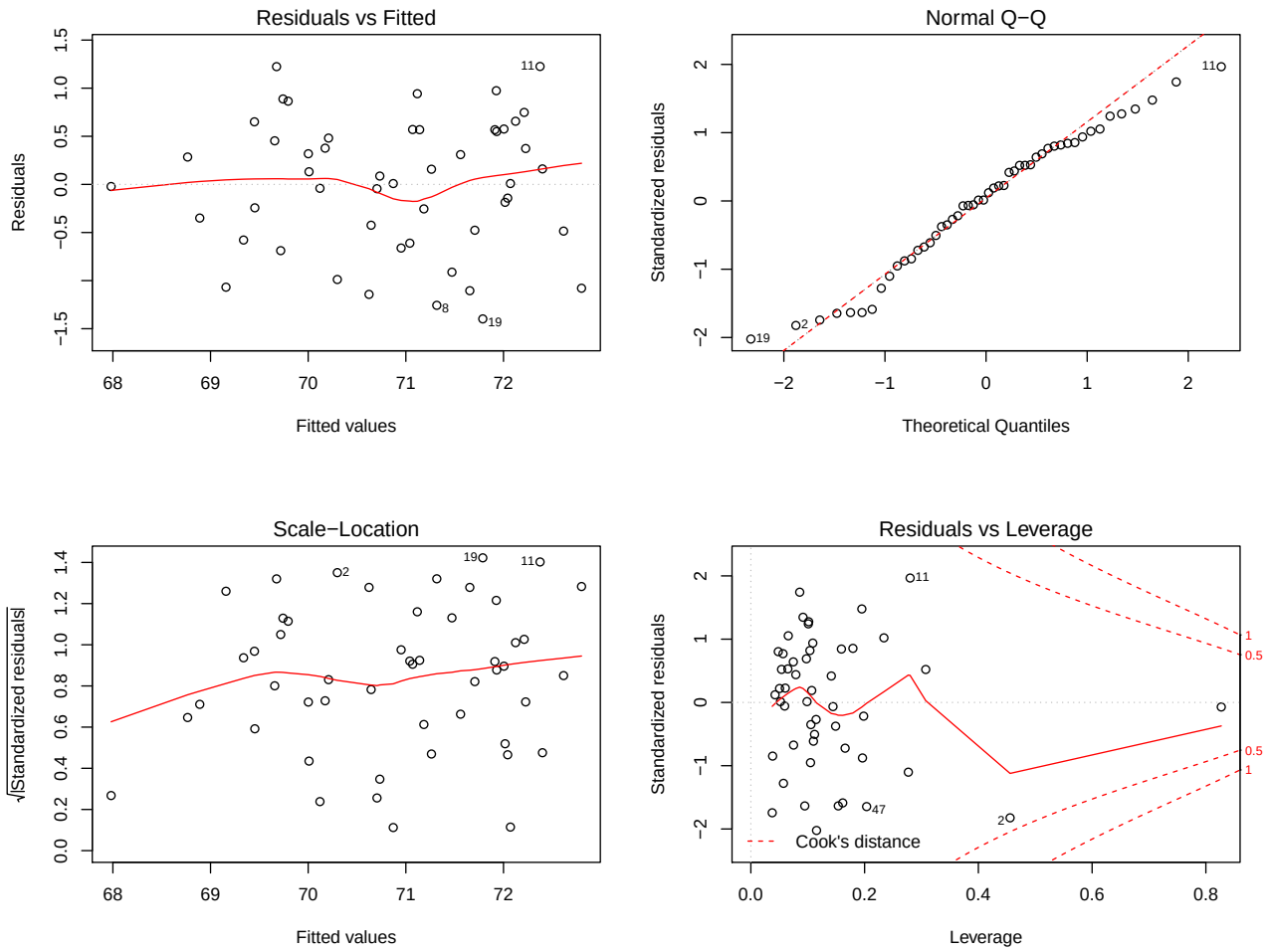

Figure 2: Análise de Diagnóstico do Modelo Saturado

O AIC é inversamente proporcional à log-verossimilhança do modelo e diretamente proporcional ao número de parâmetros, ou seja, o critério busca um modelo parcimonioso, penalizando modelos com um número excessivo de parâmetros. Apresentaremos no quadro 1 a última iteração do algoritmo stepwise com as variáveis selecionadas e as descartadas no modelo.

