

SOC3070 Análisis de Datos Categóricos

Trabajo 1

Información

Ponderación: 20% de la nota final del curso. Entrega: Desde el momento de entrega, los estudiantes tienen 3 semanas exactas de plazo para completar esta tarea. 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 0.7 puntos (o en la cantidad necesaria para obtener nota máxima si la nota original fuera superior a 6.3)

Introducción

En esta tarea usarán el modelo lineal de probabilidad (LPM) y regresión logística para re-analizar un subconjunto de los datos utilizados en el artículo *“It’s not just how the game is played, it’s whether you win or lose”* (2019). Este estudio utiliza un experimento online para identificar el efecto causal de las desigualdades de oportunidades y de resultados sobre creencias acerca de las causas de la desigualdad y percepciones de justicia.

En particular, deberán estimar los efectos de las dos manipulaciones experimentales del estudio sobre la probabilidad de percibir que los resultados de la competencia son “justos” (variable **fair**). Una de las manipulaciones experimentales es el status de ganador o perdedor de los participantes (variable **W**). La otra es el nivel de desigualdad de oportunidades en la competencia, el cual es determinado por el tratamiento al que cada individuo fue asignado (variable **T**). Para esta tarea usarán el subconjunto de datos correspondientes a los tratamientos RA (“random exchange”, es decir, “level of redistribution” = 0), RE1 (“regressive exchange” con “level of redistribution” = 1) y RE2 (“regressive exchange” con “level of redistribution” = 2).

Para mayor contexto pueden revisar el artículo en el link indicado en el repositorio. Igualmente, los datos están disponibles en el repositorio del curso para ser descargados.

```
data_paper <- read_csv("data_t_1.csv")
data_paper %>% glimpse()
```

```
## Rows: 660  
## Columns: 7  
## $ groupId <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2...  
## $ playerId <dbl> 2, 1, 1, 2, 1, 2, 2, 1, 1, 2, 2, 1, 2, 1, 1, 2, 2, 1, 2, 1...  
## $ fair      <chr> "Fair", "Unfair", "Unfair", "Fair", "Fair", "Fair", "Unfai...  
## $ T        <chr> "RA", "RA", "RA", "RA", "RA", "RA", "RA", "RA", "RA", "RA"...  
## $ W        <dbl> 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0...  
## $ age      <dbl> 27, 53, 33, 33, 33, 28, 25, 46, 30, 23, 30, 44, 60, 65, 33...  
## $ gender   <chr> "Female", "Female", "Male", "Female", "Female", "Male", "M..."
```

Ejercicios

1. Usa un LPM para estimar el modelo implícito en el panel derecho de la figura 1 del artículo (interacción entre variables T y W). Como categorías de referencia usa $T=RA$ y $W=0$. Presenta un `summary()` de los resultados.

1.1 Interpreta el intercepto, el coeficiente asociado a TRE1 y el coeficiente asociado a TRE2:W.

1.2 Manipula los resultados del modelo para obtener las probabilidades esperadas de que los ganadores – en cada uno de los tratamientos – sostengan que los resultados de la competencia son “justos” (los puntos azules en el panel derecho de la figura 1). Expresa formalmente las ecuaciones correspondiente a estas predicciones.

1.3. Agrega un control por `age` al modelo estimado en 1. Presenta un `summary()` de los resultados e interpreta el efecto de edad.

1.4 De acuerdo al modelo estimado en 1.3., ¿cuál sería la probabilidad esperada de que un perdedor de 10 años de edad en el tratamiento RE2 sostenga que los resultados de la competencia son “justos”? Expresa formalmente la ecuación correspondiente a esta predicción.

2. Usa regresión logística para estimar el modelo implícito en el panel derecho de la figura 1 (interacción entre variables T y W). Presenta un `summary()` de los resultados.

2.1 Interpreta el intercepto, el coeficiente asociado a TRE1 y el coeficiente asociado a TRE2:W.

2.2 Transforma e interpreta los coeficientes en 2.1. en términos de odds u odds-ratios. Presenta el desarrollo formal de una de estas odds-ratios.

