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 que socialmente suelen ser 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 películas que pertenecen a más de un género, lo hacemos de forma proporcional o pesada. Es decir, si una película tiene dos géneros asociados, su calificación sumará dividida por dos al cálculo del promedio (le damos un peso de $\frac{1}{2}$) y en el n total, en lugar de tomarla como 1 película, la sumaremos como $\frac{1}{2}$ de película. Esto lo podemos ver en el panel a. de la Figura 2, donde el n (que sería la suma de todos los pesos) ahora puede ser fraccional. En principio 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

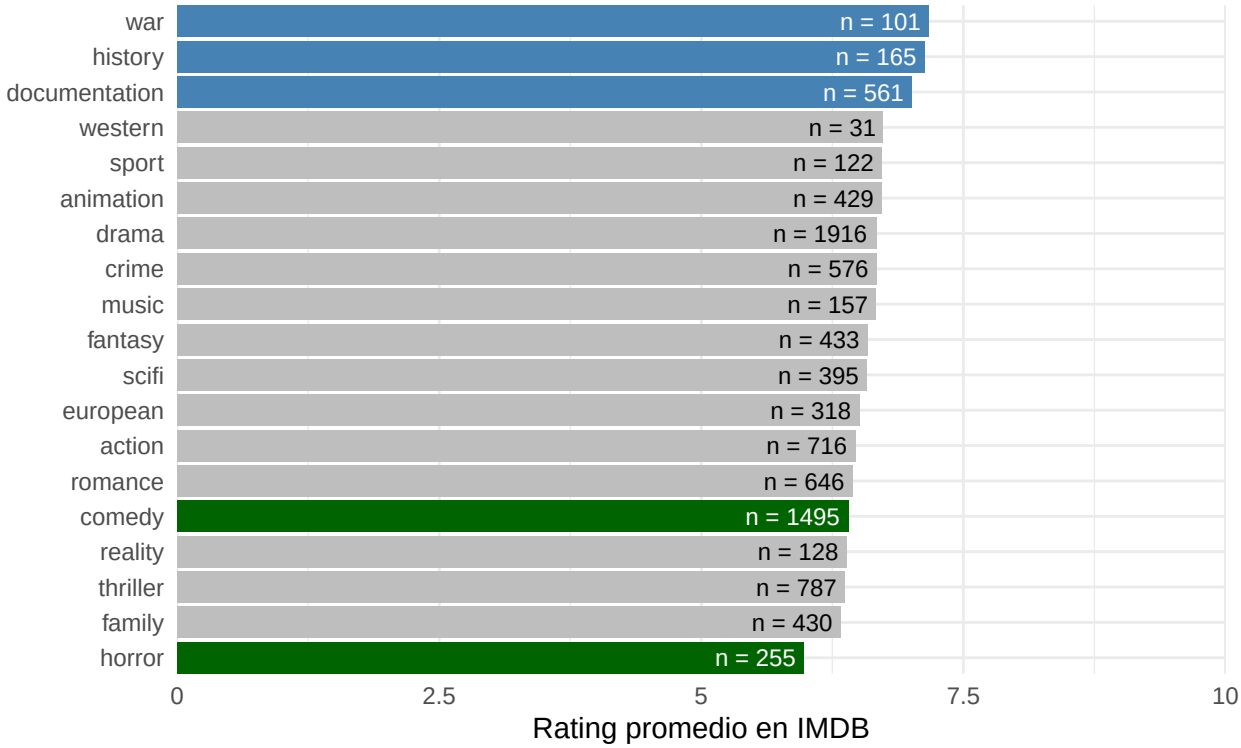


Figura 1: Relación entre el género de las películas y la calificación promedio en IMDB. El n indica la cantidad de títulos perteneciente a cada género.

sería $n_{gen,j}/n_{gen}$, donde $n_{gen} = \sum_j n_{gen,j}$ es la cantidad de calificaciones totales para ese género. Pero, en este caso, también pesaremos por la cantidad de géneros (G_j) a los que pertenece la película, multiplicando su puntaje por el siguiente peso:

$$\frac{\frac{1}{G_j} * n_{gen,j}}{\sum_i \frac{1}{G_i} * n_{gen,i}}.$$

El n sigue representando la cantidad de títulos[1] (teniendo en cuenta que las películas compartidas por varios géneros suman como fracción), 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 muestran en el panel b de la Figura 2.