Quadro 1: Algoritmo Stepwise para seleção de variáveis

```
## Step:  AIC=4.03
## expvi ~ crime + estud + densi + ndias
##
##         Df Sum of Sq    RSS      Cp F value    Pr(>F)
## <none>               23.766  4.0341
## - densi  1     1.606 25.372  5.0100  3.0412  0.088005 .
## + renda  1     0.544 23.221  5.0256  1.0313  0.315413
## + analf  1     0.058 23.708  5.9275  0.1068  0.745403
## - estud  1     4.294 28.060  9.9911  8.1316  0.006546 **
## - ndias  1     4.682 28.447 10.7086  8.8648  0.004669 **
## - crime  1    33.342 57.108 63.8122 63.1333 4.136e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pelo quadro 1 podemos verificar que as variáveis renda per capita e proporção de analfabetos foram descartadas do modelo (note o sinal positivo ao lado das variáveis) resultando no modelo com AIC = 4.03, sugerido pelo algoritmo. Ainda neste quadro foram exibidos as significâncias dos efeitos das variáveis considerando o teste F, perceba que mesmo adotando este outro critério as variáveis renda per capita e proporção de analfabetos seriam retiradas do modelo ao nível de significância de 10%.

Nas próximas subseções seguiremos as análises considerando agora somente as variáveis taxa de criminalidade, porcentagem de estudantes do segundo grau, densidade demográfica e número de dias com temperatura abaixo de 0°C na cidade mais importante do estado para especificação do modelo.

3.4 Modelo proposto

Novamente especificaremos o modelo aditivo mas agora considerando somente as variáveis selecionadas na subseção anterior. O modelo adotado é:

$$\begin{aligned} Y|\underset{\sim}{X} &\sim Normal(\mu_{\hat{y}|\underset{\sim}{x}}, \sigma^2) \\ \mu_{\hat{y}|\underset{\sim}{x}} &= \hat{\beta}_0 + \hat{\beta}_1 crime + \hat{\beta}_2 estud + \hat{\beta}_3 densi + \hat{\beta}_4 ndias \end{aligned} \tag{3}$$

No quadro 2 apresentaremos o resumo do modelo, com parâmetros estimados e respectivos testes marginais (considerando a distribuição *t-student*) para verificar a significância dos efeitos estimados.

Quadro 2: Resumo do modelo (2): estimativas, erros-padrão e significâncias

```
##
## Call:
## lm(formula = expvi ~ crime + estud + densi + ndias, data = da)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56186 -0.53406  0.06514  0.54981  1.23075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.501488   0.998428  71.614  < 2e-16 ***
## crime       -0.285873   0.035979  -7.946 4.14e-10 ***
## estud        0.043649   0.015307   2.852  0.00655 **
## densi       -0.456832   0.261960  -1.744  0.08800 .
## ndias       -0.007141   0.002399  -2.977  0.00467 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7267 on 45 degrees of freedom
## Multiple R-squared:  0.7309,Adjusted R-squared:  0.7069
## F-statistic: 30.55 on 4 and 45 DF,  p-value: 2.609e-12
```

No quadro 2 percebemos que, mesmo após utilizado o critério de AIC para selecionar as variáveis, temos o efeito da variável densidade demográfica não significativo, considerando o teste t ao nível de significância de 5%, então optaremos por abandonar esta variável do modelo.

O novo modelo terá

$$\mu_{\hat{y}|\underset{\sim}{x}} = \hat{\beta}_0 + \hat{\beta}_1 crime + \hat{\beta}_2 estud + \hat{\beta}_3 ndias \tag{4}$$

E novamente apresentaremos o quadro resumo do novo modelo especificado:

Quadro 3: Resumo do modelo (3): estimativas, erros-padrão e significâncias