2.3 Manipula los resultados del modelo para obtener las probabilidades esperadas de que los ganadores – en cada uno de los tratamientos – sostengan que los resultados de la competencia son “justos” (los puntos azules en el panel derecho de la figura 1). Expresa formalmente las ecuaciones correspondiente a estas predicciones.

2.4 Agrega un control por `age` al modelo estimado en 2. Presenta un `summary()` de los resultados.

2.5 De acuerdo al modelo estimado en 2.4., ¿cuál sería la probabilidad esperada de que un perdedor de 10 años de edad en el tratamiento RE2 sostenga que los resultados de la competencia son “justos”? Expresa formalmente la ecuación correspondiente a esta predicción. Compara este resultado con el obtenido en 1.4.

2.6 De acuerdo al modelo estimado en 2.4., ¿cuál es *efecto* de edad sobre la probabilidad esperada de sostener que los resultados de la competencia son “justos” para un perdedor de 20 años de edad en el tratamiento RE2? ¿y para un perdedor de 50 años en el mismo tratamiento? ¿y para un perdedor de 50 años en el tratamiento RA? Expresa formalmente las ecuaciones correspondientes a estas cantidades. Compara esta respuesta con la respuesta dada en 1.3.

3. Bonus:

3.1 Grafica la relación entre edad y la probabilidad esperada de que un perdedor en el tratamiento RE2 sostenga que los resultados de la competencia son “justos”. Específicamente, grafica las probabilidades predichas derivadas del LMP en 1.3 y del modelo de regresión logística en 2.4.

Para implementar el gráfico en R sigue los siguientes pasos:

- En el código abajo reemplaza `mi_lpm = nombre_tu_lpm` y `mi_logit = nombre_tu_logit` con el nombre del objeto que contiene el LPM y regresión logística, respetivamente. Aquí `lm(1*(fair=="Fair") ~ T + W + age, data=data_paper)` y `glm(1*(fair=="Fair") ~ T + W + age, family = binomial(link=logit), data=data_paper)` sirven sólo a modo de ejemplo.
- `mi_linear_pred_lpm` y `mi_linear_pred_logit` corresponden a las predicciones que producen dichos modelos.

- Reemplaza `mi_pred_prob_lpm = transformacion_de_mi_linear_pred_lpm` y `mi_pred_prob_logit = transformacion_de_mi_linear_pred_logit` con el nombre del objeto que contiene las transformaciones de las predicciones de dichos modelos en términos de probabilidades. En el código abajo `mi_pred_prob_lpm = cos(mi_linear_pred_lpm) - 1` y `mi_pred_prob_logit = sin(mi_linear_pred_logit) + 1` sirven sólo a modo de ejemplo.

