SOL3070 Análisis de Datos Categóricos

Trabajo 1

Información

- Ponderación: 20% de la nota final del curso.
- 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)

Introducción

En esta tarea usarán el modelo lineal de probabilidad (LPM) y regresión logística para re-analizar 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. Para mayor contexto pueden revisar el artículo en el link indicado en el repositorio.

Específicamente, deberán trabajar con un modelo de regresión perteneciente a la sección "Supplementary Materials". Tanto el artículo como el material suplementario se encuentran disponibles en el repositorio. En ambos documentos encontrarán destacadas las partes relevantes para desarrollar esta tarea.

Datos

Los datos están disponibles en el repositorio del curso para ser descargados. Visualización rápida de la base de datos:

```
path <- "/Users/Mauricio/Library/Mobile Documents/com~apple~CloudDocs/Teaching/ISUC/2021_2
setwd(path)
```

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

```
Rows: 857
Columns: 17
$ pid      <dbl> 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0~
$ wt       <dbl> 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1~
$ t        <chr> "RA2", "RA2", "RA2", "RA2", "RA1", "RA1", "RA1", "PR2", "PR2"~
$ score    <dbl> 551, 569, 606, 484, 566, 595, 528, 575, 569, 558, 565, 601, 5~
$ hs       <dbl> 49.19643, 50.80357, 55.59633, 44.40367, NA, 52.98308, 47.0169~
$ ioo      <dbl> 0.008035714, 0.008035714, 0.055963303, 0.055963303, NA, 0.029~
$ wg       <dbl> 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0~
$ fair     <chr> "Unfair", "Unfair", "Unfair", "Fair", "Unfair", "Fair", "Unfa~
$ most     <chr> "luck", "luck", "talent", "luck", "luck", "luck", "luck", "ru~
$ feelings <chr> "angry", "indifferent", "happy", "indifferent", "sad", "happy~
$ gender   <chr> "Male", "Male", "Male", "Male", "Female", "Female", "Female",~
$ age      <dbl> 21, 29, 25, 42, 30, 27, 53, 33, 25, 48, 45, 42, 32, 38, 46, 3~
$ religion <chr> "none", "catholic", "other", "protestant", "protestant", "non~
$ race     <chr> "white", "white", "white", "white", "white", "white", "white"~
$ inc      <chr> "$50,000-$75,000", "$0-$25,000", "$50,000-$75,000", "$25,000-~
$ pol      <dbl> 1, 6, 6, 8, 8, 4, 3, 1, 4, 6, 6, 5, 1, 6, 6, 8, 1, 6, 6, 4, 3~
$ educ     <chr> "Some College", "College Degree", "High School", "Some Colleg~
```

Descripción de variables relevantes

- `wg` indica si el jugador ganó (`wg`=1) o perdió el juego (`wg`=0).
- `pid` indica si el jugador es "player 1" (`pid`=0) o "player 2" (`pid`=1).
- `hs` corresponde a "hand strength". La variable utilizada acá es equivalente a la usada en el artículo pero multiplicada por 100 para facilitar la interpretación de resultados. Valores cercanos a cero indican que el jugador tenía carta débiles y valores cercanos a 100 indican que el jugador tenía cartas fuertes.
- `wt` indica si el jugador ganó (`wt`=1) o perdió la sesión de entrenamiento (`wt`=0).

Ejercicios

1. Usa un LPM para estimar el modelo implícito en la Figura S3 de la sección "Supplementary Materials" del artículo. Puedes encontrar mayores detalles sobre la especificación del modelo en la ecuación 4 y en la Tabla S3, columna 1. Escribe la escuación de regresión y presenta un `summary()` de los resultados.