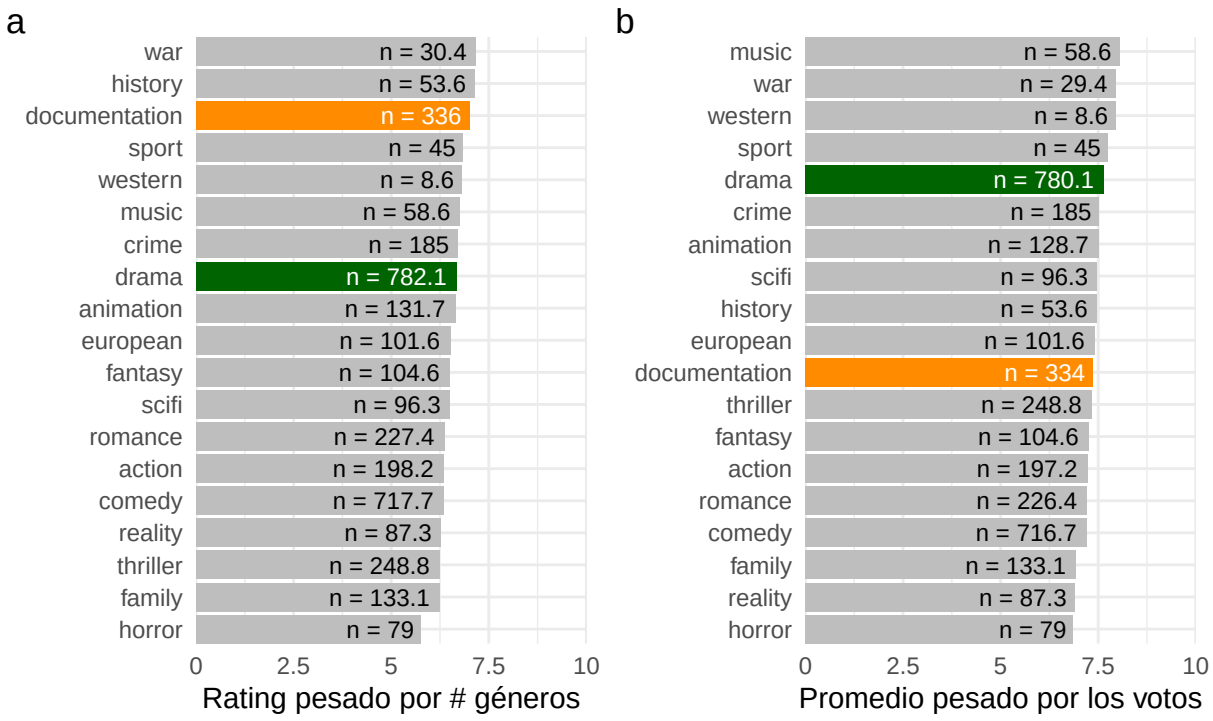


Figura 2: Relación entre el género de las películas y la calificación promedio en IMDB a) Teniendo en cuenta que hay títulos que son clasificados en más de un género y b) teniendo en cuenta que no todas las películas recibieron el mismo número de votos (promedio pesado por la cantidad de votos).

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 películas con muchos votos y calificaciones altas, lo que hace que pesen más en el promedio pesado y mejoren la

[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 sí tienen calificación pero no tienen cantidad de votos.

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).

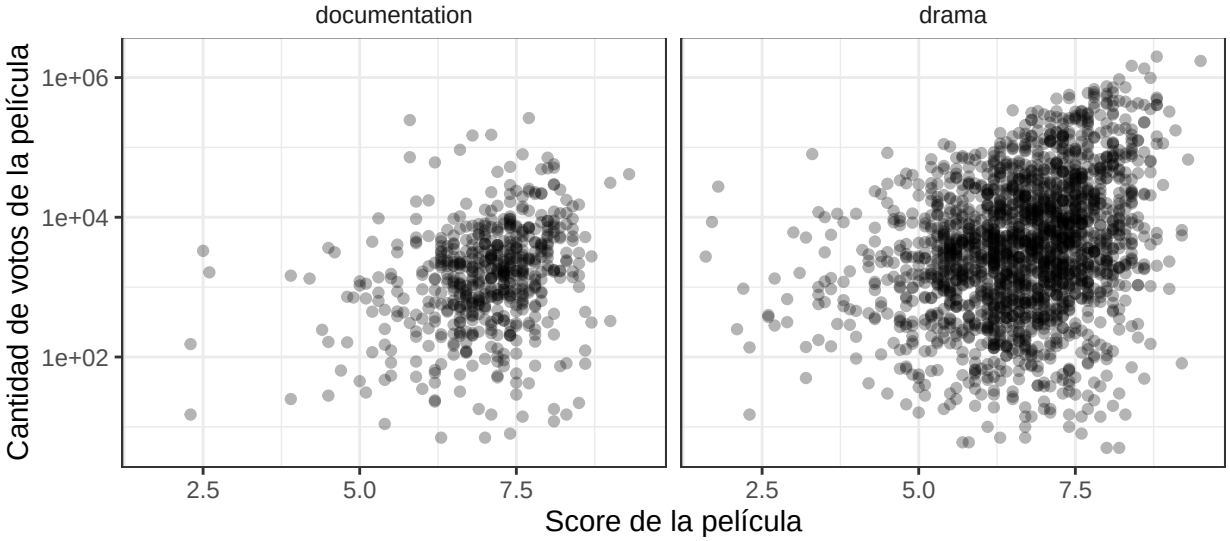


Figura 3: Cantidad de votos vs. calificación de la película para los títulos pertenecientes a los géneros Documentales y Drama.