```
##
## Call:
## lm(formula = expvi ~ crime + estud + ndias, data = da)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5015 -0.5391  0.1014  0.5921  1.2268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
## crime       -0.283065   0.036731  -7.706 8.04e-10 ***
## estud        0.049949   0.015201   3.286  0.00195 **
## ndias       -0.006912   0.002447  -2.824  0.00699 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7427 on 46 degrees of freedom
## Multiple R-squared:  0.7127,Adjusted R-squared:  0.6939
## F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

Aqui percebemos que todas as variáveis inclusas no modelo são significativas ao nível de significância de 1%. Além disto percebemos, pelas estimativas dos parâmetros, que a relação entre expectativa de vida e taxa de criminalidade é decrescente, ou seja, quanto maior a taxa de criminalidade menor a expectativa de vida, o mesmo acontece com o número de dias com temperatura abaixo de zero, somente observamos uma relação crescente entre expectativa de vida e percentual de estudantes concluintes do segundo grau, ou seja, quanto maior o percentual de estudantes concluintes do segundo grau maior a expectativa de vida, estas relações serão abordadas com detalhe na seção 3.6.

3.5 Análise de diagnóstico

Antes de validar qualquer análise previamente feita com o modelo devemos realizar uma análise de diagnóstico. A análise de diagnóstico tem o papel de verificar a adequação do modelo aos dados, averiguar se todos os pressupostos considerados são atendidos e verificar observações influentes.

Faremos a análise gráfica em três etapas. Na primeira etapa verificaremos os pressupostos, na segunda se existe a necessidade de fatores quadráticos e na terceira se há observações isoladas influenciando significativamente o modelo.

- Verificação dos pressupostos

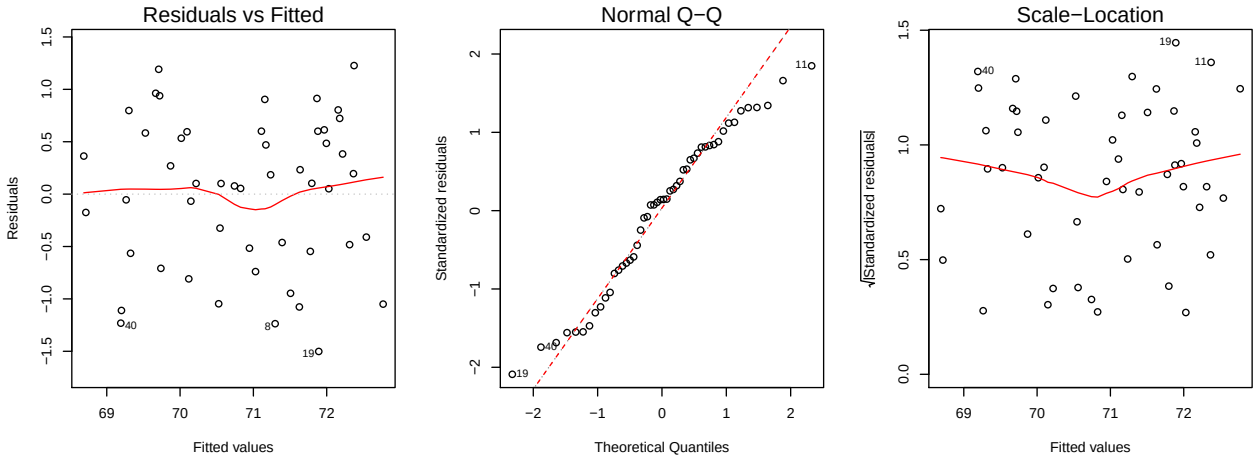

Figure 3: Análise de Diagnóstico do Modelo Proposto - Pressupostos

Estes gráficos apresentam a mesma interpretação da figura 2 não há evidências de fuga de pressupostos.

- Necessidade de fatores quadráticos

Quadro 4: Adequação do modelo com relação a fatores quadráticos

```
##            Test stat Pr(>|t|)
## crime          0.509    0.613
## estud         -0.748    0.458
## ndias          0.090    0.929
## Tukey test     0.020    0.984
```

No quadro 4 são exibidos as estatísticas t, calculadas para o fator quadrático se incluído no modelo apresentando também o p-valor associado ao fator quadrático. Com isso concluímos que não há a necessidade de inclusão de fatores quadráticos no modelo.

