TP 2 - Herramientas de Modelado Estadístico

Jésica Charaf e Ignacio Spiousas

4 de agosto de 2024

Modelos mixtos, Splines penalizados y Causalidad

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.

```
credits_train <- read_csv(here("modelado_estadístico/TP2/data/credits_train.csv"),
                          show_col_types = FALSE, col_select = -1)
```

```
## New names:
## * `` -> `...1`
```

```
titles_train <- read_csv(here("modelado_estadístico/TP2/data/titles_train.csv"),
                         show_col_types = FALSE, col_select = -1) |>
  mutate(genres = str_replace_all(genres, "\\[|\\]", ""),
         genres = str_replace_all(genres, "'", ""),
         production_countries = str_replace_all(production_countries, "\\[|\\]", ""),
         production_countries = str_replace_all(production_countries, "'", ""))
```

```
## New names:
## * `` -> `...1`
```

1. Hacer un análisis exploratorio de estos datos.

Lo primero que vamos a hacer es explorar cómo se relaciona el género de una película con su calificación en IMDB. Para esto vamos a calcular el puntaje promedio por género y mostrarlo en una gráfica de barras.

El principal problema que tienen nuestros datos para llevar adelante este tipo de análisis es que las películas pueden estar asociadas a más de un género. De momento lo vamos a resolver duplicando los títulos de películas y contando las calificaciones para cada uno de los géneros. Por ejemplo, si

una película es comedia y musical, su calificación será tenida en cuenta tanto al calcular el promedio de calificaciones del género comedia como el de musicales.

```
titles_train_by_genre <- titles_train |>
  separate_rows(genres, sep = ", ") |>
  filter(genres != "") |>
  drop_na(imdb_score)

head(titles_train_by_genre) |>
  dplyr::select(c("title", "genres")) |>
  knitr::kable()
```

title	genres
Monty Python and the Holy Grail	comedy
Monty Python and the Holy Grail	fantasy
Life of Brian	comedy
The Exorcist	horror
Dirty Harry	thriller
Dirty Harry	crime

Estos resultados los podemos ver en la Figura 1 como una gráfica de barras de acuerdo al promedio del género y con la información de la cantidad de películas que son clasificadas como pertenecientes a ese género (como n). Puede verse que los tres géneros mejor calificados son Guerra, Historia y Documentales (en azul), mientras que géneros considerados menores, o menos prestigiosos, como Terror y Comedia, se encuentran muy por debajo (en verde).

Sin embargo, podemos ver que los géneros con más producciones son drama y, justamente, comedia. Entonces: ¿No es injusto que Western con 32 producciones esté por arriba de comedia con 1575? Esta idea la vamos a desarrollar más en detalle cuando veamos las calificaciones por director.

Algo que podemos investigar rápidamente es qué pasa si en lugar de sumar de igual forma las palículas que pertecen a más de un género, lo hacemos de forma proporcional, es decir, si una película tiene dos géneros asociados, su calificación sumará dividida por dos al calculo del promedio. Esto lo podemos ver en la Figura 2.a, donde el n ahora puede ser fraccional. De momento esta propuesta no parecería haber cambiado demasiado el orden de los géneros.

Un último paso es pesar la calificación de cada película por la cantidad de votos. Es decir, si una película j perteneciente al género gen tiene $n_{gen,j}$ calificaciones, su peso en el promedio pesado será $n_{gen,j}/n_{gen}$, donde $n_{gen} = \sum n_j$ es la cantidad de calificaciones totales para ese género. En este caso también pesaremos por la cantidad de géneros a los que pertenece la palícula. En este caso el n

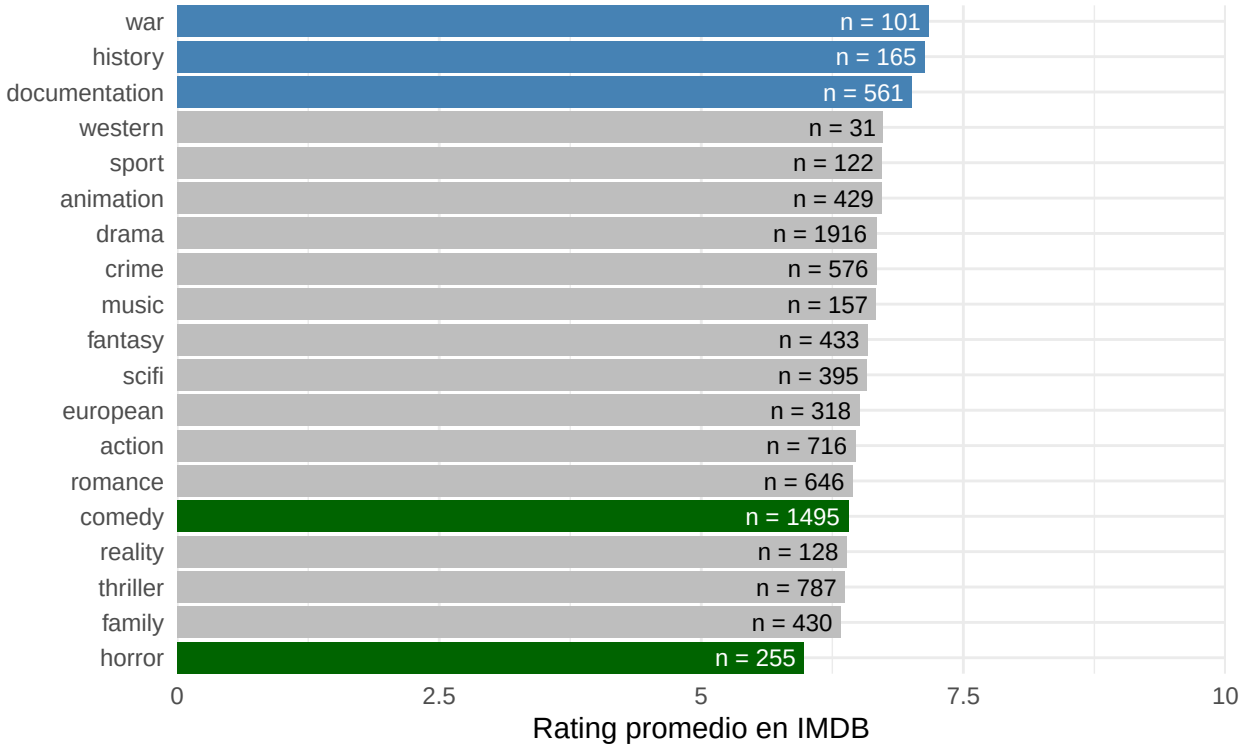


Figura 1: Relación entre el género de las películas y la calificación promedio en IMDB. n indica la cantidad de títulos perteneciente a cada género.

sigue representando la cantidad de títulos[1], mientras que el n_{gen} es la cantidad de votos recibidos para ese género. Los resultados de esta nueva forma de calcular la calificación promedio se muestra en el panel b de la Figura 2.