Ahora vamos a explorar la asociación de Director con el rating promedio de la película. En la Figura 4 podemos ver un ranking del promedio de las calificaciones por director a partir de un gráfico de barras (en este caso se muestran solo los mejores puntuados y los peores, ya que hay 2533 directores), donde también indicamos con n la cantidad de películas que corresponden a cada director.

Se observan grandes diferencias en los scores promedios para los distintos directores, pero pareciera que hay algo raro, ¿no?. Tanto los directores mejor rankeados como los peores tienen 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 restaurante 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:

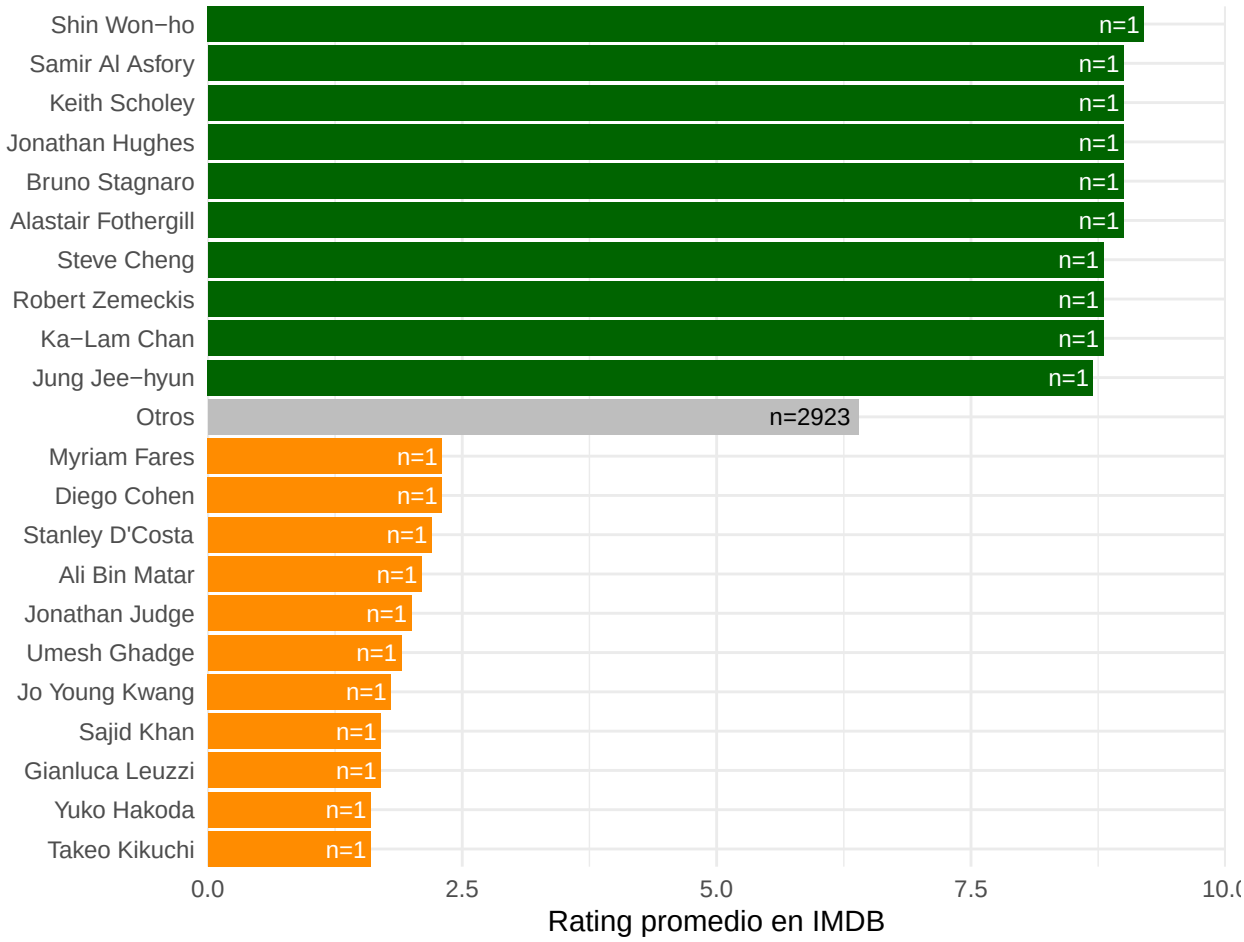


Figura 4: Relación entre el director de las películas y la calificación promedio en IMDB. El n indica la cantidad de títulos que pertenecen a cada director.

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

donde R_i^b es la calificación modificada del director i, r_{ji} es la calificación de la película j del director 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 las películas del director. 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 restaurante, 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 tres formas: 1) $R = 5$, tomando el valor medio de nuestra escala de calificación como punto de partida; 2) $R = \frac{1}{D} \sum_i \overline{r_i}$, es decir, el promedio de los ratings individuales de cada director $\overline{r_i}$ (donde D es la cantidad total de directores) y; 3) $R = Med_i(\overline{r_i})$, es decir, la mediana de los ratings individuales por director. 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 dos 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 rating promedio para varios valores de W y así ver cómo impacta en el ranking de directores.