- Observações Influentes

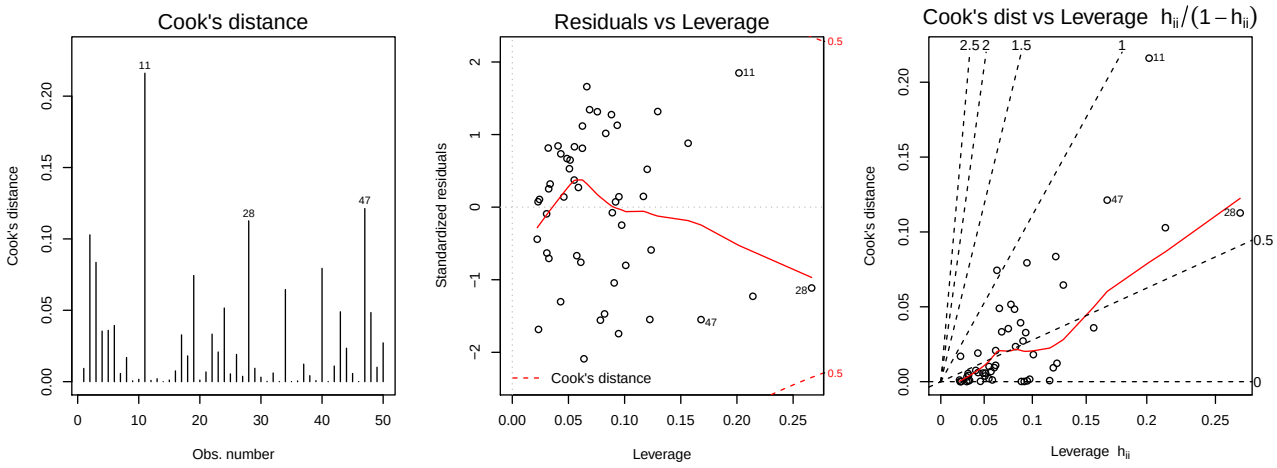

Figure 4: Análise de Diagnóstico do Modelo Proposto - Observações Influentes

Na figura 4 temos três gráficos sendo que o primeira avalia a medida distância de Cook, usualmente adota-se a regra de observações com distância de Cook maior que 1 como suspeitas a outliers, no gráfico temos somente a observação 11 com distância de Cook próxima a 0.2, portanto sem indícios de pontos influentes neste gráfico. Os outros dois gráficos levam em conta o poder de alavancagem de cada observação e novamente não temos observações suspeitas.

É importante lembrar que a análise de diagnóstico foi subdivida, mas as três etapas não são independentes, geralmente quando uma das etapas aponta alguma incoerência na análise os demais também a indicam.

3.6 Predições marginais

O modelo proposto após as análises contém três variáveis explicativas e uma de interesse caracterizando a modelagem em um hiperplano de quatro dimensões, portanto a visualização gráfica conjunta é impossível. Porém podemos visualizar graficamente as relações marginais, variando uma das variáveis explicativas e fixando as outras duas. Os gráficos marginais são exibidos na figura 5:

