**Modelos Lineares Generalizados**

LISTA DE EXERCÍCIOS nº 01

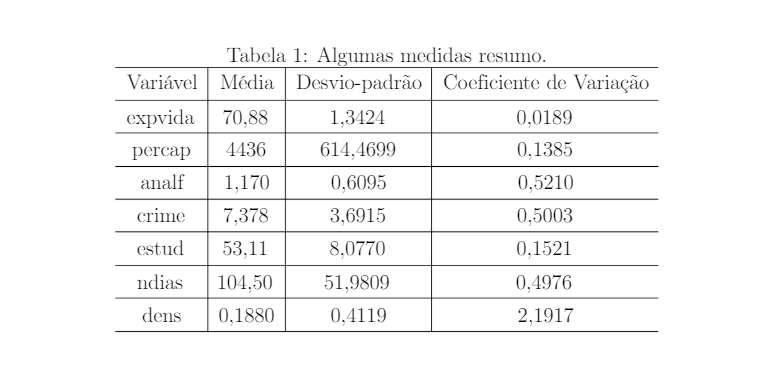
**1. No arquivo reg3.dat são descritas as seguintes variáveis referentes a 50 estados norte-americanos: (i) estado (nome do estado), (ii) pop (população estimada em julho de 1975), (iii) percap (renda per capita em 1974 em USD), (iv) analf (proporção de analfabetos em 1970), (v) expvida (expectativa de vida em anos 1969-70), (vi) crime (taxa de criminalidade por 100000 habitantes 1976), (vii) estud (porcentagem de estudantes que concluem o segundo grau 1970), (viii) ndias (número de dias do ano com temperatura abaixo de zero grau Celsius na cidade mais importante do estado) e (ix) area (área do estado em milhas quadradas). O objetivo do estudo é tentar explicar a variável expvida usando um modelo de regressão normal linear dadas as variáveis explicativas percap, analf, crime, estud, ndias e dens, em que dens=pop/area. Inicialmente faça uma análise descritiva dos dados, por exemplo com boxplots de cada uma das variáveis que serão consideradas no estudo e com diagramas de dispersão com as respectivas tendências entre as variáveis explicativas e a variável resposta. Comente essa parte descritiva. Posteriormente, ajuste o modelo de regressão normal linear com todas as variáveis explicativas e através do método AIC faça uma seleção de variáveis. Uma vez selecionado o modelo faça uma análise de diagnóstico e apresente as interpretações dos coeficientes estimados do modelo final.**

