Trabalho final

Nessa avaliação vocês deverão modelar quatro conjuntos de dados distintos utilizando alguns dos modelos nas classes de regressão linear, regressão linear generalizada (incluindo modelos de quasiverossimilhança) e suas versões com efeitos aleatórios. É permitido considerar transformações das variáveis resposta em cada situação. Muito cuidado ao utilizar como referência artigos ou outros materiais de terceiros pela internet ou com LLMs. É bastante comum essas fontes tratarem observações discrepantes removendo-as ou substituindo-as por valores médios no conjunto de dados. Durante a correção, essa maneira de tratar observações discrepantes será bastante penalizada e pode prejudicar consideravelmente a nota final! Ênfase: em 99% dos casos é **muito** errado fazer isso! Vale a máxima de que não alteramos os dados para se ajustarem à modelagem, mas sim o modelo para se ajustar aos dados.

Vocês podem utilizar os conjuntos de dados enviados como arquivos .csv pelo Sigaa ou outras versões disponíveis em outras fontes, como no R, Python, Kaggle, e afins, sem problemas.

Embora não seja necessàrio para a nota nesse trabalho, sugiro apresentar uma breve análise exploratória dos dados para auxiliar a escolha do modelo a ser utilizado. A entrega pode ser feita como arquivo .R ou .py com o código comentando, arquivo .Rmd ou .ipynb ou similar, em formato .pdf como relatório (opcional) ou de qualquer forma que eu possa ler o código. Para cada modelo inclua uma breve interpretação de pelo menos um dos parâmetros estimados.

Lembre-se de sempre avaliar o fenômeno estudado e as relações entre as variáveis consideradas antes de inspecionar os dados, caso contrário pode ser introduzido viés na modelagem.

Qualidade do ar

O conjunto de dados disponível no arquivo airquality.csv apresenta informações sobre a qualidade do ar em um determinado local e determinado período. Na versão enviada pelo Sigaa há a concentração de ozônio no ar e esse é um indicativo de poluição. Modele a concentração de ozônio em função das outras variáveis disponíveis. Não é necessário utilizar todas as variáveis, embora seja possível. Caaso você use alguma versão desse conjunto de dados que não contenha a concentração de ozônio, indique qual poluente está sendo analisado.

Estudo sobre o sono

O conjunto de dados em sleep.csv traz o tempo de reação em milisegundos a um determinado estimulo em indivíduos submetidos a privação de sono. A variável days é codificada da seguinte maneira:

- Os valores 0 e 1 correspondem a dias de treinamento para a reação ao estimulo.
- O valor 2 representa o dia de referência, quando o indivíduo não foi submetido à privação de sono ainda.

- Os valores 3 ou mais representam a quantidade de dias de privação de sono (3 horas de sono por noite). Perceba que o 3 corresponde ao primeiro dia de privação de sono.

A principal questão a ser respondida é: "qual o efeito da privação de sono no tempo de reação dos indivíduos?".

Metanálise - viés de publicação

O arquivo metanalise.csv traz dados retirados de 100 artigos sobre a mudança comportamental de indivíduos em relação ao consumo de carne após serem informados sobre questões relacionadas ao bem-estar animal. Em cada artigo foi reportada uma medida de razão de chances em relação a mudança de comportamento após a intervenção. Por exemplo: "há um aumento de 10% nas chances de uma pessoa reduzir ou parar o consumo de carne após a intervenção". Essa razão de chances é, na maioria das vezes, estimada a partir de um modelo de regressão. Um erro-padrão deve ser reportado junto dessa estimativa a fim de verificar a significância estatística do efeito da intervenção. O arquivo .csv traz a estimativa do logaritmo da razão de chances (coluna yi) e o erro-padrão associado a esse número (coluna vi).

Dizemos que ocorre viés de publicação quando as publicações de uma determinada área tendem a apresentar apenas os estudos reportando grandes tamanhos de efeito. No caso desse conjunto de dados, uma preocupação é que apenas estudos que apresentem estimativas elevadas das razões de chances sejam publicados, inflando assim o efeito reportado pela comunidade de pesquisadores. Um indicativo de que isso ocorre pode ser obtido ao analisar a relação entre o tamanho do efeito estimado (yi) e seu erro-padrão (vi). Se houver evidência de que o erro-padrão influencie no tamanho do efeito, há indícios de viés de publicação. A intuição seria a seguinte: um erro-padrão maior sugere uma maior variabilidade do efeito (razão de chances) estimado. Com isso há uma chance de serem observados valores altos para o efeito por acaso.

Para avaliar se há indícios de viés de publicação nessa área, considere uma regressão linear entre as variáveis yi e vi. Qual sua conlusão?

Dados de uso do Instagram

O arquivo instagram.csv contém as informações sobre o uso do Instagram coletadas através do Google Forms. A variável resposta para essa análise é o tempo, em horas, de uso do aplicativo do Instagram em um dia. O conjunto de dados está apresentado da forma como foi recebido e precisa ser tratado antes da modelagem. Quais hipóteses iniciais sobre o uso do aplicativo você formulou e quais foram verificadas?