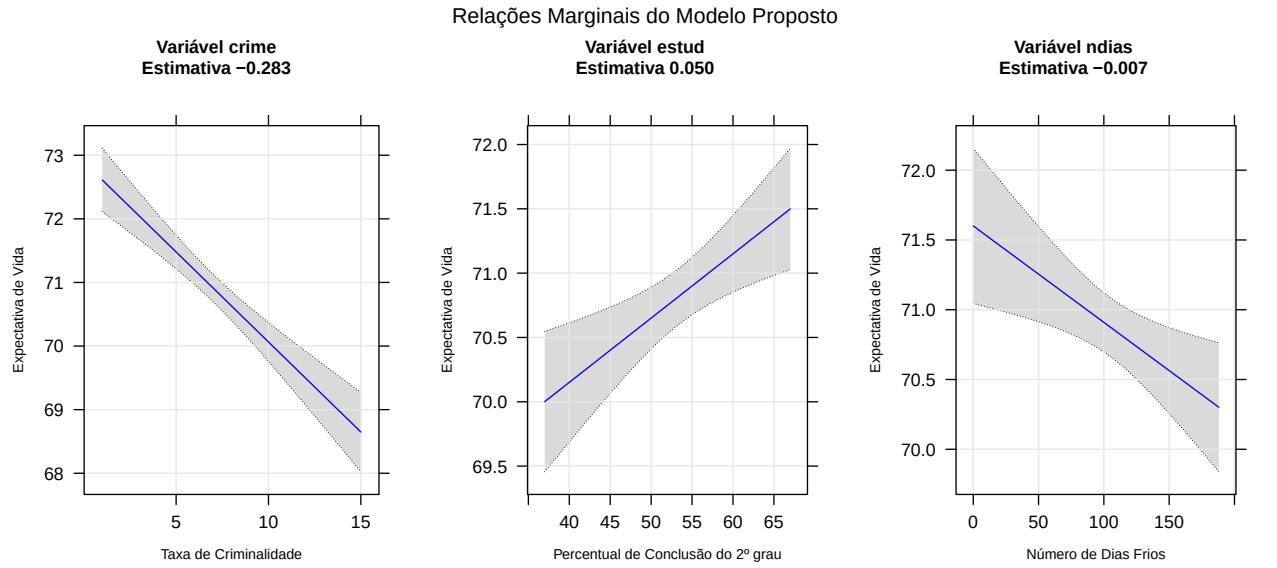

Figure 5: Gráficos Marginais do Modelo Proposto

Na figura 5 podemos visualizar as relações mencionadas na seção 3.4, os gráficos foram construídos fixando as variáveis taxa de criminalidade e porcentagem de estudantes concluintes do segundo grau em suas respectivas médias e a variável número de dias no ano com temperatura abaixo de 0°C na cidade mais importante do estado

em 115, que equivale ao valor arredondado de sua mediana. Para cada um dos gráficos variamos somente o valor do eixo x e verificamos o valor predito pelo modelo no eixo y, também são apresentadas as bandas de confiança para a média com 95% de confiança.

A interpretação das relações (crescente ou decrescente) já foi mencionada, mas aqui complementaremos com a interpretação do parâmetros que também estão expostos nos gráficos da figura 5. O valor estimado para β_0 não está no gráfico mas equivale a 71.036 e é interpretado como o valor estimado para a expectativa de vida quando as outras variáveis são fixadas em 0, mas na prática não podemos realizar esta predição, pois a predição não pode extrapolar o intervalo de valores utilizado para a modelagem e conforme pode ser visto na tabela 2 somente a variável número de dias no ano com temperatura abaixo de 0°C na cidade mais importante do estado tem o valor 0 presente. Para o valor de β_1 podemos interpretá-lo como o valor médio estimado de decréscimo da variável resposta a cada uma unidade acrescida na taxa de criminalidade quando fixado as outras variáveis explicativas. Analogamente a interpretação de β_1, para β_2 e β_3 temos a mesma interpretação, porém alterando os valores: estima-se pelo modelo proposto que para uma unidade acrescida na porcentagem de alunos concluintes do segundo grau em média teremos um aumento de 0.05 na expectativa de vida do estado, mantendo fixados os valores da taxa de criminalidade e do número de dias com temperatura abaixo de 0°C e para uma unidade a mais no número de dias com temperatura abaixo de 0°C espera-se em média um decréscimo de 0.007 na expectativa de vida nas mesmas condições.

4 Conclusões

Pelo estudo foi possível constatar as relações que inicialmente desejávamos. Porém os pequenos valores das estimativas não são tão expressivos ao olhar somente a estimativa isolada, mas considerando também as medidas de dispersão percebe-se que as relações estão bem caracterizadas. Como já esperado a expectativa de vida nos estados norte-americanos decresce com relação a taxa de criminalidade e número de dias com temperatura abaixo de 0ºC na cidade mais importante do estado e apresenta um crescimento com relação ao percentual de estudantes que concluem o segundo grau já com relação as outras variáveis renda per capita, proporção de analfabetos e densidade demográfica não foi possível identificar uma relação expressiva com a expectativa de vida. Um ponto que pode ser questionável está na obtenção dos dados, temos dados provenientes de períodos bem distintos o que pode acarretar em interpretações não válidas na realidade.