En la Figura 5 podemos ver los directores mejor rankeados para $R = 6.5$ y $W = 1$, 2, 5 y 10. Podemos observar que al incorporar este modelo con $W = 1$ ya aparecen 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 famoso 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.

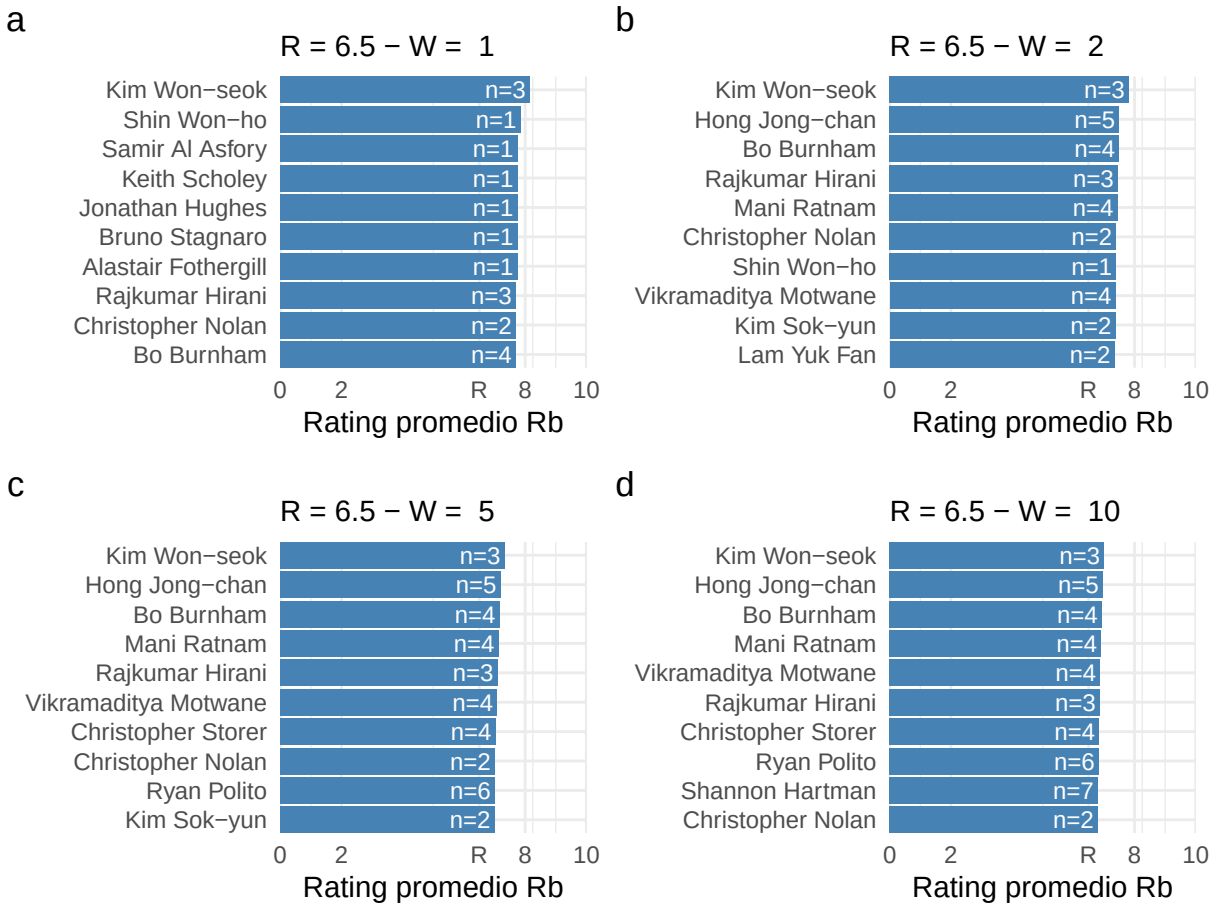


Figura 5: Relación entre el director de la película y la calificación promedio en IMDB considerando el rating actualizado Rb .

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 pertenecen a más de un país, es decir, son una coproducción. En este sentido, la solución propuesta va a ser duplicar las filas que tengan coproducción para cada país. En este caso es menos influyente que en el caso de los géneros ya que pasamos de 4000 filas a 4470 filas.

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

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.

