Departamento de Matemática

Licenciatura em Estatística Aplicada

Modelos Lineares e Aplicações
Trabalho de Análise/Modelação de Dados

Docente: Raquel Menezes

Abril-Maio 2022

1 Objectivos

Comprovar que o aluno é capaz de realizar a sequência de análise habitual em estudos de **Regressão Linear Múltipla**, utilizando as ferramentas informáticas mostradas durante o curso. O aluno deverá ser capaz de avaliar críticamente os resultados obtidos. Também se verificará a capacidade de utilizar as diversas funções disponíveis em ambiente R para regressão linear.

O objetivo principal é encontrar o modelo adequado com o menor número de variáveis que descreva adequadamente a relação entre variáveis. Para tal, o aluno poderá recorrer a uma das técnicas apresentadas em sala de aula: *stepwise*, *backward* ou *forward*.

Nesta análise, deve-se detetar a existência de *outliers*, pontos com *leverages* elevados ou observações influentes; e verificar se a eliminação dessas observações é adequada. Por último, para o modelo eleito, deve-se proceder com a análise de resíduos.

2 Instruções detalhadas

Um grupo de trabalho será compostos por dois alunos, devendo cada grupo escolher uma base de dados, distinta dos restantes grupos. Sugerem-se as seguintes base de dados:

1. "caranguejos.txt", disponível na Blackboard (v.a. resposta "CL" e 6 possíveis covariáveis)
2. "rios.txt", disponível na Blackboard (v.a. resposta "Y=%nitrogénio" e 4 possíveis covariáveis)
3. "arvores.txt", disponível na Blackboard (v.a. resposta "altura" e 4 possíveis covariáveis)
4. "salinidade.txt", disponível na Blackboard (v.a. resposta "Y" e 3 possíveis covariáveis)
5. "swiss", automaticamente disponível em ambiente R (v.a. resposta "Fertility" e 4 possíveis covariáveis)
6. "xangai.txt", disponível na Blackboard (v.a. resposta "Prod" e 3 possíveis covariáveis)
7. "diabetes", disponível em ambiente R debaixo da *library* "faraway" (v.a. resposta "chol", i.e. colesterol, e escolher no máximo 4 possíveis covariáveis)
8. "diabetes", disponível em ambiente R debaixo da *library* "faraway" (v.a. resposta "stab.glu", i.e. glucose, e escolher no máximo 4 possíveis covariáveis)
9. "fat", disponível em ambiente R debaixo da *library* "faraway" (v.a. resposta "density", i.e. densidade de gordura, e escolher no máximo 4 possíveis covariáveis)

Note-se que um grupo poderá também propor uma base de dados diferente das anteriormente sugeridas (inclusivamente, a *library* "faraway" apresenta diversas alternativas). Neste caso, o grupo deverá confirmar com o docente de ML&A a adequabilidade dos dados escolhidos, antes de proceder com a sua análise e modelação.

3 Apresentação do trabalho e prazos

Está agendada para o dia **19 de Maio**, uma **apresentação oral informal** sobre a base de dados escolhida. Esta apresentação deverá incluir resultados de uma análise preliminar dos dados, e os primeiros resultados obtidos através da regressão linear (por exemplo, utilizando apenas uma covariável).

Relativamente à apresentação do trabalho final, sugere-se a preparação de um pequeno relatório (cerca de 10 páginas) em formato PDF, com gráficos, comentários e conclusões que considere oportunos para a compreensão do trabalho desenvolvido. Toda a informação relevante deve estar contida neste relatório. Adicionalmente, deverá também ser preparado um ficheiro TXT onde se incluam os comandos utilizados, que apenas será consultado em caso de dúvida na interpretação de algum resultado apresentado no relatório PDF.

O último dia de entrega do **relatório PDF** será **25 de maio**, podendo o trabalho ser enviado por email para `rmenezes@math.uminho.pt`.