**Dados: https://www.ime.usp.br/~giapaula/reg3.dat**

Variável dependente: expvida

Linear models no SAS

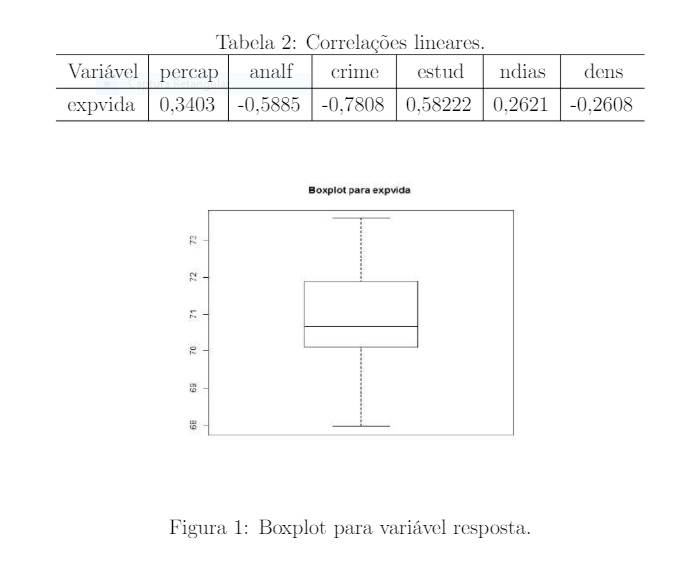
Análise Descritiva: Statistics no SAS



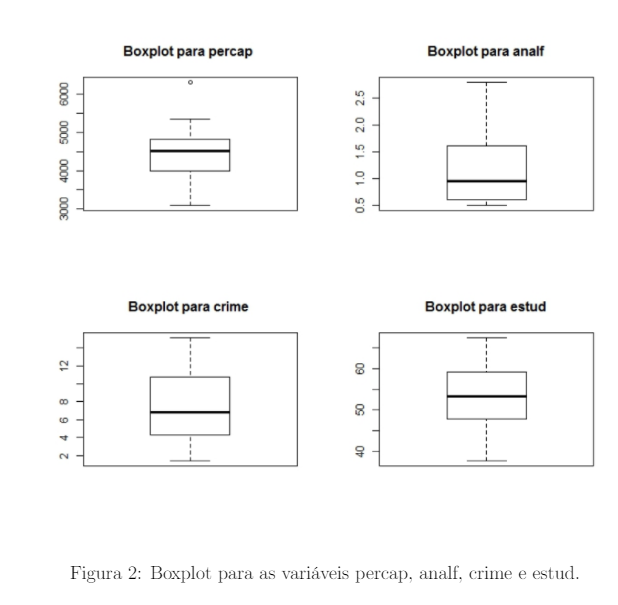
* Os indivíduos norte-americanos

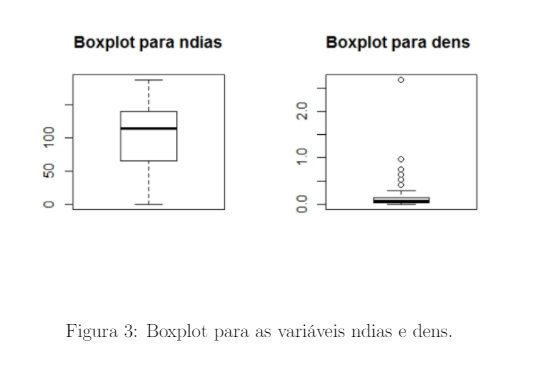
no período de 1969-1970 vivem, em média, 70,88 anos;

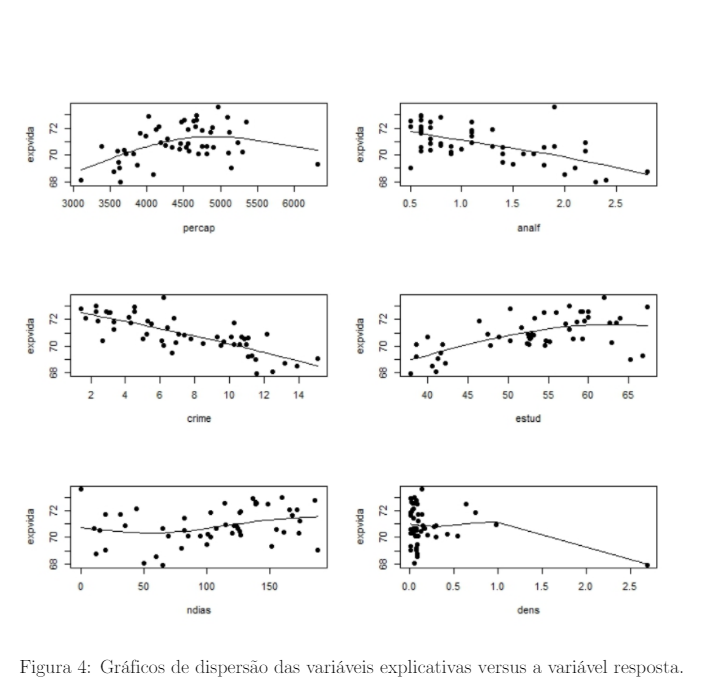
* A variação das expectativas de vida são em torno de 1,3424 (baixa) e que o coeficiente de variação é igual a 0,0189, ou seja média amostral da expectativa de vida é representativa;
* A renda per capita dos indivíduos, em média, é de 4436(representativa) com variação de 614,46999 ;
* A proporção média de analfabetos nos anos de 1970 é de 1,17 (média amostral não representativa) com variação de 0,6095;
* Em média, a taxa de criminalidade dos estados por 100000 habitantes em 1976 é de 7,3780 (média representativa)com variação de 3,6915;
* A porcentagem de estudantes que concluem o segundo grau 1970, em média, nos estados norte-americanos é de 53,11% (média representativa);
* O número médio de dias do ano com temperatura abaixo de zero grau Celsius na cidade mais importante do estado, é de 104,5 dias (média representativa) com variação de 1,9809 dias;
* A densidade média populacional dos estados norte-americanos é 0,1880 habitantes por milhas quadradas (média não representativa).