A partir de este dataset vamos a ajustar dos modelos: 1- `fixed_countries`, un modelo de efectos fijos donde cada país tiene un parámetro asociado (β_j); y 2- `random_countries`, un modelo de efectos mixtos donde se ajusta un *intercept* global (β) y cada país tiene a la vez un *intercept* aleatorio ($\mu_j \sim \mathcal{N}(0, \sigma^2_{between})$), de modo que cada predicción para un país dado j será: $\hat{\beta} + \hat{\mu}_j$.

```
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 (que coinciden con las estimaciones de los efectos de cada modelo) y graficarlas. En la Figura 7 podemos ver las predicciones de ambos modelos para cada país junto con el promedio de todos los scores del conjunto de datos como una línea punteada.

En la figura se observan varias cosas. La más notoria es que a medida que disminuye 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 considera una muestra de una distribución centrada en el promedio global (la línea punteada). De esta forma, cuanto más grande es el número de títulos

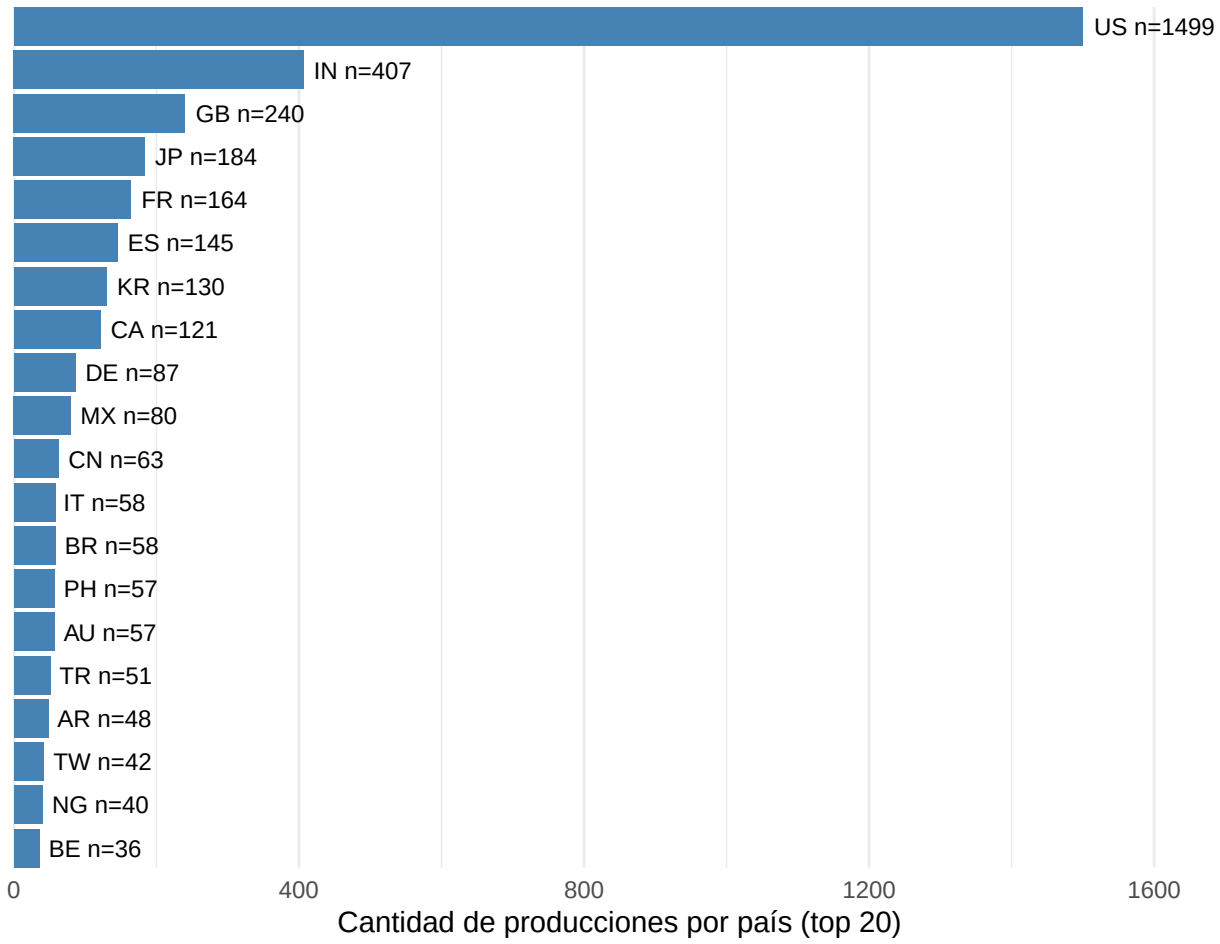


Figura 6: Cantidad de producciones por país para los 20 países con más producciones en el dataset de entrenamiento.

de un país más cerca estará la estimación del modelo de efectos aleatorios al promedio del mismo (que a su vez es la estimación de efectos fijos), mientras que si n es más chico, la estimación de efectos aleatorios 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.

A continuación vamos a ajustar un modelo con un spline cúbico variando la cantidad de nodos k que se utilizan y dejando fijo el λ de penalización en 0, con el objetivo de predecir la popularidad a partir de la variable `release_year`.

Para ello, vamos a considerar la cantidad de nodos sugeridos en la consigna a partir de $k = 3$ (es decir, tomamos $k = 3, 5, 10, 20, 50$) dado que con $k = 1$ y 2 no es posible realizar el ajuste por una cuestión de la dimensión del espacio de funciones que estamos utilizando. A su vez, para analizar las predicciones tomaremos una grilla de años entre el mínimo y el máximo de los valores que toma la variable `release_year` en nuestros datos.