La ecuación de regresión es: $\mathbb{P}(\text{wg} \mid \text{pid, hs, wt}) = y = \beta_0 + \beta_{pid}\text{pid} + \beta_{hs}\text{hs} + \beta_{wt}\text{wt}$

Implementación en `R`:

```
lpm_1 <- lm(wg ~ factor(pid) + hs + factor(wt) , data=data_paper)
summary(lpm_1)
```

```
Call:
lm(formula = wg ~ factor(pid) + hs + factor(wt), data = data_paper)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80571 -0.36301 -0.00011  0.35995  0.76643

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.373112   0.137740 -17.229  < 2e-16 ***
factor(pid)1 -0.215152   0.027016  -7.964 5.57e-15 ***
hs            0.059134   0.002705  21.858  < 2e-16 ***
factor(wt)1   0.050784   0.027050   1.877   0.0608 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3871 on 818 degrees of freedom
  (35 observations deleted due to missingness)
Multiple R-squared:  0.4036,    Adjusted R-squared:  0.4014
F-statistic: 184.5 on 3 and 818 DF,  p-value: < 2.2e-16
```

1.1 Interpreta el coeficiente asociado a `wt`.

El coeficiente asociado a `wt` indica que, manteniendo los otros factores constantes, la probabilidad esperada de que un ganador de la sesión de entrenamiendo gane el juego es 5 punto porcentuales mayor que la de un jugador que perdió en el entrenamiento.

1.2 De acuerdo al modelo estimado en 1., ¿cuál es el efecto marginal de "hand strength" sobre la probabilidad esperada de ganar el juego?

El coeficiente asociado a `hs` indica que, manteniendo los otros factores constantes, un aumento en 1 unidad de "hand strength" se traduce en un aumento de 6 puntos porcentuales en la probabilidad esperada ganar el jugo.

1.3 En base al modelo usado en 1., calcula las probabilidades esperadas de que un "player 1" que ganó el entrenamiento gane el juego si su "hand strength" es 50 y 60, respectivamente. Expresa formalmente las ecuaciones correspondiente a estas predicciones.

Formalmente:

- $\mathbb{P}(\text{wg} \mid \text{pid=0, hs=50, wt=1}) = \beta_0 + \beta_{hs} * 50 + \beta_{wt} = -2.373112 + 0.059134 * 50 + 0.050784 = 0.634372$
- $\mathbb{P}(\text{wg} \mid \text{pid=0, hs=60, wt=1}) = \beta_0 + \beta_{hs} * 60 + \beta_{wt} = -2.373112 + 0.059134 * 60 + 0.050784 = 1.225712$

Implementación en R:

```
newx <- data_paper %>% data_grid(pid=0,hs=c(50,60),wt=1,.model=lpm_1)
newy <- newx %>% mutate(pred_prob = predict(lpm_1, newdata = newx))
print(newy)
```

```
# A tibble: 2 x 4
    pid    hs    wt pred_prob
  <dbl> <dbl> <dbl>     <dbl>
1     0    50     1     0.634
2     0    60     1     1.23
```

1.4. Agrega una interacción entre **hs** y **wt** al modelo estimado en 1. Escribe la ecuación de regresión y presenta un `summary()` de los resultados. Interpreta el efecto de "hand strength" estimado en este modelo.

La ecuación de regresión es: $\mathbb{P}(\text{wg} \mid \text{pid, hs, wt}) = y = \beta_0 + \beta_{pid}\text{pid} + \beta_{hs}\text{hs} + \beta_{wt}\text{wt} + \beta_{hs:wt}\text{hs} * \text{wt}$

Implementación en R:

```
lpm_2 <- lm(wg ~ factor(pid) + hs*wt , data=data_paper)
summary(lpm_2)
```

```
Call:
lm(formula = wg ~ factor(pid) + hs * wt, data = data_paper)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8060 -0.3630 -0.0001  0.3599  0.7662
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.3680716  0.1957315 -12.099  < 2e-16 ***
factor(pid)1 -0.2151308  0.0270389  -7.956  5.9e-15 ***
hs            0.0590335  0.0038814  15.209  < 2e-16 ***
wt            0.0409646  0.2721146   0.151    0.880
hs:wt         0.0001964  0.0054154   0.036    0.971
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3873 on 817 degrees of freedom
  (35 observations deleted due to missingness)
Multiple R-squared:  0.4036,    Adjusted R-squared:  0.4007
F-statistic: 138.2 on 4 and 817 DF,  p-value: < 2.2e-16
```

