Universidade Federal do Paraná - Departamento de Estatística

CE071 - Análise de Regressão Linear

Prof. Cesar Augusto Taconeli

Lista de exercícios - diagnóstico e medidas corretivas

Nota: Os exercícios desta lista foram extraídos do livro **Linear models with R**, de Julian J. Faraway. Todas as bases de dados estão disponíveis no pacote `faraway` do `R`. Antes de iniciar qualquer análise, consulte a documentação da base, verifique o contexto dos dados e a descrição das variáveis. Adicionalmente, todas as questões devem ser precedidas por uma análise descritiva/exploratória, composta por gráficos e medidas descritivas pertinentes.

Exercício 1- Usando o conjunto de dados `sat`, ajuste um modelo de regressão linear múltipla com o escore SAT total como resposta e as variáveis `expend`, `salary`, `ratio` e `takers` como explicativas. Conduza o diagnóstico do ajuste e responda às questões apresentadas. Utilize os gráficos que julgar relevantes. Apresente possíveis melhorias ou medidas corretivas para o modelo, quando apropriado.

a) Cheque a suposição de variância constante para os erros;
b) Cheque a suposição de normalidade;
c) Cheque a existência de pontos de alavanca;
d) Cheque se há outliers;
e) Cheque a existência de pontos influentes;
f) Cheque a estrutura da relação entre a resposta e as variáveis explicativas.

Exercício 2- Semelhante ao exercício 1, mas desta vez para o conjunto de dados `prostate`, tomando a variável `lpsa` como resposta e as demais variáveis como explicativas.

Exercício 3- Semelhante ao exercício 1, mas desta vez para o conjunto de dados `happy`, tomando a variável `happy` como resposta e as demais variáveis como explicativas.

Exercício 4- Semelhante ao exercício 1, mas desta vez para o conjunto de dados `divusa`, tomando a variável `divorce` como resposta e as demais variáveis como explicativas (exceto `year`). Adicionalmente:

a) Análise possível autocorrelação temporal nos resíduos;
b) Obtenha os valores dos VIFs e analise as correlações entre as variáveis. Há evidência de que colinearidade é a justificativa para algumas variáveis explicativas serem não significativas? Justifique.
c) Investigue se a remoção de variáveis explicativas não significativas do modelo reduz a colinearidade.

Exercício 5- Considere o conjunto de dados `fat` e o ajuste do modelo de regressão linear múltipla da variável resposta `brozek` em função de `age`, `weight`, `height`, `neck`, `chest`, `hip`, `abdom`, `thigh`, `ankle`, `biceps`, `forearm` e `wrist`.

a) Investigue possível colinearidade nos dados;
b) As observações 39 e 42 são atípicas. Reajuste o modelo sem essas observações. Comente as diferenças observadas para o modelo ajustado com todos os dados;
c) Ajuste o modelo com `brozek` como resposta e `age`, `weight` e `height` como preditores. Faça a análise de colinearidade e compare os resultados aos produzidos pelo modelo original;
d) Obtenha o intervalo de predição 95% para `brozek` considerando os valores medianos para `age`, `weight` e `height`;
e) Obtenha o intervalo de predição 95% para `brozek` para `age=40`, `weight=200` e `height=73`. Como esse intervalo se compara ao obtido no item anterior?

f) Obtenha o intervalo de predição 95% para `brozek` para `age=40`, `weight=130` e `height=73`. Esses valores para os preditores são atípicos? Compare os intervalo obtido aos produzidos nos itens anteriores.

Exercício 6- Use o conjunto de dados `cheddar` para essa questão.

a) Ajuste o modelo de regressão linear múltipla com `taste` como resposta e as outras três variáveis como explicativas. Analise os resíduos parciais e verifique a necessidade de transformação nas variáveis explicativas;

b) use o método de Box-Cox para identificar uma transformação ótima para a resposta. Seria razoável manter a resposta na escala original?

c) Ajuste novamente o modelo do item (a), agora com a resposta transformada. Analise novamente os resíduos parciais. Houve alguma mudança nos resultados?

Exercício 7- Pesquisadores do NIST (National Insitutes of Standards and Technology) coletaram dados referentes à profundidade de defeitos avaliados em diferentes pontos de um oleoduto no Alasca primeiramente no campo e, num segundo momento, num laboratório. As medidas de profundidade obtidas no laboratório são mais acuradas que aquelas obtidas no campo, mas são mais caras e demandam mais tempo. Deseja-se ajustar uma equação de regressão para corrigir as medidas obtidas no campo. Os dados estão disponíveis no conjunto de dados `pipeline` do pacote 'faraway.

a) Ajuste um modelo de regressão para `Lab~Campo`. Cheque a homogeneidade das variâncias;
b) Vamos agora usar pesos para acomodar a variância não constante. Divida os dados originais em 12 grupos de tamanho nove (exceto para o oitavo grupo, que terá apenas oito observações), com base nos valores ordenados das medidas tomadas no campo. Dentro de cada grupo, calcule a média da medida de campo (`medcampo`) e a variância das medidas tomadas no laboratório (`varlab`).

Suponha que a variância da resposta seja relacionada à medida de campo da seguinte forma:

$$\text{Var(Lab)} = \alpha_0 \text{Campo}^{\alpha_1} \tag{1}$$

Ajuste a regressão de log(varlab) em função de log(medcampo). Use os resultados desse ajuste para determinar pesos apropriados para estimação por mínimos quadrados ponderados no ajuste de Lab~Campo. Você deverá usar como pesos os inversos dos valores ajustados para `varlab`. Analise o sumário da regressão, e compare com o sumário do ajuste obtido no item anterior, em que não se usa ponderação.

c) Uma alternativa à ponderação é transformar os dados. Encontre a transformação de Lab e/ou Campo tal que na escala transformada a relação seja linear com variância aproximadamente constante. Você pode restringir sua escolha às transformações raiz quadrada, logarítmica e inversa.

Exercício 8- Usando o conjunto de dados `stackloss`, ajuste o modelo de regressão linear múltipla para `stack.loss` em função das demais variáveis usando os seguintes métodos:

a) Mínimos quadrados;
b) Least absolute deviations;
c) Huber method;
d) Bisquare method.

Compare os resultados. Voltando ao ajuste por mínimos quadrados, faça o diagnóstico do ajuste e detect outliers e pontos influentes. Remova essas observações e use novamente mínimos quadrados. Compare os resultados.