```
train_data <- titles_train %>%
  dplyr::select(c("imdb_score", "release_year"))

# k=1 y 2 no deja ajustar, por la dimension del espacio
Ks <- c(3, 5, 10, 20, 50)

pred_k <- tibble(release_year=min(train_data$release_year):max(train_data$release_year))

# para cada K ajustamos el modelo y predecimos en una grilla
for (K in Ks) {
  fit <- gam(imdb_score ~ s(release_year, k = K , sp = 0, bs = 'cr'), data = train_data)
  pred <- predict(fit,
newdata = tibble(release_year = min(train_data$release_year):max(train_data$release_year)))
  var <- paste0("pred_k", K)
  pred_k <- pred_k %>%
    bind_cols(tibble(var = pred))
}
```

```
## New names:
## New names:
## New names:
## New names:
```

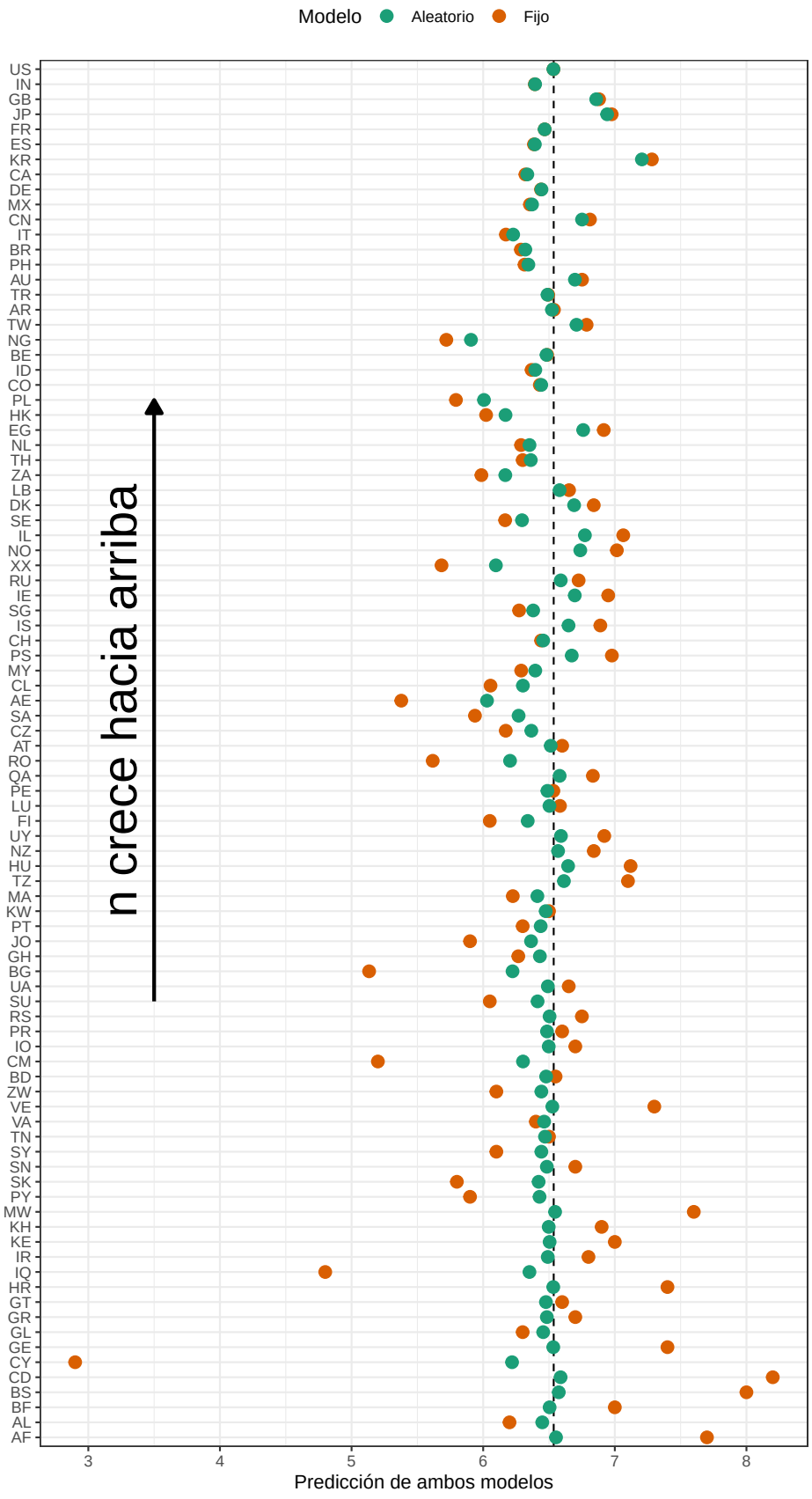


Figura 7: Predicciones de los modelos de efectos fijos y aleatorios para los países presentes en el dataset de entrenamiento (ordenados según cantidad de películas decreciente de arriba hacia abajo). La línea punteada vertical indica el promedio global de todas las calificaciones sin importar el país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