Al añadir una interacción entre **hs** y **wt** permitimos que el efecto de "hand strength" dependa de si el jugador ganó o no la sesión de entrenamiento (y viceversa). Es decir, no hay un único efecto de "hand strength" si no dos. Para los perdedores de la sesión de entrenamiento la probabilidad esperada de ganar el juego viene dada por:

- $\mathbb{P}(\text{wg} \mid \text{pid, hs, wt=0}) = \beta_0 + \beta_{pid}\text{pid} + \beta_{hs}\text{hs}$.

Por tanto, el efecto de **hs** sobre la probabilidad de ganar el juego es: $\frac{\partial \mathbb{P}(\text{wg}|\text{pid, hs, wt=0})}{\partial \text{hs}} = \beta_{hs}$

Por su parte, para los ganadores de la sesión de entrenamiento la probabilidad esperada de ganar el juego viene dada por:

- $\mathbb{P}(\text{wg} \mid \text{pid, hs, wt=1}) = \beta_0 + \beta_{pid}\text{pid} + \beta_{hs}\text{hs} + \beta_{wt} + \beta_{hs:wt}\text{hs} = \beta_0 + \beta_{pid}\text{pid} + \beta_{wt} + (\beta_{hs} + \beta_{hs:wt})\text{hs}$

Por tanto: $\frac{\partial \mathbb{P}(\text{wg}|\text{pid, hs, wt=1})}{\partial \text{hs}} = \beta_{hs} + \beta_{hs:wt}$

1.5 En base a los resultados del modelo en 1.4. calcula las probabilidades esperadas de que un "player 1" que ganó el entrenamiento gane el juego si su "hand strength" es 60. Expresa formalmente la ecuación correspondiente a esta predicción.

- $\mathbb{P}(\text{wg} \mid \text{pid=0, hs=60, wt=1}) = \beta_0 + \beta_{wt} + (\beta_{hs} + \beta_{hs:wt}) * 60 = -2.3680716 + 0.0409646 + (0.0590335 + 0.0001964) * 60 = 1.226687$

Implementación en R:

```r
newx <- data_paper %>% data_grid(pid=0,hs=60,wt=1,.model=lpm_2)
newy <- newx %>% mutate(pred_prob = predict(lpm_2, newdata = newx))
print(newy)
```

```
# A tibble: 1 x 4
    pid    hs    wt pred_prob
  <dbl> <dbl> <dbl>     <dbl>
1     0    60     1      1.23
```

2. Usa una regresión logística para estimar el modelo implícito en la Figura S3 de la sección "Supplementary Materials" del artículo. Puedes encontrar mayores detalles sobre la especificación del modelo en la ecuación 4 y en la Tabla S3, columna 1. Escribe la escuación de regresión y presenta un `summary()` de los resultados.

La ecuación de regresión es: $\ln \frac{\mathbb{P}(\text{wg} \mid \text{pid, h, wt})}{1-\mathbb{P}(\text{wg} \mid \text{pid, h, wt})} = \beta_0 + \beta_{pid}\text{pid} + \beta_{hs}\text{hs} + \beta_{wt}\text{wt}$

Implementación en `R`:

```r
logit_1 <- glm(wg ~ factor(pid) + hs + wt, family = binomial(link=logit), data=data_paper)
summary(logit_1)
```

```
Call:
glm(formula = wg ~ factor(pid) + hs + wt, family = binomial(link = logit),
    data = data_paper)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -33.18254    2.94112 -11.282  < 2e-16 ***
factor(pid)1  -1.57179    0.20472  -7.678 1.62e-14 ***
hs             0.67661    0.05935  11.401  < 2e-16 ***
wt             0.30697    0.19394   1.583    0.113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1139.53  on 821  degrees of freedom
Residual deviance:  650.91  on 818  degrees of freedom
  (35 observations deleted due to missingness)
AIC: 658.91

Number of Fisher Scoring iterations: 6
```

2.1 Interpreta el coeficiente asociado a `wt`.

El coeficiente asociado a `wt` indica que, manteniendo los otros factores constantes, las log-odds de ganar el juego de los ganadores de la sesión de entrenamiento son 0.31 puntos mayores que las de los perdedores.

