SOC3070 Análisis de Datos Categóricos

Trabajo final

Ponderación: 30% de la nota final del curso. Entrega: hasta el día 15 de Diciembre a las 23:59 p.m.

BONUS: Responder la pregunta *bonus* NO es un requisito necesario para obtener puntaje completo. Responder incorrectamente la pregunta *bonus* no afectará negativamente la nota obtenida, pero responderla correctamente mejorará la nota obtenida en un máximo de 0.5 puntos (o en la cantidad necesaria para obtener nota máxima si la nota original fuera superior a 6.5)

Descripción

En el trabajo final del curso trabajarán con el artículo "Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the "Many Analysts, One Data Set" Project" (Auspurg and Brüderl 2021), que revisita la hallazgos del artículo "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results" (Silberzahn et at 2018). Ambos artículos están disponibles en el repositorio del curso: Auspurg and Brüderl 2021 y Silberzahn et at 2018

Datos

Ambos estudio utilizan una misma base de datos, también disponible en el repositorio: Datos

Se obtuvieron datos generados por una empresa de estadísticas deportivas y fotos de perfil de todos los futbolistas (N = 2053) que jugaron en las primeras divisiones masculinas de Inglaterra, Alemania, Francia y España en la temporada 2012-2013. Además, estos datos incluyen un registro de todos los árbitros (N = 3147) que dirigieron partidos en los que estos jugadores participaron durante su carrera profesional. A partir de estos registros se creó una base de datos que contiene díadas (pares) de árbitro-jugador y atributos relevantes de dichas díadas.

El tono de piel de los jugadores fue codificado por dos evaluadores independientes. Basándose en su foto de perfil. Estos evaluadores categorizaron a los jugadores en una escala de 5 puntos que van desde "piel muy clara" (0) a "piel muy oscura" (1), con "piel ni oscura ni clara" como valor central (0.5).

Además, se calcularon escalas que mide sesgo racial implícito y explícito en el país de cada árbitro. Valores más altos en ambas escalas indica preferencia por jugadores blancos frente a jugadores negros. Ambas medidas se crearon agregando datos obtenidos online a partir de tests realizados a usuarios pertenecientes a los países de los árbitros.

La base de datos datos tiene un total de 146028 díadas de jugadores-árbitros. La siguiente lista describe en detalle las variables de la base de datos:

- `playerShort`: identificador único del jugador
- `player`: nombre del jugador
- `club`: equipo del jugador

- `leagueCountry`: país del equipo del jugador (England, Germany, France, Spain)
- `birthday`: fecha de nacimiento del jugador
- `height`: altura del jugador (en cm)
- `weight`: peso del jugador (en kg)
- `position`: posición en la que juega el jugador
- `games`: número de partidos en los que participó la diada jugador-árbitro
- `victories`: número de victorias en los que participó la diada jugador-árbitro
- `ties`: número de empates en los que participó la diada jugador-árbitro
- `defeats`: número de derrotas en los que participó la diada jugador-árbitro
- `goals`: número goles marcado por el jugador en la diada jugador-árbitro
- `yellowCards`: número de tarjetas amarillas que el árbitro en la diada jugador-árbitro le da al jugador.
- `yellowReds`: número de tarjetas amarillas o rojas que el árbitro en la diada jugador-árbitro le da al jugador.
- `redCards`: número de tarjetas rojas que el árbitro en la diada jugador-árbitro le da al jugador.
- `rater1`: evaluación del color de piel del jugador, evaluador 1
- `rater2`: evaluación del color de piel del jugador, evaluador 2
- `refNum`: identificador único del árbitro (anonomizado por privacidad)
- `refCountry`: identificador único del país del árbitro (anonomizado por privacidad)
- `meanIAT`: promedio de sesgo racial implícito en el país del árbitro
- `nIAT`: tamaño muestral en test de sesgo racial implícito en el país del árbitro
- `seIAT`: error estándar de promedio de sesgo racial implícito en el país del árbitro
- `meanExp`: promedio de sesgo racial explícito en el país del árbitro
- `nExp`: tamaño muestral en test de sesgo racial explícito en el país del árbitro
- `seExp`: error estándar de promedio de sesgo racial explícito en el país del árbitro

Instrucciones:

1. Leer "Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the "Many Analysts, One Data Set" Project" (Auspurg and Brüderl 2021) – AB2021 de aquí en adelante.
2. Usa el siguiente código para importar los datos a `R`:

```
library("tidyverse")
library("readr")
path <- url("https://raw.githubusercontent.com/mebucca/cda_soc3070/master/homework/tf/redcard_data.csv")
data_redcards <- read.csv(path)

data_redcards %>% glimpse()
```

3. A partir de los datos disponibles, selecciona dos variables de interés focal (una dependiente y una independiente) cuya relación estudiarás en este trabajo.