Os box plots mostram que as variáveis expvida , analf, crime e dens possuem distribuições assimétricas à direita. Já as variáveis percap e ndias possuem distribuições assimétricas à esquerda, já a variável estud aparenta se distribuir simetricamente.







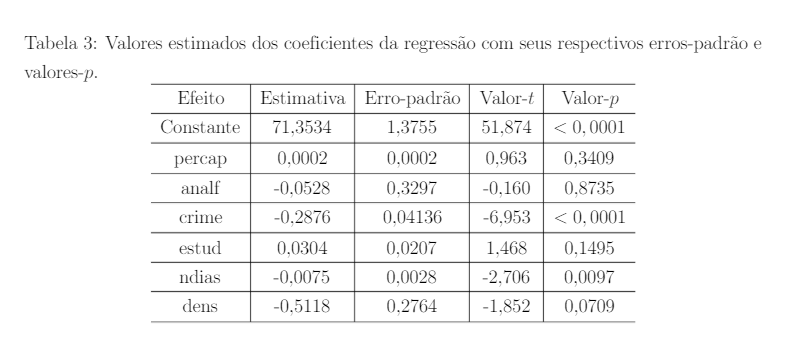
Os gráficos de dispersão acima representam cada variável explicativa versus a resposta expvida com suas respectivas tendências, as quais indicam o quanto cada variável independente é relacionada linearmente com a variável expvida. Observemos que as tendências nos levam a acreditar que as únicas variáveis que parecem não trazer informação para explicar a expectativa de vida são percap e ndias, o que pode ser feito através da análise da correlação linear que cada variável explicativa possui com expvida.

Modelo de Regressão: Linear models do SAS

yi=α+β1×percapi,+β2×analfi+β3×crimei+β4×estudi+β5×ndiasi+β6×densi+ǫi,

em que Yi denota a expectativa de vida do i-ésimo estado norte-americano da amostra, ǫi∼

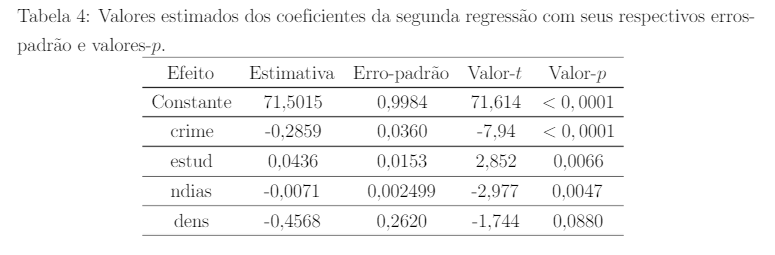
N(0; σ2)eα, β1, β2, β3, β4, β5, β6 correspondem aos coeficientes do intercepto e das variáveis explicativas, respectivamente, os quais devem ser estimados. Pelos dados apresentados, temos que as variáveis percap, analf e estud não se mostraram significativas ao nível de 10% para explicar expvida. Para realizar uma seleção de covariáveis que devem entrar no modelo, utilizamos o método de Akaike.



Utilizando o Critério de Akaike para seleção de variáveis explicativas, o modelo fica:

yi=α+β1×crimei+β2×estudi+β3×ndiasi+β4×densi+ǫi

em que as estimativas para os parâmetros do modelo estão apresentadas na Tabela 4.