2.2 Transforma e interpreta el coeficiente de interés de 2.1. en términos de odds-ratios. Presenta el desarrollo formal.

Transformamos los coeficientes en odds ratios exponenciando los coeficientes originales:

Implementación en `R`:

```
exp(summary(logit_1)$coefficients[4,1])
```

```
[1] 1.359306
```

Formalmente: Dado la ecuación de regresión logística,

$\ln \frac{\mathbb{P}(\text{wg} \mid \text{pid, h, wt})}{1-\mathbb{P}(\text{wg} \mid \text{pid, h, wt})} = \beta_0 + \beta_{pid}\text{pid} + \beta_{hs}\text{hs} + \beta_{wt}\text{wt}$

podemos re-expresar el modelo en términos de odds de ganar el juego:

$\frac{\mathbb{P}(\text{wg} \mid \text{pid, h, wt})}{1-\mathbb{P}(\text{wg} \mid \text{pid, h, wt})} = e^{\beta_0} e^{\beta_{pid}\text{pid}} e^{\beta_{hs}\text{hs}} e^{\beta_{wt}\text{wt}}$

Por tanto, las odds de ganar el juego para un ganador de la sesión de entrenamiento son:

$\text{odds}_{\text{wt=1}} : \frac{\mathbb{P}(\text{wg} \mid \text{pid, h, wt=1})}{1-\mathbb{P}(\text{wg} \mid \text{pid, h, wt=1})} = e^{\beta_0} e^{\beta_{pid}\text{pid}} e^{\beta_{hs}\text{hs}} e^{\beta_{wt}}$

y las odds de ganar el juego para un perdedor de la sesión de entrenamiento son:

$\text{odds}_{\text{wt=0}} : \frac{\mathbb{P}(\text{wg} \mid \text{pid, h, wt=0})}{1-\mathbb{P}(\text{wg} \mid \text{pid, h, wt=0})} = e^{\beta_0} e^{\beta_{pid}\text{pid}} e^{\beta_{hs}\text{hs}}$

Luego, la razón de las odds de ganar el juego entre un ganador y un perdedorde la sesión de entrenamiento es:

$\frac{\text{odds}_{\text{wt=1}}}{\text{odds}_{\text{wt=0}}} = \frac{e^{\beta_0} e^{\beta_{pid}\text{pid}} e^{\beta_{hs}\text{hs}} e^{\beta_{wt}}}{e^{\beta_0} e^{\beta_{pid}\text{pid}} e^{\beta_{hs}\text{hs}}} = e^{\beta_{wt}}$

Es decir, las odds de ganar el juego de un ganador de la sesión de entrenamiento son 1.36 veces las odds de un perdedor del entrenamiento (36% más altas).

2.3 De acuerdo al modelo estimado en 2, calcula las probabilidades esperadas de que un "player 1" que ganó el entrenamiento gane el juego si su "hand strength" es 50 y 60, respectivamente. Expresa formalmente las ecuaciones correspondiente a estas predicciones. Compara este resultado con el obtenido en 1.3.