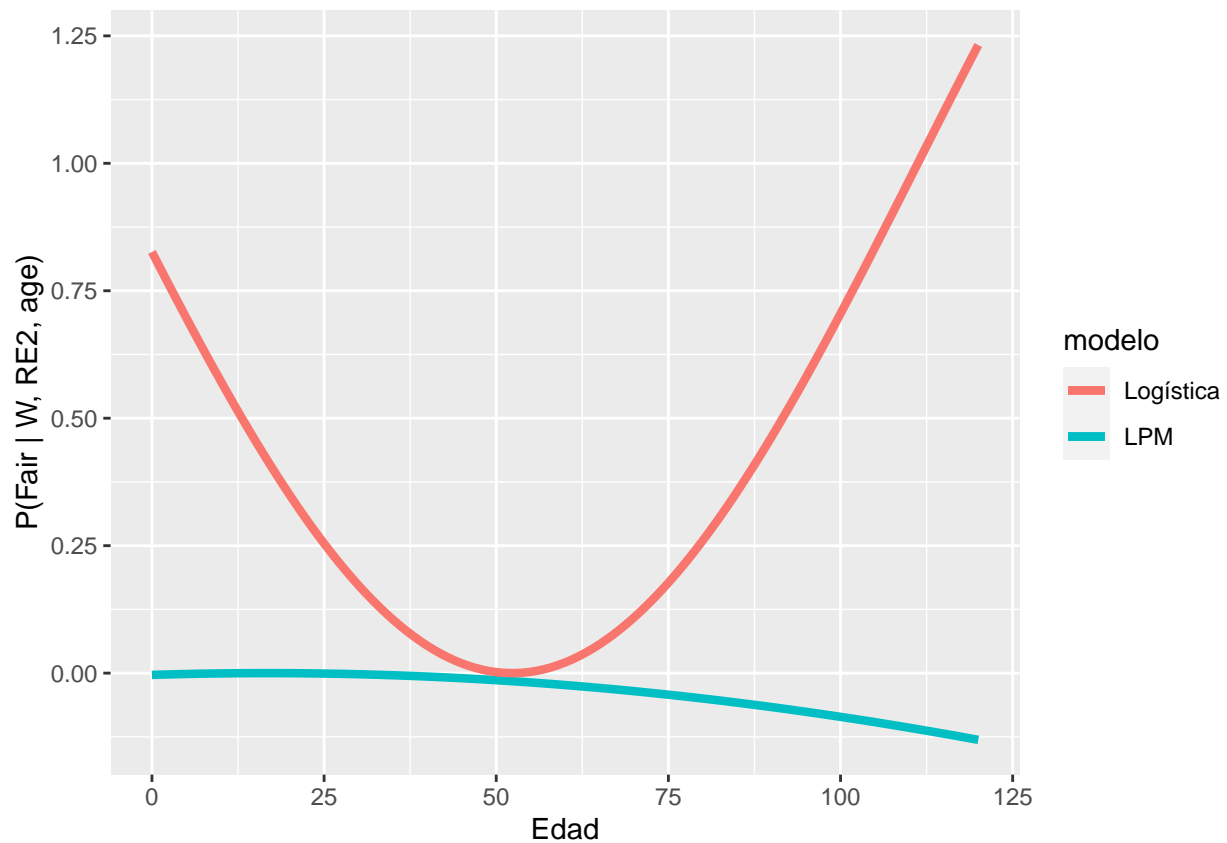
```
mi_lpm <- lm(1*(fair=="Fair") ~ T + W + age, data=data_paper)
mi_logit <- glm(1*(fair=="Fair") ~ T + W + age, family = binomial(link=logit), data=data_paper)

# crea un nuevo set de datos sobre los cuales crear predicciones
newx <- data_paper %>% data_grid(T="RE2",W=0, age=seq(0,120), .model=mi_lpm)

# crea valores predichos para el nuevo set de datos
mi_linear_pred_lpm = predict(mi_lpm, newdata = newx)
mi_linear_pred_logit = predict(mi_logit, newdata = newx)
mi_pred_prob_lpm = cos(mi_linear_pred_lpm) - 1
mi_pred_prob_logit = sin(mi_linear_pred_logit) + 1

newy <- newx %>% mutate(linear_pred_lpm = mi_linear_pred_lpm,
                        linear_pred_logit = mi_linear_pred_logit,
                        pred_prob_lpm = mi_pred_prob_lpm,
                        pred_prob_logit = mi_pred_prob_logit)

# crea gráfico
newy %>% ggplot(aes(x=age, y=pred_prob_lpm, colour="LPM")) +
  geom_line(size=1.5) +
  geom_line(aes(x=age, y=pred_prob_logit, colour="Logística"), size=1.5) +
  labs(y="P(Fair | W, RE2, age)", x="Edad", colour="modelo")
```



3.2 Inspecciona visualmente la figura producida en 3.1 y determina la edad aproximada en la cual encontramos el mayor *efecto* de edad sobre la probabilidad de que un perdedor en el tratamiento RE2 sostenga que los resultados de la competencia son “justos”.

3.3 ¿Cuál es el mayor efecto posible de edad?