```
## * `var` -> `var...2`
## * `var` -> `var...3`
```

En la Figura 8 podemos ver en distintos colores las curvas obtenidas al predecir el score en función del año de lanzamiento, junto con las observaciones de nuestro conjunto de datos (en grises).

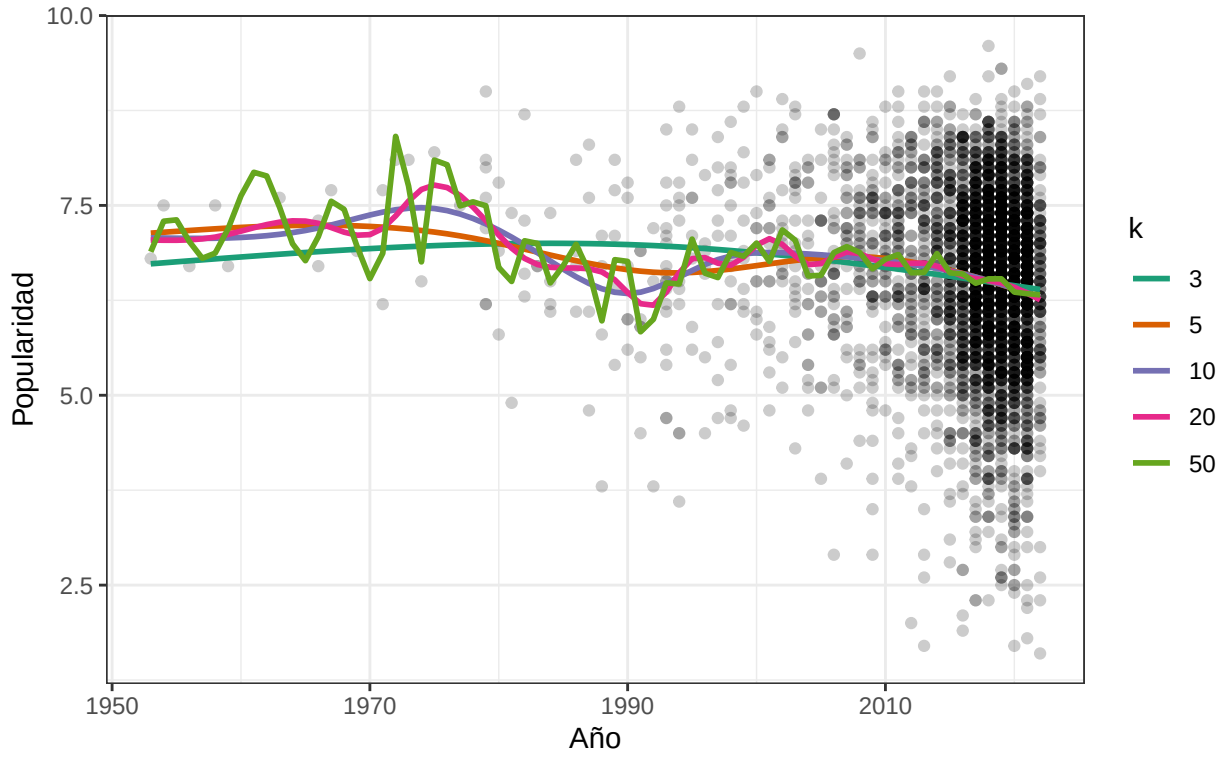


Figura 8: Curvas estimadas para la popularidad en función del año de lanzamiento, según la cantidad de nodos utilizados en el spline cúbico.

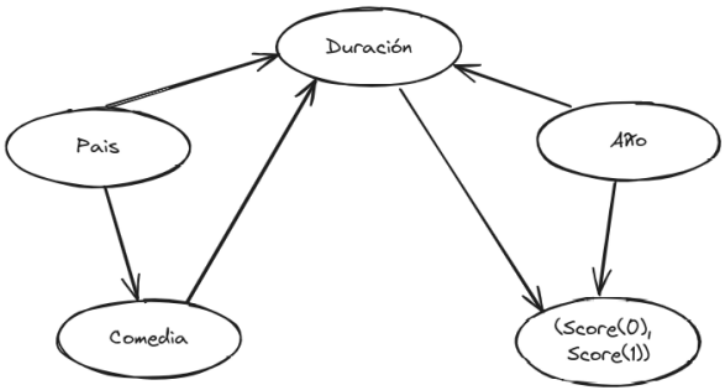
Lo primero que podemos observar es que la gran mayoría de los datos se encuentran en los años más recientes (post 2010) y que en cada uno de esos años hay mucha variabilidad en la popularidad de los títulos. Es decir, para un mismo año tenemos observaciones con puntajes muy diversos, por lo que, a simple vista, la variable `release_year` no parece ser muy buena para explicar la popularidad de una película.

Siguiendo con el análisis de las predicciones, en la figura vemos que a medida que aumentamos la cantidad de nodos las curvas parecen tener más fluctuaciones y se sobreajustan más a los datos que con los valores de k más chicos. De esta manera, para valores de k más grandes el modelo tiene mayor flexibilidad y, por lo tanto, las estimaciones resultarán más variables según el conjunto que se utilice para realizar el ajuste. En cambio, para valores de k más chicos las curvas estimadas tendrán una mayor estabilidad con la contraparte de aumentar el sesgo.