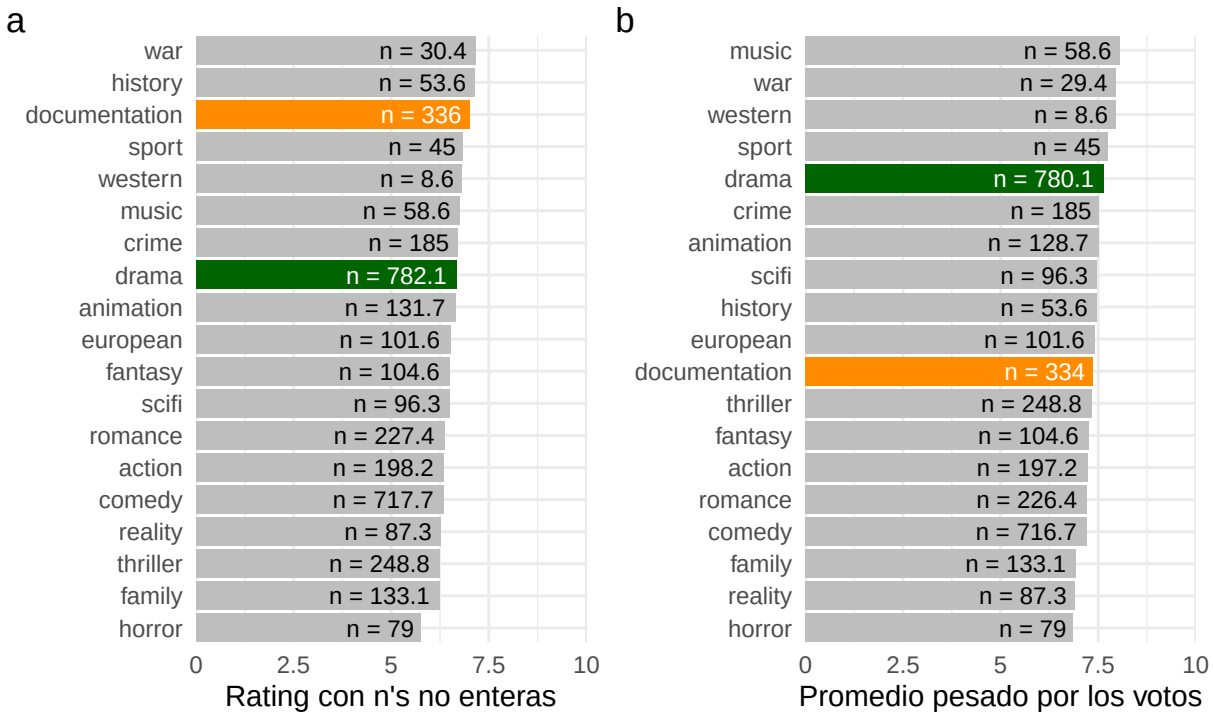


Figura 2: Relación entre el género de las películas y la calificación promedio en IMDB a) Teniendo en cuenta que hay títulos que son clasificados en más de un género y b) teniendo en cuenta que no todas las películas recibieron el mismo número de votos (promedio pesado por la cantidad de votos).

Vemos que esta nueva propuesta de cálculo del promedio sí cambia algunas cosas. Los ejemplos más claros son los géneros Documentales y drama que pasan de las posiciones 3 a la 11 y de la 8 a la 5, respectivamente. Una posible explicación para esto sería que el género drama tiene película con muchos votos y calificaciones altas lo que have que pesen más en el promedio pesado y mejoren la calificación promedio. Algo de esto puede verse en la Figura 3, donde vemos que el género Drama tiene más títulos en la esquina superior derecha de la figura (muchos votos y calificaciones altas).

De hecho, podemos ver que las 5 películas con más peso relativo en Drama tienen una calificación notablemente más alta que en Documentales.

[1]Puede parecer extraño que los valores de n difieran entre los paneles a y b de la Figura 2, pero esto se debe a que hay 11 películas que si tienen calificación pero no tienen cantidad de votos.

