Segundo bimestre de 2025 TP 2: Modelos mixtos, splines penalizados y causalidad

Herramientas de Modelado Estadístico

Reglas del TP:

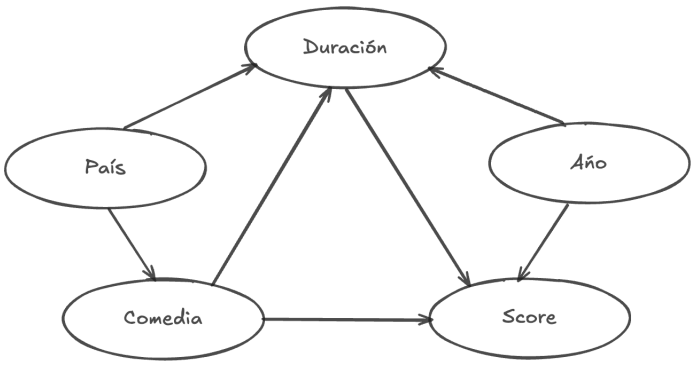
- Este trabajo debe hacerse de a dos o de a tres.
- Deben enviarse el script (un .Rmd o .R) y el informe que contenga las respuestas a todas las preguntas y los gráficos pedidos, en lo posible en .pdf. No hace falta explicar en el informe que es lo que hace cada una de las funciones del script.

En el archivo `titles_train.csv` se presentan 4000 títulos de una plataforma de streaming. El archivo `credits_train.csv` contiene los actores y directores para estas películas y series. La idea de este trabajo es poder predecir la calificación de IMDB a partir de otras covariables para cada título. Siempre, a lo largo de este trabajo, se va a considerar la pérdida cuadrática como forma de evaluar modelos.

1. (Opcional, no va a ser evaluado pero puede ser útil para los dos últimos items) Hacer un análisis exploratorio de estos datos. Algunas ideas:
   (a) ¿Hay algún género que parezca estar más asociado con el puntaje del título?
   (b) ¿Hay algún actor o director asociado con mayores o menores puntajes?
   (c) (Más complicada) Hay palabras de la descripción o del nombre del título que estén asociadas con un mayor/menor puntaje?
2. (a) Plantear un modelo de efectos fijos para predecir el puntaje de IMDB únicamente en función del país de origen.
   (b) Plantear un modelo de efectos aleatorios para predecir el puntaje de IMDB únicamente en función del país de origen.
   (c) Mostrar las estimaciones de los efectos de ambos modelos en un mismo gráfico e interpretar cómo se diferencian.

3. Usando únicamente la variable `release_year`, predecir la popularidad de cada título (usando un tipo de modelo que crea adecuado) con un spline cúbico. Usar $k = 1, 2, 3, 5, 10, 20, 50$ nodos fijando el λ (penalización de rugosidad) en 0, y comparar todas las curvas estimadas en un mismo gráfico.

4. Se tiene el siguiente DAG:



   donde `Comedia` es una variable binaria que indica si el género del título incluye comedia. A qué subconjunto de las variables `Año`, `Duración` y `País` de debe condicionar para estimar el efecto causal promedio de la variable `Comedia` sobre el `Score`? Dar todas las posibilidades.

5. Dividir al conjunto de datos en entrenamiento y testeo (tambien puede usar otra técnica, como validación cruzada). Con todas las variables que tiene disponibles, probar al menos 3 modelos diferentes y elegir el que minimice el error cuadrático médio de predicción para el rating de IMDB.

6. En los archivos `titles_test.csv` y `credits_test.csv` aparecen 1806 nuevos títulos, para los cuales no aparece el rating de IMDB (pero yo sí los tengo). A partir del modelo elegido en el item anterior, producir un archivo `predicciones.csv` que tenga una sola columna que contenga, en la fila i, la predicción del rating de IMDB para el título de la fila i. (tiene que tener 1806 filas). A partir de estas predicciones, yo voy a computar el error cuadrático medio de predicción. El equipo que tenga el menor error cuadrático medio gana el título de *Rey de la Estadística*.