Estatística Aplicada

Estatística Aplicada em Ciências e Engenharia

Trabalho FCUP, 10 dez 2021

Instruções:

- O trabalho deve ser realizado num grupo constituído por **três estudantes**. Apenas excecionalmente, e só após autorização da Professora, será permitida uma resolução numa estrutura diferente.
- Cada grupo deve comunicar a sua constituição e o ficheiro de dados que pretende utilizar abrindo um tópico no fórum "Trabalho EA/EACE", da UC, no Moodle. **Não devem existir dois grupos diferentes a tratar o mesmo problema.**
- O trabalho escrito não deverá exceder 15 páginas e terá de ser submetido no Moodle até ao dia **2 de janeiro**, juntamente com o script em R e o ficheiro de dados correspondentes.
- O horário das apresentações orais será combinado posteriormente com os alunos.
- Considere um nível de significância de 0.05.
- No relatório, deverão apresentar e assinar a seguinte declaração, do código de ética e conduta académica da Universidade do Porto: "*Declaro que o presente relatório é de minha autoria e não foi utilizado previamente noutro curso ou unidade curricular, desta ou de outra instituição. As referências a outros autores (afirmações, ideias, pensamentos) respeitam escrupulosamente as regras da atribuição, e encontram-se devidamente indicadas no texto e nas referências bibliográficas, de acordo com as normas de referenciação. Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico*".

1. Encontre um conjunto de dados que se enquadre num contexto de regressão linear e que satisfaça as seguintes condições:
   - ter pelo menos 6 variáveis explicativas e não mais de 8
   - ter pelo menos 2 variáveis explicativas categóricas
   - ter pelo menos 1 variável explicativa categórica com mais de 2 categorias
   - ter pelo menos 2 variáveis explicativas contínuas.

   De forma a satisfazer os requisitos sobre o tipo de variáveis a incluir, pode categorizar variáveis contínuas do ficheiro de dados original bem como considerar subconconjuntos de dados.

   Deve indicar explicitamente a forma como obteve os dados. Pode utilizar um conjunto de dados próprio ou algum disponível na web; pode consultar, por exemplo:

   - http://www.cs.toronto.edu/~delve/data/datasets.html
   - http://archive.ics.uci.edu/ml/
   - http://www.umass.edu/statdata/statdata/index.html
   - http://libdatabase.uchc.edu/wang/search.asp
   - http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/DataSets
   - http://vincentarelbundock.github.io/Rdatasets/datasets.html
   - http://www.cs.waikato.ac.nz/ ml/weka/datasets.html
   - https://www.kaggle.com/datasets.html
   - https://www.rdocumentation.org/packages/AER/versions/1.2-9

   Construa um relatório final que inclua as respostas às questões seguintes.

   (a) Formule, de forma clara e precisa, o problema que se propõe resolver neste trabalho.

   (b) Efetue uma descrição estatística, numérica e gráfica, dos dados no contexto do problema formulado.

   (c) Efetue uma discussão devidamente fundamentada sobre a seleção do modelo final, incluindo uma análise à qualidade do ajustamento e à satisfação dos pressupostos do modelo.
   Nota: Só deve analisar, quanto à qualidade do ajustamento e satisfação dos pressupostos, o modelo final; não os modelos intermédios. O modelo final deverá ser o mais completo possível.

(d) Apresente a equação matemática do modelo final.

(e) Para uma variável contínua X_1 e uma variável categórica com mais de 2 categorias X_2 que constem do modelo final[1]:

(e.1) interprete o efeito bruto de X_1 e o efeito ajustado de X_2.

(e.2) determine, graficamente, bandas de confiança e de predição em função dos valores de X_1, fixando os restantes preditores contínuos nos seus valores medianos e os categóricos nas respetivas modas.

(e.3) interprete o efeito provocado na resposta por uma mudança da terceira categoria de X_2 para a segunda, e indique um intervalo de confiança a 95% para esse efeito.

(e.4) interprete o efeito provocado por um aumento em X_1 correspondente a dois desvios padrão dos seus valores.

(e.5) averigue a existência de uma interação significativa entre X_1 e X_2. Independentemente da sua significância estatística, interprete os efeitos estimados nessa interação.

[1] se o modelo final não contiver alguma das variáveis X_1 ou X_2, considere a sua inclusão no modelo, apenas para esta alínea