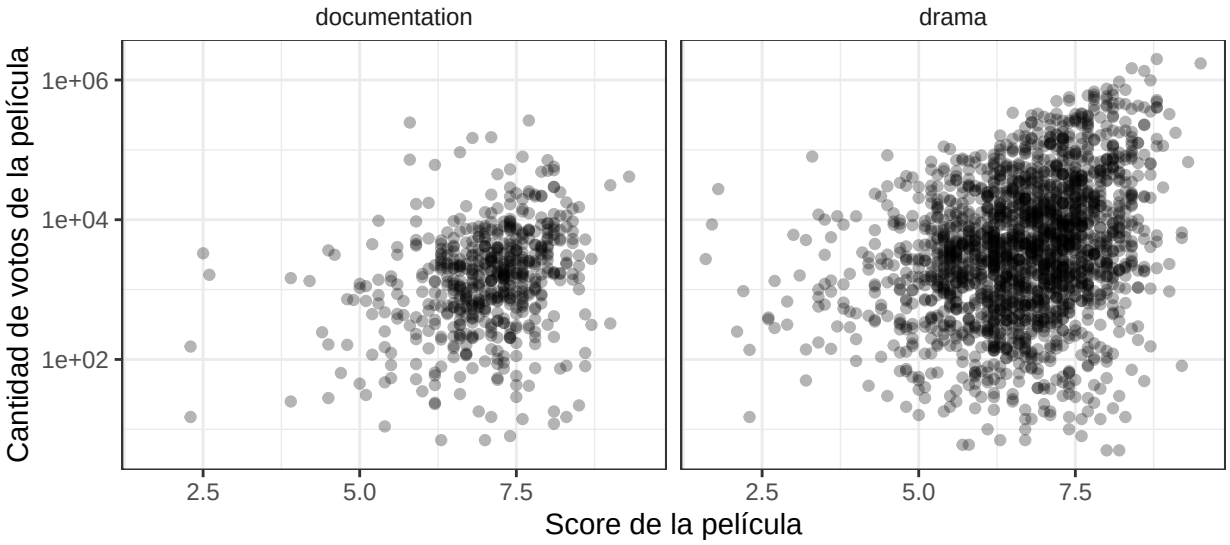
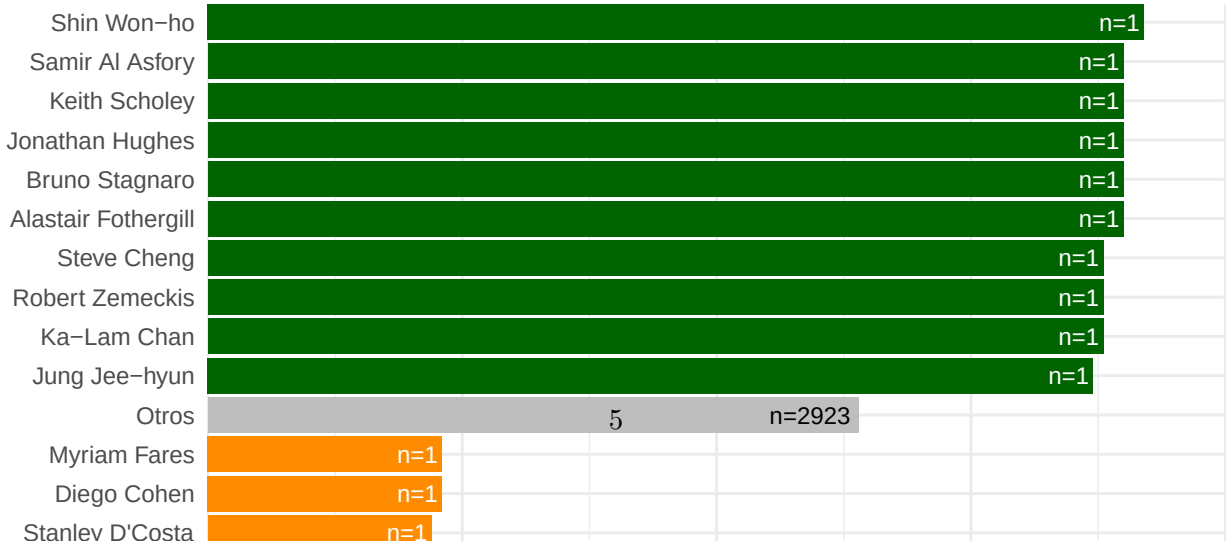


Figura 3: Cantidad ed votos versus calificación de la película para los títulos pertenecientes a los géneros Documentales y Drama.

title	weight	genres	imdb_score
Forrest Gump	0.0317042	drama	8.8
Breaking Bad	0.0274618	drama	9.5
Django Unchained	0.0234081	drama	8.4
Saving Private Ryan	0.0213950	drama	8.6
Stranger Things	0.0157216	drama	8.7
Road to Perdition	0.0812428	documentation	7.7
Battleship	0.0759326	documentation	5.8
The Gift	0.0468171	documentation	7.1
The Lake House	0.0456948	documentation	6.8
Jackass: The Movie	0.0284908	documentation	6.6

Ahora vamos a explorar la asocición de Director con el ranking promedio de la película.



Pareciera que hay algo raro, ¿No?. Tanto los directores mejor o peor rankeados tiene una sola película en el dataset. Esto si pensamos en promedios tiene sentido pero, ¿Cuán confiable es el promedio de una sola película? Por ejemplo, lo tenemos al queridísimo Bruno Stagnaro que, aunque no necesitamos de un modelo estadítico para entender que es un capo total, justifica su posición en el quinto lugar sólo con la calificación de la obra maestra que es Okupas. Este es un problema que en nuestra vida cotidiana a menudo enfrentamos y tenemos en cuenta. Por ejemplo, estás de vacaciones buscando un restaurant para almorzar en Google Maps y aparecen dos opciones: Uno con 5 estrellas y dos calificaciones y uno con 4.6 y cinco mil calificaciones. ¿Cuál elegirías? Probablemente el segundo, ¿No?

Ahora bien, ¿Cómo lidiamos con este problema? Vamos a explorar una alternativa que está íntimamente relacionada con la actualización bayesiana. Es decir, vamos a partir de una "creencia inicial" (R) con un deteminado peso (W) y a partir de eso actualizamos el rating de la siguiente forma:

$$R_i^b = \frac{R \times W + \sum_i^{n_j} r_{ji}}{W + n_i} = \frac{RW + \overline{r_i}}{W + n_i}$$

donde R_i^b es la calificación modificada del director i, r_{ji} es el calificación de la película j del diractor i y n_i es la cantidad de películas que tiene calificadas el director i. Esta nueva magnitud R_i^b podemos pensarla como si fuera el promedio pesado de W calificaciones R y las calificaciones de la película. De esta forma, va a "costar más" alejar el promedio de R y vamos a necesitar un n_i mayor para hacerlo.

Si pensamos en el ejemplo del resturante, este modelo está muy relacionado con el tipo de razonamiento que hacemos intuitivamente. Elegimos un restaurante con una calificación promedio más baja pero a la que le tenemos más confianza, haciendo una estimación interna de la incerteza de ese promedio.

Ahora viene la siguiente pregunta: ¿Cómo eligimos a R y W?. La elección de R podríamos hacerla de 3 formas: 1- $R = 5$, tomando el valor medio de nuestra escala de calificación como punto de partida; 2- $R = \overline{r_i}$, es decir, el promedio de los ratings individuales y; 3 - $R = Med(r_{ij})$, es decir, la mediana de los ratings individuales. Vamos a elegir este último valor ya que lo vamos a considerar como la "calificación más veces entregada". En cuanto a W, vamos a tomar un camino similar y calcularla como la mediana de los n_i.