Por último, si miramos los lanzamientos más recientes (el último tramo de las curvas) podemos

visualizar que en todas las estimaciones hay un comportamiento levemente descendiente de los puntajes.

4. Se tiene el siguiente DAG:



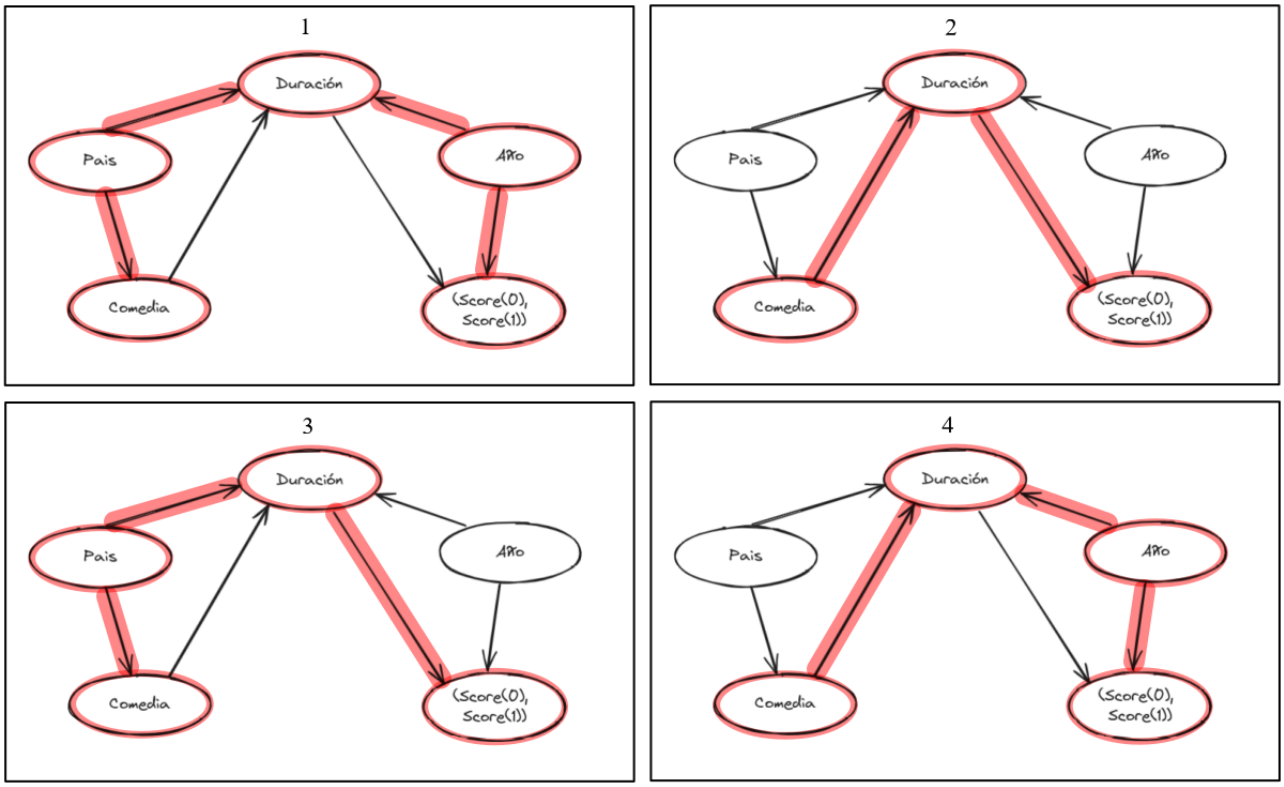
donde `Comedia` es una variable binaria que indica si el género del título incluye comedia y `(Score(0), Score(1))` son los puntajes potenciales del título si no fuera y si fuera de comedia, respectivamente. ¿A qué subconjunto de las variables `Año`, `Duración` y `País` de debe condicionar para estimar 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), 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`$\rightarrow$ `Duración` $\leftarrow$ `(Score(0), Score(1))`) y en Z no está el nodo central del *cofounder* (`Duración` $\leftarrow$ `Año` $\rightarrow$ `(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` $\leftarrow$ `Año` $\rightarrow$ `(Score(0), Score(1))`).
  - El camino 2 está bloqueado ya que `Duración` pertenece a Z y es el nodo central de la *chain* (`Comedia` $\rightarrow$ `Duración` $\rightarrow$ `(Score(0), Score(1))`).
  - El camino 3 está bloqueado porque `Duración` pertenece a Z y es el nodo central de la *chain* (`País` $\rightarrow$ `Duración` $\rightarrow$ `(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`$\}$.

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.