Formalmente:

- $\mathbb{P}(\text{wg} \mid \text{pid=0, hs=50, wt=1}) = \frac{1}{1+e^{-(\beta_0+\beta_{hs}*50+\beta_{wt})}} = 1/(1+\exp(-(-33.18254+0.67661*50+0.30697))) = 0.7221056$

- $\mathbb{P}(\text{wg} \mid \text{pid=0, hs=60, wt=1}) = \frac{1}{1+e^{-(\beta_0+\beta_{hs}*60+\beta_{wt})}} = 1/(1+\exp(-(-33.18254+0.67661*60+0.30697))) = 0.9995568$

Implementación en `R`:

```
newx <- data_paper %>% data_grid(pid=0,hs=c(50,60),wt=1,.model=logit_1)
newy <- newx %>% mutate(pred_prob = predict(logit_1, newdata = newx, type="response"))
print(newy)
```

```
# A tibble: 2 x 4
    pid    hs    wt pred_prob
  <dbl> <dbl> <dbl>     <dbl>
1     0    50     1     0.722
2     0    60     1     1.00
```

A diferencia del LPM, el sigmoide $1/(1+e^{-x})$ garantiza que las probabilidades predichas siempre serán restringidas al rango 0-1.

2.4. De acuerdo al modelo estimado en 2., ¿cual es el efecto marginal de "hand strength" sobre la probabilidad esperada de ganar el juego? Calcula esta cantidad para un "Player 1" que ganó el entrenamiento si su "hand strength" es 50 y si su "hand strength" es 60. Expresa formalmente las ecuaciones correspondientes a estas cantidades. Compara esta respuesta con la respuesta dada en 1.2.

Recordar que: $\frac{\partial p_i}{\partial \text{hs}} = \beta_{\text{hs}} \times p_i(1-p_i)$

donde $p_i = \mathbb{P}(\text{wg} \mid \text{pid, hs, wt}) = \frac{1}{1+e^{-(\beta_0+\beta_{pid}\text{pid}+\beta_{hs}\text{hs}+\beta_{wt}\text{wt})}})$

De 2.3. sabemos que para un "player 1" que ganó el entrenamiento y cuyo "hand strength" es 50, $\mathbb{P}(\text{wg} \mid \text{pid=0, hs=50, wt=1}) = 0.7221056$. Por tanto,

- $\frac{\partial p_i}{\partial \text{hs}} = \beta_{\text{hs}} \times p_i(1-p_i) = 0.67661 * 0.7221056 * (1-0.7221056) = 0.1357747$

Es decir, incrementar desde `hs=50` a `hs=51` implica un aumento de 14 puntos porcentuales en la probabilidad de ganar el juego.

Por su parte, de 2.3. sabemos que para un "player 1" que ganó el entrenamiento y cuyo "hand strength" es 60, $\mathbb{P}(\text{wg} \mid \text{pid=0, hs=60, wt=1}) = 0.9995568$. Por tanto,

- $\frac{\partial p_i}{\partial \text{hs}} = \beta_{\text{hs}} \times p_i(1-p_i) = 0.67661 * 0.9995568 * (1-0.9995568) = 0.0002997406$

Es decir, incrementar desde `hs=60` a `hs=61` implica un aumento prácticamente nulo en la probabilidad de ganar el juego.

A diferencia del LPM, en el modelo de regresión logística el efecto de "hand stregth" depende del valor de la misma variable y del valor de otras covariables. Esta característica regula que

los efectos sean muy limitados cuando las probabilidades se acercan a cero o uno, mientras que pueden ser mayores lejos de dichos valores.

Bonus:

El siguiente gráfico describe las probabilidad predichas por el LMP estimado en 1. y la regresión logística estimada en 2. de que un "player 1" que ganó la sesión de entrenamiento gane el juego, para diferentes valores de "hand strength".