De esta forma, R es igual a 6.5 y W es igual a 1. El hecho de que W sea igual a 1 puede ser problemático, aunque teniendo en cuenta que hay sólo 124 directores (de 2533) que tienen más de 2 calificaciones, no suena tan raro. Sin embargo, se trata del parámetro de este modelo más "difícil" de determinar ya que es el que nos dice cuál es el peso relativo de la evidencia de que la calificación de la película es R. Por eso, vamos a calcular el promedio para varios valores de W y así ver

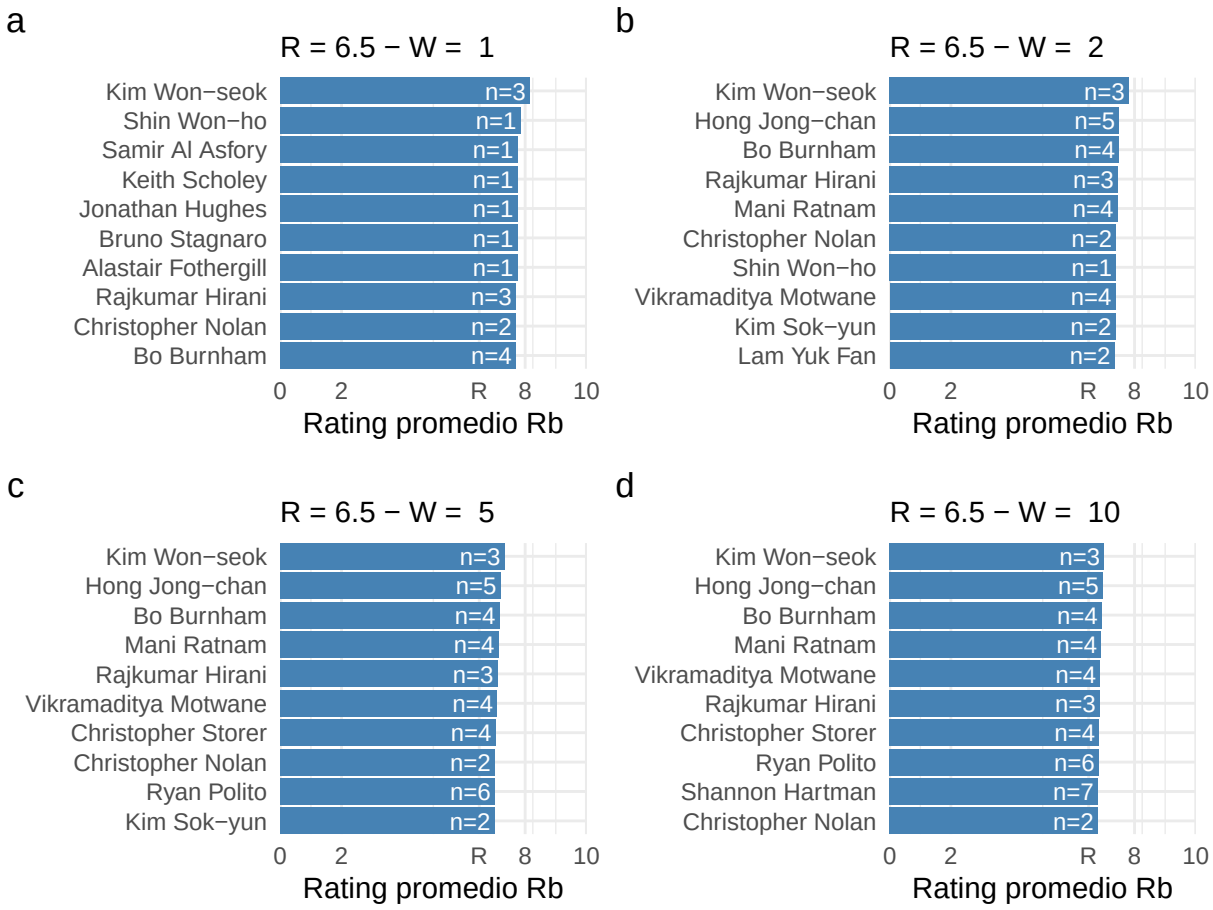

a
R = 6.5 − W = 1
Kim Won−seok n=3
Shin Won−ho n=1
Samir Al Asfory n=1
Keith Scholey n=1
Jonathan Hughes n=1
Bruno Stagnaro n=1
Alastair Fothergill n=1
Rajkumar Hirani n=3
Christopher Nolan n=2
Bo Burnham n=4
0 2 R 8 10
Rating promedio Rb
b
R = 6.5 − W = 2
Kim Won−seok n=3
Hong Jong−chan n=5
Bo Burnham n=4
Rajkumar Hirani n=3
Mani Ratnam n=4
Christopher Nolan n=2
Shin Won−ho n=1
Vikramaditya Motwane n=4
Kim Sok−yun n=2
Lam Yuk Fan n=2
0 2 R 8 10
Rating promedio Rb
c
R = 6.5 − W = 5
Kim Won−seok n=3
Hong Jong−chan n=5
Bo Burnham n=4
Mani Ratnam n=4
Rajkumar Hirani n=3
Vikramaditya Motwane n=4
Christopher Storer n=4
Christopher Nolan n=2
Ryan Polito n=6
Kim Sok−yun n=2
0 2 R 8 10
Rating promedio Rb
d
R = 6.5 − W = 10
Kim Won−seok n=3
Hong Jong−chan n=5
Bo Burnham n=4
Mani Ratnam n=4
Vikramaditya Motwane n=4
Rajkumar Hirani n=3
Christopher Storer n=4
Ryan Polito n=6
Shannon Hartman n=7
Christopher Nolan n=2
0 2 R 8 10
Rating promedio Rb

Figura 5: .

En la Figura 5 podemos ver los directores mejor rankeados para $R = 6.5$ y $W = 1$, 2, 5 y 10. Podemos ver que al incorporar este modelo con $W = 1$ ya aprecen en el "top ten" directores con $n > 1$, es decir, tener más películas mejor calificadas los aleja más de R. Como es de esperarse, al aumentar W cada vez hay directores con n más grande, pero también los R_i^b se comprimen alrededor de R. Esto último es esperable ya que le estamos dando, por ejemplo, un peso relativo de 10 películas a esa creencia inicial de 6.5. El director favorito indiscutido es Kim Won-seok, de nacionalidad coreana y famosos por dirigir telenovelas muy populares.

Se podría seguir explorando en la mejor forma de combinar toda esta información pero vamos a continuar con los modelos.

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.

Con los países de origen tenemos el mismo problema que con los géneros, hay películas que perteneces a más de un país, son una coproducción. De momento la solución propuesta va a ser duplicar las filas que tengan coproducción para ambos países. En este caso es menos influyente que en el caso de los géneros ya que pasamos de 4000 filas a 4470 filas.

Lo primero que vamos a hacer es ver la cantidad de producciones por país. En la figura 6 podemos ver el top 20 y como es de esperarse Estados Unidos lidera cómodamente este ranking, seguido de India, Gran Bretaña y Japón. Nuestro cine aparece en la posición 17, nada mal.