Por ejemplo, a los investigadores que participaron del proyecto "Many Analysts, One Data Set" (Silberzahn et at 2018) se les pidió responder la siguiente pregunta: **"Are soccer players with dark skin tone more likely to receive red cards from referees than players with light skin tone?"**. Ustedes pueden usar estas mismas variables de interés (tarjetas rojas y color de piel) u otras (ejemplos: número del goles marcados y altura del jugador; posición en la que juega el jugador y color de piel, etc.).

4. AB2021 señalan que, a pesar de haber sido instruidos con la misma pregunta de investigación, diferentes grupos interpretaron esta pregunta de formas distintas. En las páginas 3-5 AB2021 describen en detalle cuatro interpretaciones posibles de la misma pregunta.

Para este trabajo deberán elegir entre una interpretación del tipo 2 –estudiar el efecto de la variable independiente sobre la variable dependiente neto de posibles "confounders"– o una interpretación del tipo 3 –desarrollar un modelo predictivo para la variable dependiente de interés. Una vez seleccionado el tipo de pregunta, expresa tu pregunta empírica de manera clara. Ejemplos:

- pregunta tipo 1: "¿Tiene la altura de los jugadores un efecto sobre la cantidad de goles que marcan?"
- pregunta tipo 2: "¿Qué factores predicen la cantidad goles marcados por un jugador?"

5. Selecciona un conjunto de controles/predictores para incluir en tu modelo de regresión y explica breve pero claramente la razones para su inclusión. Por simplicidad, elige un máximo de 2 controles/predictores. Ejemplos:

- modelo tipo 1: modelar la cantidad de goles marcados un jugador como función de su altura, controlado por peso. Razonamiento: los jugadores más altos tienden a pesar más. Al mismo tiempo, el peso podría aumentar negativamente la probabilidad de marcar goles. Por tanto, al controlar por edad se estudiaría el efecto de la altura entre jugadores de igual peso. Por tanto, se espera que este efecto sea mayor que la asociación bivariada entre altura y goles marcados.
- modelo tipo 2: modelar la cantidad de goles marcados como función de todos los factores observados que podrían tener un efecto: posición, altura, peso, edad, etc. Razonamiento: los goles marcados dependerían tanto de la edad como de la posición en la que juegan los jugadores, así como también de su altura y peso.

6. Decide la forma funcional de tu modelo de regresión. Es decir, la relación algebraica entre variables dependiente e independientes (términos cuadráticos, transformaciones logarítmicas, interacciones, etc.). Expresa formalmente la ecuación correspondiente y justifica esta decisión. Ejemplo

- modelo tipo 1/2: $f(\text{goles marcados}) = \alpha + \beta_1 * \text{altura} + \beta_2 * \text{edad} + \beta_3 * \text{edad}^2$

Justificación: La edad podría tener un efecto positivo dentro del rango de edades bajas porque jugadores más experimentados tienen mayores chances de hacer goles, pero pasado cierta edad el efecto podría ser negativo porque jugadores más viejos tienden a ser más lentos.

7. Decide el tipo de modelo de regresión (cubierto en este curso) que mejor se ajusta al tipo de datos con usarás para responder tu pregunta empírica. Justifica esta decisión y escribe la ecuación de regresión correspondiente. Nota que la ecuación de arriba usa genéricamente $f(.)$, sin especificar la función de enlace correspondiente a un modelo de regresión en particular. La ecuación que escribas acá debe considerar tanto la forma funcional del modelo (punto 6) como las características propias del tipo de regresión escogido.

8. Ajusta el modelo de regresión planteado en el punto 7 y reporta los resultados. La información relevante a reportar variará dependiendo de la opción tomada en el punto 4.

9. Interpreta los resultados. Los resultados relevantes y su interpretación variarán dependiendo de la opción tomada en el punto 4.

Bonus: todo análisis adicional destinado a mejor la calidad y precisión de los resultados reportados será considerado en la nota del trabajo (ejemplos: gráficos, definir cantidades de interés, realizar inferencia via bootstrap, cross-validation predictiva, estimar diferentes modelos para evaluar robustez, u otros.).

La no realización de análisis adicionales no será penalizada.

Formato:

- El trabajo final debe ser entregado en formato de reporte de investigación.
- El reporte debe contener todos los elementos señalados en las instrucciones. No es necesario (ni esperable) que las preguntas sean respondidas punto por punto.
- El reporte debe tener un máximo de 4 páginas. Si fuera necesario, incluir análisis suplementarios en un apéndice.
- Los resultados estadísticos deben ser reportados de manera clara pero no es requerido el uso de tablas o gráficos de calidad publicable.