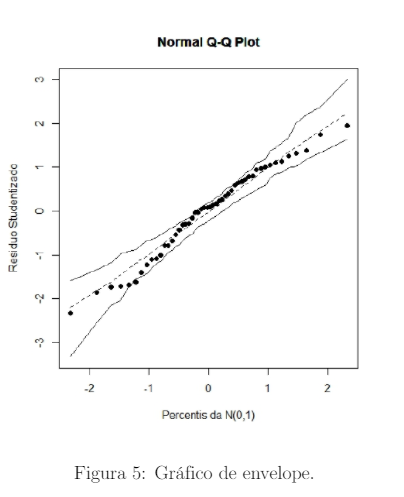
todas as variáveis são significativas ao nível de 10%. Logo, o modelo que

consideramos agora é dado por:

ˆyi= ˆµi= 71,5015 −0,2859 ×crimei+ 0,0436 ×estudi−0,0071 ×ndiasi−0,4568 ×densi

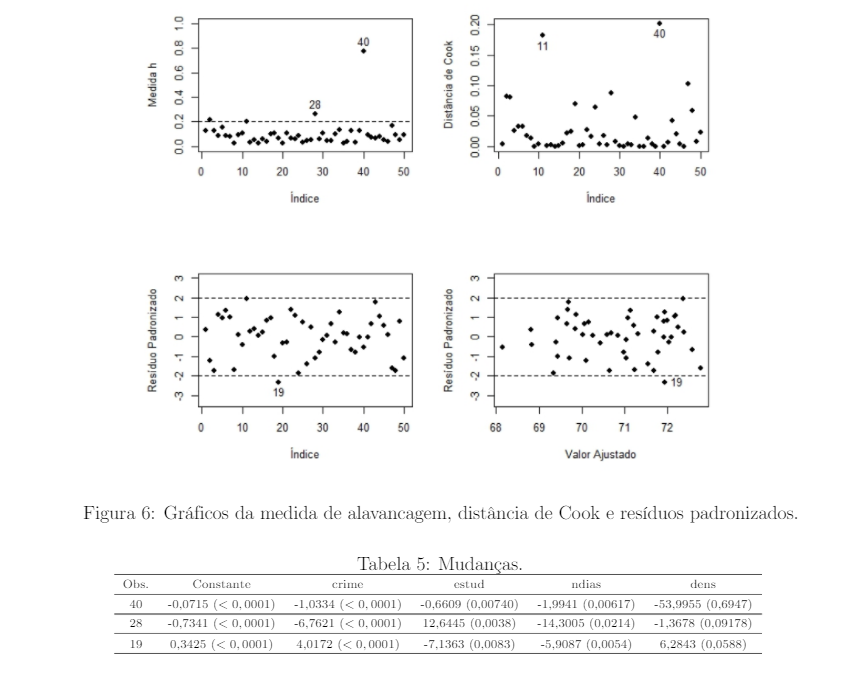
Diagnóstico:

O gráfico de envelope mostra que todas as observações estão dentro do envelope, confirmando que a suposição de normalidade neste caso é válida.



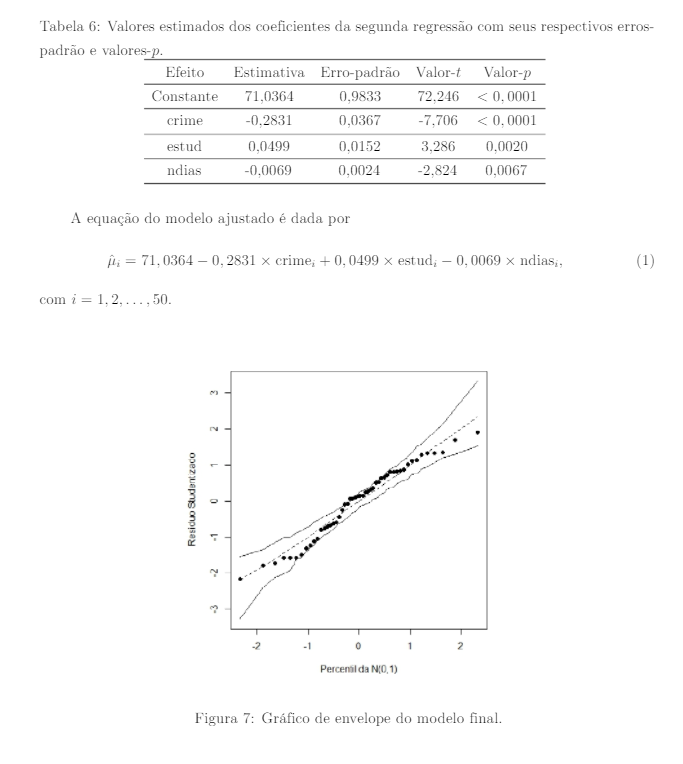
Os gráficos de índices e valores ajustados versus os resíduos padronizados não há fortes indícios de fugas da suposição de homocedasticidade do modelo. Há dois possíveis pontos de alavanca e influência (28 e 40)podem estar interferindo na estimação dos coeficientes da regressão, como também, podem causar mudança inferencial.