Una vez que tenemos este dataset vamos a ajustar dos modelos: 1- `fixed_countries`, un modelo de efectos fijos donde cada país tiene un asociado parámetro; y 2- `random_countries`, un modelo de efectos mixtos donde se ajusta un intercept y cada país tiene un intercept aleatorio.

```
fixed_countries <- lm(imdb_score ~ production_countries,
                      data = titles_train_by_country)
random_countries <- lmer(imdb_score ~ (1|production_countries),
                         data = titles_train_by_country)
```

Una vez que tenemos estos modelos ajustados vamos a generar predicciones para los países incluidos en el dataset de entrenamiento y graficarlos. En la Figura 7 podemos ver las predicciones de ambos modelos para cada país junto con el promedio de todos los scores del dataset como una línea punteada.

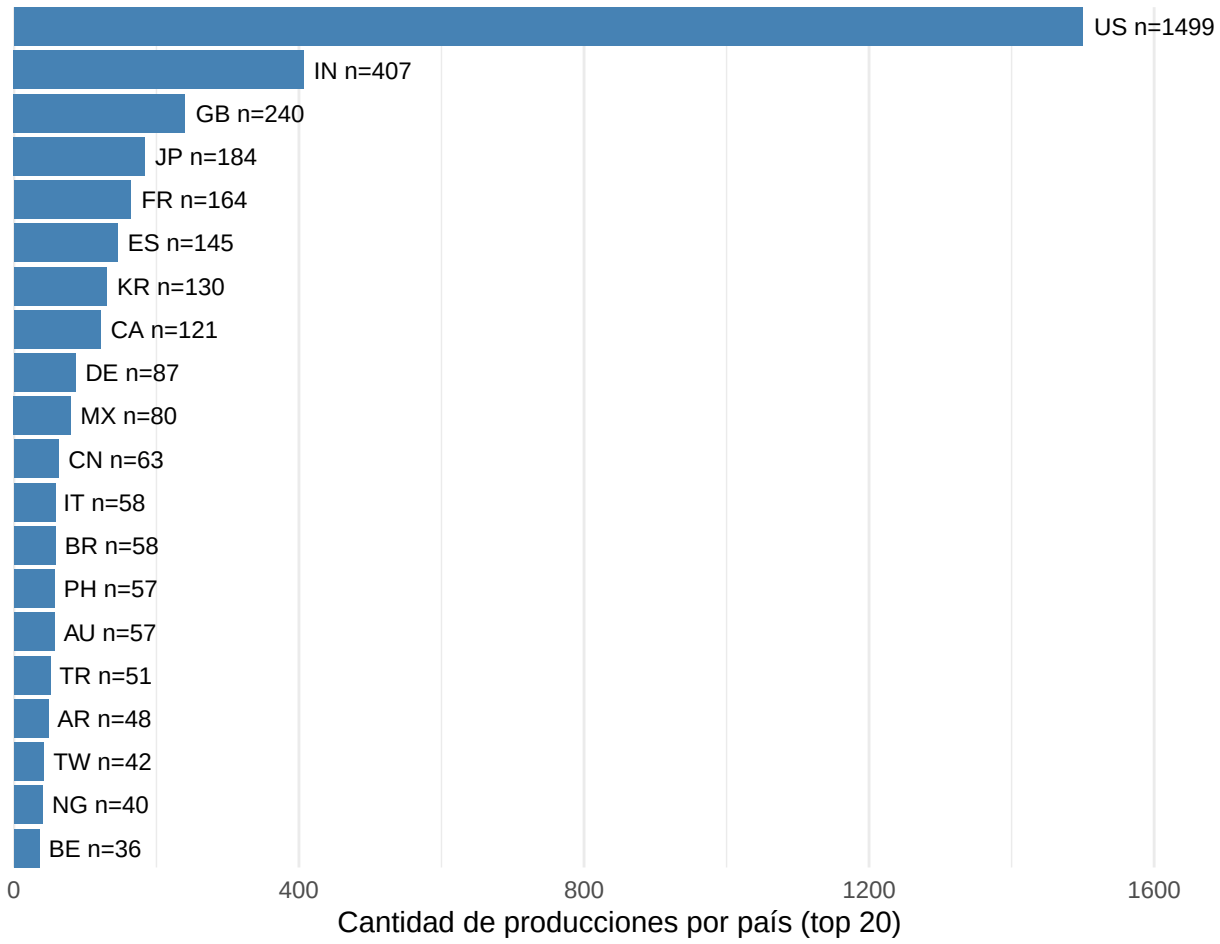


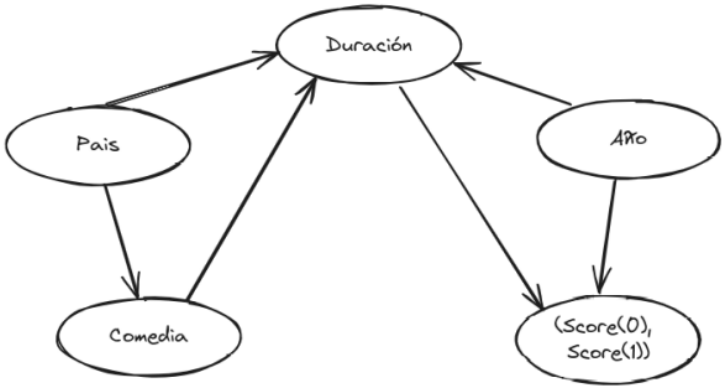
Figura 6: Cantidad de producciones por país para los 20 países con más producciones en el dataset de entrenamiento.

En la figura se ven varias cosas. La más notoria es que a medida que aumenta la cantidad de títulos por país (de arriba hacia abajo) hay más diferencia entre las predicciones de ambos modelos. Esto ocurre porque el modelo de efectos fijos predice a cada país como el promedio de los títulos del mismo mientras que el de efectos mixtos es una muestra de una distribución centrada en el promedio global (la línea punteada). Cuanto más grande es el número de títulos de un país más cerca estará su estimación del modelo de efectos aleatorios del promedio del mismo (que a su vez es la estimación de efectos fijos), mientras que si n es más chico, la estimación de efectos aleateatorios estará más cercana al promedio global que al promedio de ese país.