Implementación en R:

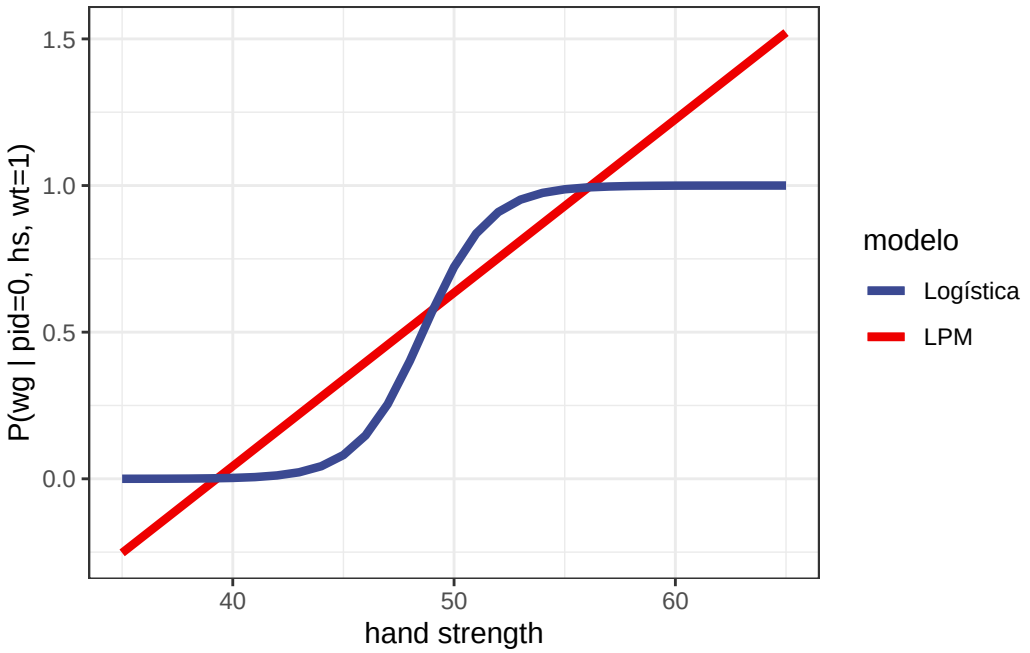
```
# crea un nuevo set de datos sobre los cuales crear predicciones
newx <- data_paper %>% data_grid(pid=0,hs=seq(35,65,by=1),wt=1, .model=lpm_1)

# crea valores predichos para el nuevo set de datos
xb_lpm = predict(lpm_1 , newdata = newx)
xb_logit = predict(logit_1, newdata = newx)
prob_lpm = xb_lpm
prob_logit = 1/(1 + exp(-xb_logit))

newy <- newx %>% mutate(prob_lpm = prob_lpm, prob_logit = prob_logit)

# crea gráfico
newy %>% ggplot(aes(x=hs, y=prob_lpm, colour="LPM")) +
  geom_line(size=1.5) +
  geom_line(aes(x=hs, y=prob_logit, colour="Logística"), size=1.5) +
  labs(y="P(wg | pid=0, hs, wt=1)", x="hand strength", colour="modelo") +
  scale_color_aaas() + theme_bw()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



3.1 Inspecciona visualmente la figura y determina la "hand strength" aproximada en la cual encontramos el mayor *efecto* de "hand strength" sobre la probabilidad de que un "player 1" que ganó el entrenamiento gane el juego.

El mayor efecto marginal de "hand strength" se da cuando la "hand strength" es aproximádamente 48.

3.2 ¿Cuál es el mayor efecto posible de "hand strength"?

El mayor efecto de "hand strength" ocurre cuando $p_i = 0.5$, por tanto el mayor efecto marginal de "hand strength" es $\frac{\partial p_i}{\partial \text{hs}} = \beta_{\text{hs}}/4 = 0.67661/4 = 0.17$

3.3 Verdadero o falso: "el efecto no lineal de `hs` en el modelo de regresión logística (linea azul) se debe a que la especificación del modelo permite tal no-linealidad (por ejemplo, usando un término cuadrático o cúbico)". Justifica tu respuesta.

Falso. La especificación del modelo no considera un efecto no-lineal de "hand strength". La relación no lineal respecto de la probabilidad de ganar el juego es inducida por el la "link function" logit.