em que $\mathrm{D}(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$.

19. 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 percapita 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 Celsus na cidade mais importante do estado) e (ix) `area` (área do estado em milhas quadradas).

    O objetivo do estudo é tentar explicar e 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.

20. (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 variá-

veis 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.

21. (Wood, 1973). No arquivo **reg4.dat** estão os dados referentes à produção de gasolina numa determinada refinaria segundo três variáveis observadas durante o processo e uma quarta variável que é uma combinação das três primeiras. A resposta é o número de octanas do produto produzido. A octanagem é a propriedade que determina o limite máximo que a gasolina, junto com o ar, pode ser comprimida na câmara de combustão do veículo sem queimar antes de receber a centilha vinda das velas. As melhores gasolinas têm uma octanagem alta. Em grandes refinarias, o aumento de um octana na produção de gasolina pode representar um aumento de alguns milhões de dolares no custo final

da produção. Assim, torna-se importante o controle dessa variável durante o processo de produção. Use o método `AIC` para selecionar as variáveis explicativas significativas. Faça uma análise de diagnóstico com o modelo selecionado. Comente.

22. (Narula e Stangenhaus, 1988, pgs. 31-33). No arquivo **imoveis.dat** são apresentados dados relativos a uma amostra de 27 imóveis. Na ordem são apresentados os valores das seguintes variáveis: (i) imposto do imóvel (em 100 USD), (ii) área do terreno (em 1000 pés quadrados), (iii) área construída (em 1000 pés quadrados), (iv) idade da residência (em anos) e (v) preço de venda do imóvel (em 1000 USD). Ajuste um modelo normal linear do preço de venda contra as demais variáveis. Use o método `AIC` para selecionar as variáveis explicativas. Faça uma análise de diagnóstico com o modelo selecionado. Interprete os coeficientes estimados.

23. (Ryan e Joiner, 1994, p. 299). No arquivo **trees.dat** é apresentado um conjunto de dados que tem sido analisado sob diversos pontos de vista por vários pesquisadores (ver, por exemplo, Jørgensen, 1989). As variáveis observadas são o diâmetro (d), a altura (h) e o volume (v) de uma amostra de 31 cerejeiras numa floresta do estado da Pensilvânia, EUA. A relação entre diâmetro, altura e volume de uma árvore depende da forma da mesma e pode-se considerar duas possibilidades

$$v = \frac{1}{4}\pi d^2 h$$

para forma cilíndrica e

$$v = \frac{1}{12}\pi d^2 h$$

para forma cônica. Em ambos os casos a relação entre $\log v$, $\log d$ e $\log h$

é dada por

$$\log v = a + b\log d + c\log h.$$

Supor inicialmente um modelo linear em que $\epsilon \sim \mathrm{N}(0, \sigma^2)$. Faça uma análise de diagnóstico e verifique se é possível melhorar o modelo, por exemplo incluindo algum termo quadrático.

24. (Ruppert, 2004). No arquivo **capm.dat** estão os seguintes dados: Tbill (taxa de retorno livre de risco), retorno Microsoft, SP500 (retorno do mercado), retorno GE e retorno FORD de janeiro de 2002 a abril de 2003. Todos os retornos são diários e estão em porcentagem. Faça inicialmente os diagramas de dispersão entre os excessos de retorno $(y_{rt} - r_{ft})$ de cada uma das empresas Microsoft, GE e FORD e os excessos de retorno do mercado $(r_{mt} - r_{ft})$, em que y_{rt} denota o retorno da ação da empresa, r_{mt} é o retorno do mercado e r_{ft} indica a taxa livre de risco durante o t-ésimo período. Posteriormente, ajuste o seguinte modelo de regressão:

$$y_{rt} - r_{ft} = \alpha + \beta(r_{mt} - r_{ft}) + \epsilon_t,$$

em que $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$. Verifique a significância do parâmetro α e compare e interprete as estimativas intervalares para β. Faça uma análise de diagnóstico para cada modelo ajustado.

25. O conjunto de dados descrito na tabela abaixo refere-se a um estudo cujo objetivo foi tentar prever o preço de venda de um imóvel (em mil USD) dada a área total (em mil pés quadrados) numa região de Eugene, EUA (Gray, 1989). Esses dados estão armazenados no arquivo externo **reg1.dat**.

    Tente inicialmente ajustar uma regressão normal linear para explicar o preço dada a renda. Faça uma análise de diagnóstico e proponha

algum modelo alternativo (se for o caso) a fim de reduzir as eventuais influências de observações discrepantes bem como afastamentos de suposições feitas para o modelo. Interprete as estimativas obtidas para os coeficientes do modelo proposto.

Área	800	950	910	950	1200	1000	1180	1000
Preço	30,6	31,5	33,3	45,9	47,4	48,9	51,6	53,1
Área	1380	1250	1500	1200	1600	1650	1600	1680
Preço	54,0	54,3	55,2	55,2	56,7	57,9	58,5	59,7
Área	1500	1780	1790	1900	1760	1850	1800	1700
Preço	60,9	60,9	62,4	63,0	64,5	66,0	66,3	67,5
Área	1370	2000	2000	2100	2050	1990	2150	2050
Preço	68,4	68,4	68,7	69,6	70,5	74,7	75,0	75,3
Área	2200	2200	2180	2250	2400	2350	2500	2500
Preço	79,8	80,7	80,7	83,4	84,0	86,1	87,0	90,3
Área	2500	2500	2680	2210	2750	2500	2400	3100
Preço	96,0	101,4	105,9	111,3	112,5	114,0	115,2	117,0
Área	2100	4000						
Preço	129,0	165,0						