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 $\mathbf{k = 1, 2, 3, 5, 10, 20, 50}$ nodos fijando el λ (penalización de rugosidad) en 0, y comparar todas las curvas estimadas en un mismo gráfico.**

```
## New names:
## New names:
## New names:
## New names:
## * `var` -> `var...2`
## * `var` -> `var...3`
```

4. Se tiene el siguiente DAG:



donde `Comedia` es una variable binaria que indica si el género del título incluye comedia y `(Score(0), Score(1))` son los puntajes potenciales del título si no fuera y si fuera de comedia, respectivamente. ¿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.

Es decir, hallar los conjuntos Z tales que Comedia es independiente de `(Score(0),`

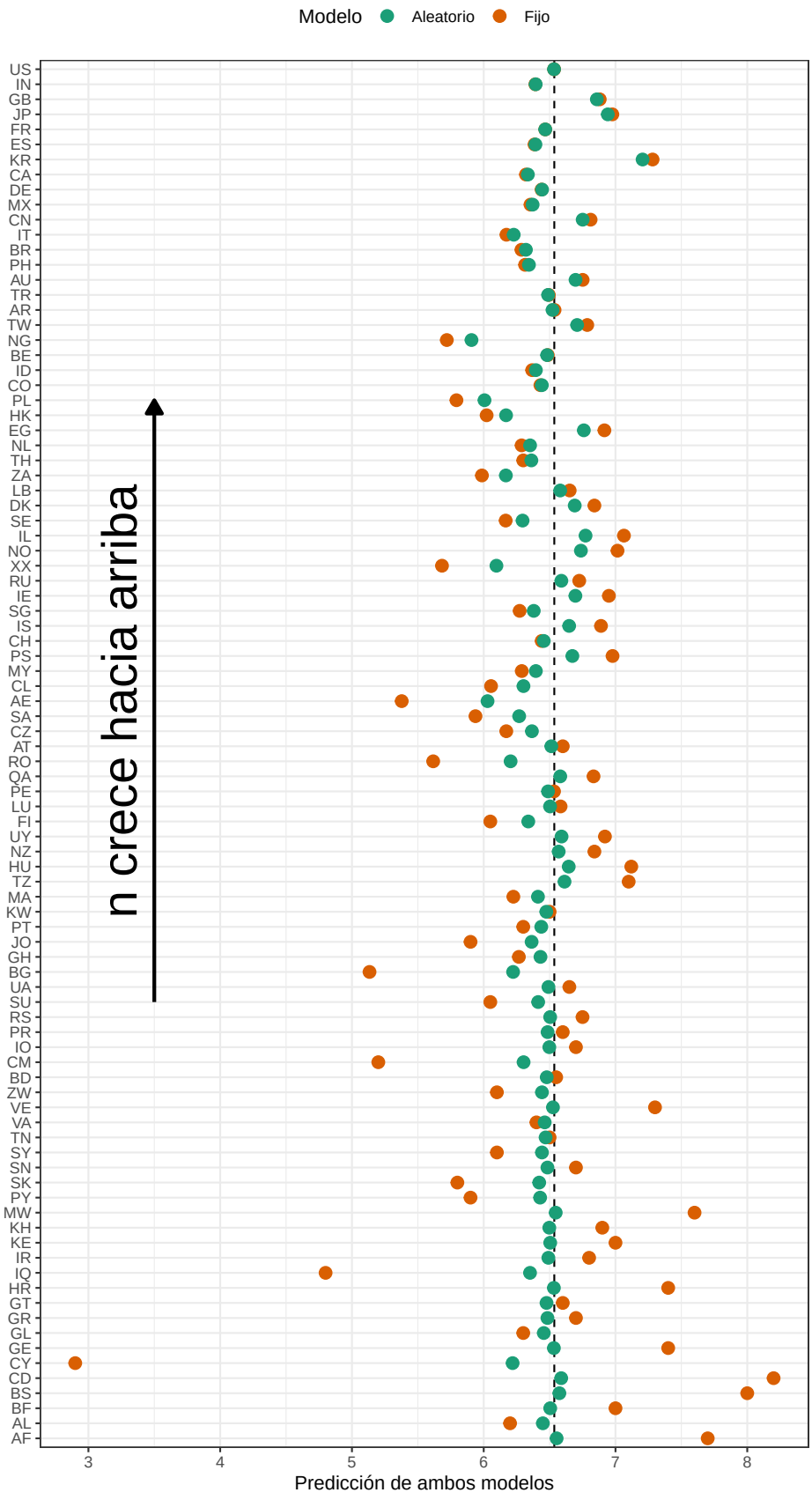


Figura 7: Predicciones de los modelos de efectos fijos y aleatorios para los títulos presentes en el dataset de entrenamiento. Los países esetán ordenados con cantidad de películas decreciente de arriba hacia abajo. La línea puntea vertical indica el promedio glbal de todas las calificaciones sin importar el país (promedio full pooleado).