O modelo foi ajustado retirando cada ponto identificado, e calculando a mudanças relativa nos nos coeficientes estimados e observamos os valores-p(em parênteses)apresentadas na Tabela 5



A observação 40 é extremamente influente, pois além de interferir muito na estimativa do coeficiente associado a variável dens, esta observação causa mudança inferencial. Note que quando, retiramos o ponto 40, a variável dens deixa de ser significativa ao nível de 10% que era o que tínhamos obtido anteriormente. As demais observações identificadas não causam mudanças desproporcionais nas estimativas dos parâmetros nem causam mudança inferencial.

Modelo Final:



A observação 40 faz com que a variável dens entre no modelo quando a mesma não deve entrar. Identificado este problema, estimamos o modelo novamente, mas agora sem considerar a variável dens. Todas as variáveis são significativas a 5%, o gráfico de envelope para este modelo diz que não há fugas da suposição de normalidade do modelo.

A equação do modelo ajustado é dada por:

ˆµi= 71,0364 −0,2831 ×crimei+ 0,0499 ×estudi−0,0069 ×ndiasi,(1)

com i= 1,2,...,50

No modelo final dado nota-se que a média é função diretamente do componente sistemático, assim, temos que valores negativos dos coeficientes estimados indicam que quanto maior for o valor da variável explicativa menor será a expectativa de vida, obtemos que quanto maior for a taxa de criminalidade menor será a expectativa de vida, como também, quanto maior o número de dias com temperatura abaixo de zero menor será a expectativa de vida. Ambos resultados são esperados intuitivamente. Por fim, quanto maior a porcentagem de estudantes que concluem o ensino médio, maior será a expectativa de vida (coeficiente estimado positivo).

**2. (Neter et el., 1996, p. 449). No arquivo vendas.dat são descritas informações a respeito das vendas no ano anterior de um tipo de telhado de madeira em 26 filiais de uma rede de lojas de construção. As variáveis estão colocadas na seguinte ordem: (i) telhados, total de telhados vendidos (em mil metros quadrados), (ii) gastos, gastos pela loja com promoções do produto (em mil USD), (iii) clientes, número de clientes cadastrados na loja (em milhares), (iv) marcas, número de marcas concorrentes do produto e (v) potencial, potencial da loja (quanto maior o valor maior o potencial). Um dos objetivos do estudo com esse conjunto de dados é tentar prever o número esperado de telhados vendidos dadas as variáveis explicativas. Faça inicialmente uma análise descritiva construindo, por exemplo, os diagramas de dispersão de cada variável explicativa contra a variável resposta telhados. Calcule também as correlações entre as variáveis. Use os métodos stepwise e AIC para selecionar um modelo de regressão normal linear. Se o modelo selecionado for diferente pelos dois métodos, adote algum critério para escolher um dos modelos. Interprete os coeficientes estimados do modelo selecionado. Faça uma análise de diagnóstico para verificar se existem afastamentos sérios das suposições feitas para o modelo e se existem observações discrepantes.**

**Dados:** [**https://www.ime.usp.br/~giapaula/vendas.dat**](https://www.ime.usp.br/~giapaula/vendas.dat)