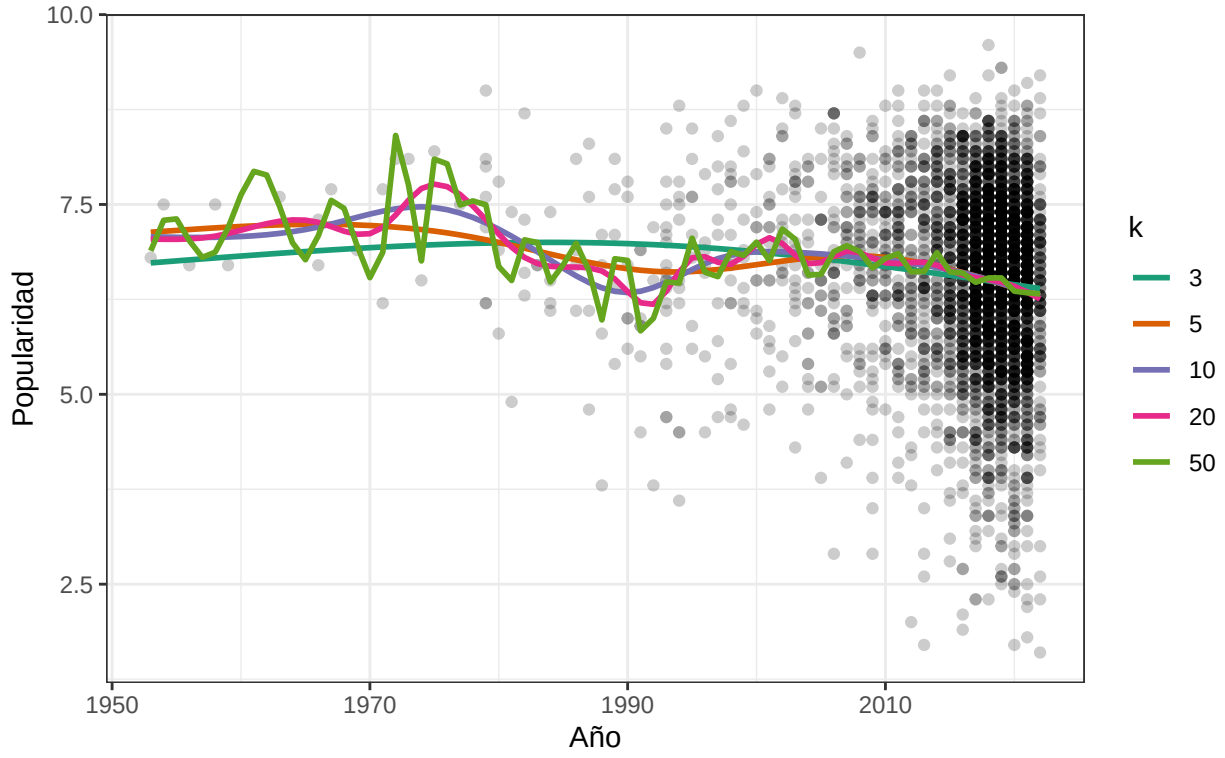


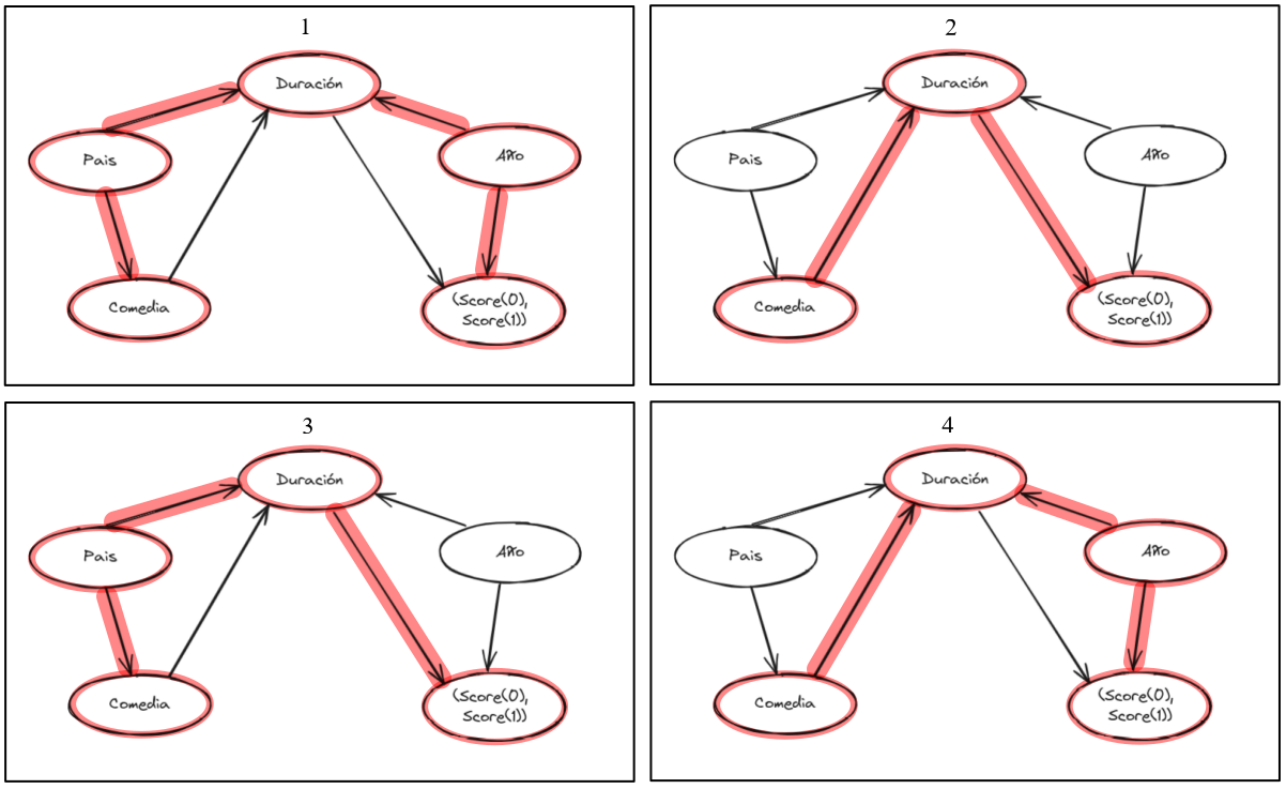
Figura 8: .

`Score(1))` condicional a Z.

En este caso, tenemos una variable binaria T que corresponde a la variable `Comedia` y nuestros *potencial outcomes* $(Y(0), Y(1))$ son los puntajes potenciales del título si no fuera y si fuera comedia, es decir, `(Score(0), Score(1))`. De esta manera, para estimar el efecto causal promedio de la variable `Comedia` sobre el `Score` nos interesa encontrar los conjuntos Z tales que

$$T \perp\!\!\!\perp (Y(0), Y(1))|_Z \, .$$

Para empezar, vamos a identificar todos los caminos que hay entre `Comedia` y `(Score(0), Score(1))`. En la siguiente imagen vemos los cuatro caminos resaltados en color rojo.



Luego, vamos a buscar los conjuntos de variables Z que aseguren que `Comedia` y `(Score(0), Score(1))` estén *d-separados*, basándonos en el siguiente teorema:

Teorema: *Si X e Y están d-separados por Z, entonces $X \perp\!\!\!\perp Y|_Z$.*

Recordemos que decimos que dos nodos X e Y de un DAG están *d-separados* por un conjunto de nodos Z si todos los caminos entre X e Y están bloqueados por Z.

De esta forma, analizaremos todos los subconjuntos de las variables `Año`, `Duración` y `País` para ver cuáles garantizan la d-separación entre `Comedia` y `(Score(0), Score(1))`.

- $Z = \emptyset$: No sirve para asegurar d-separación. Por ejemplo, falla el camino 3 que no está bloqueado ya que no tiene ningún *collider* y en Z no están ninguno de los nodos centrales de la *chain* (`País` $\rightarrow$ `Duración` $\rightarrow$ `(Score(0), Score(1))`) ni del *cofounder* (`Comedia` $\leftarrow$ `País` $\rightarrow$ `Duración`).
- $Z = \{$`País`$\}$: No sirve para asegurar d-separación. Falla el camino 2 dado que no tiene ningún *collider* y solo hay una *chain* (`Comedia` $\rightarrow$ `Duración` $\rightarrow$ `(Score(0), Score(1))`) cuyo nodo central no pertenece a Z.
- $Z = \{$`Duración`$\}$: No sirve para asegurar d-separación. El camino 1, por ejemplo, queda desbloqueado cuando agregamos `Duración` que es el nodo central del único *collider* (`País`$\rightarrow$ `Duración` $\leftarrow$ `Año`) y en Z no hay otros nodos que bloqueen el camino.

- Z = {`Año`}: No sirve para asegurar d-separación. Falla, por ejemplo, el camino 2 por los mismos motivos que mencionamos con Z = {`País`}.
- Z = {`País`, `Duración`}: No sirve para asegurar d-separación. El camino 4 no está bloqueado porque agregamos `Duración` que es el nodo central del único *collider* (`Comedia`→ `Duración` ← `(Score(0), Score(1))`) y en Z no está el nodo central del *cofounder* (`Duración` ← `Año` → `(Score(0), Score(1))`).
- Z = {`País`, `Año`}: No sirve para asegurar d-separación. Falla el camino 2 por los mismos motivos que mencionamos con Z = {`País`}.
- Z = {`Duración`, `Año`}: Sí sirve para asegurar la d-separación:
  - El camino 1 queda bloqueado ya que `Año` está en Z y es el nodo central de un *cofounder* (`Duración` ← `Año` → `(Score(0), Score(1))`).
  - El camino 2 está bloqueado ya que `Duración` pertenece a Z y es el nodo central de la *chain* (`Comedia` → `Duración` → `(Score(0), Score(1))`).
  - El camino 3 está bloqueado porque `Duración` pertenece a Z y es el nodo central de la *chain* (`País` → `Duración` → `(Score(0), Score(1))`).
  - El camino 4 queda bloqueado por el mismo motivo que el camino 1.
- Z = {`País`, `Duración`, `Año`}: Sí sirve para asegurar la d-separación, los motivos para argumentarlo son los mismos que en el conjunto anterior.

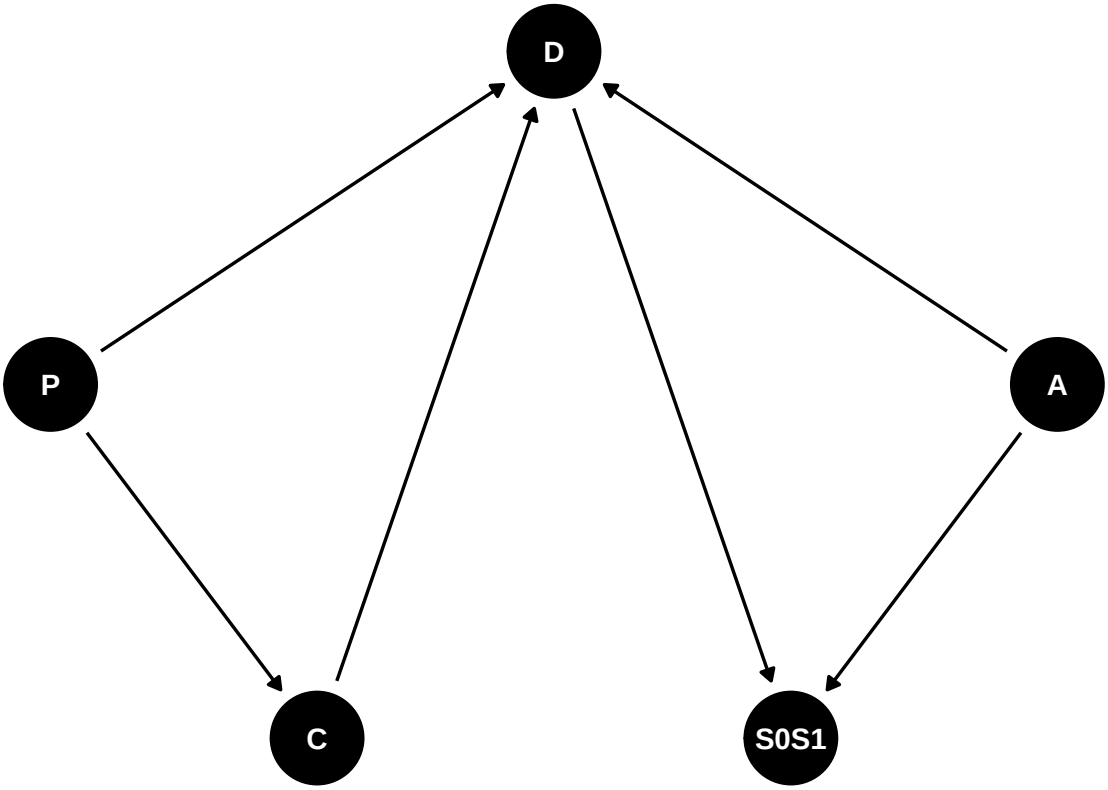
En conclusión, los posibles conjuntos Z a los cuales se debe condicionar de forma que `Comedia` y `(Score(0), Score(1))` resulten independientes son: Z = {`Duración`, `Año`} y Z = {`País`, `Duración`, `Año`}.

% COMENTARIO (BORRAR DESPUÉS): Dejo acá código para ver si el gráfico de caminos lo hacemos en R. No me salió colorear todos los caminos (el comando `ggdag_paths` solo te pone los que están abiertos) %

```
dag<- dagitty::dagitty('dag {
                       SOS1 [outcome,pos="0.4,-1.517"]
                       D [pos="0,0.631"]
                       A [pos="0.850,-0.411"]
                       P [pos="-0.850,-0.411"]
                       C [exposure,pos="-0.4,-1.517"]
                       P -> C
                       P -> D
                       A -> D
                       A -> SOS1
                       D -> SOS1
```

```
                      C -> D
                       }')


tidy_dag <- tidy_dagitty(dag)
ggdag(tidy_dag) +
  theme_dag()
```



5. Dividir al conjunto de datos en entrenamiento y testeo (también 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 medio 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 un premio sorpresa.