SOC3070 Análisis de Datos Categóricos

Tarea corta 5

Ponderación: 5% de la nota final del curso. Entrega: Desde el momento de entrega, los estudiantes tienen 1 semana exacta de plazo para completar esta tarea.

Introducción:

En su artículo *"Understanding – and Misunderstanding – Social Mobility in Britain: The Entry of the Economists, the Confusion of Politicians and the Limits of Educational Policy"* John H. Goldthorpe describe la distinción entre movilidad social absoluta y relativa, y resume los principales hallazgos al respecto en UK:

> "Sociologists attach [crucial importance] to the distinction between absolute and relative mobility rates. Absolute rates refer to the actual proportions of individuals of given class origins who are mobile to different class destinations, while relative rates compare the chances of individuals of differing class origins arriving at different class destinations and thus indicate the extent of social fluidity. [Relative mobility is a zero-sum phenomenon. If one person moves up in relative terms, another by definition must have moved down]. In these two respects, the major research findings [can be summarized] as follows.
>
> (i) Absolute rates of intergenerational class mobility, as measured in percentage terms, appear quite high. [...] Rates of upward mobility steadily increased in the course of the twentieth century, primarily as a consequence of class structural change - i.e. of the expansion of professional and managerial positions creating "more room at the top". However, immobility at the "top" also increased.
>
> (ii) Relative rates of intergenerational class mobility [...] showed a basic constancy over most of the twentieth century, or at all events no sustained directional change. [...] In other words, the strength of the association between the class positions of children and their parents, considered net of class structural effects, appeared remarkably robust.
>
> Although increasing upward mobility might create a contrary impression, Britain had not in fact become a significantly more fluid or 'open' society.

Datos:

En esta tarea usarán un subconjunto de los datos provistos por Kazuo Yamaguchi en su artículo "Models for comparing mobility tables: toward parsimony and substance" (ASR 1987) para estudiar movilidad social intergeneracional. Este subconjunto de datos corresponde a una tabla de contingencia que clasifica a padres e hijos según su clase social en USA y UK, donde tanto padres como hijos pueden pertenecer a la clase UpNM (profesionales, gerentes y funcionarios) o la clase LoM (trabajadores no agrícolas semicualificados y no cualificado).

Como se observa, la tabla tiene tres dimensiones: ocupación del hijo (filas), ocupación del padre (columnas) y país ("layer"). Para acceder a la sub-tabla 2-way correspondiente a cada país usa los índices de la tabla. Por ejemplo, `ctable[,,1]` corresponde a la tabla para USA (layer=1). `ctable[1,,1]` corresponde a la primera fila de la tabla para USA, y `ctable[,2,2]` corresponde a la columna 2 de la tabla para UK (layer=2)

```
print(ctable)
```

```
## , , Country = US
##
##       Father
## Son     UpNM  LoM
##   UpNM 1275 1159
##   LoM   272 2046
##
## , , Country = UK
##
##       Father
## Son     UpNM  LoM
##   UpNM  474  601
##   LoM   124 1789
```

Problema:

Siguiendo a Goldthorpe, podemos medir la tasa de **"movilidad absoluta"** en cada país calculando la proporción de casos que se encuentra fuera de la diagonal en cada una de las tablas. Esto nos da una estimación de la probabilidad de que un hijo alcance una clase social distinta a la de sus padres. Usando esta medida obtenemos las tasas de movilidad absoluta descritas abajo, donde la diferencia entre USA y UK es estadísticamente significativa.

```
rate_abs_immobility <- ctable %>% prop.table(3) %>% apply(.,3,diag) %>% apply(.,2,sum)
rate_abs_mobility <- 1 -  rate_abs_immobility

p_usa <- rate_abs_mobility[1]
p_uk  <- rate_abs_mobility[2]
var_usa <- (p_usa*(1-p_usa))/sum(ctable[,,1])
var_uk <- (p_uk*(1-p_uk))/sum(ctable[,,2])

ci_diff <- round((p_usa - p_uk) + c(-1.96,1.96)*sqrt(var_usa + var_uk), 2)
print("Tasas de movilidad absoluta en USA y UK")
```

```
## [1] "Tasas de movilidad absoluta en USA y UK"
```

```
print(rate_abs_mobility)
```

```
##        US        UK
## 0.3011364 0.2426372
```

```
print(paste0("95% CI diferencia movilidad absoluta USA-UK: [", ci_diff[1],",",ci_diff[2],"]"))
```

```
## [1] "95% CI diferencia movilidad absoluta USA-UK: [0.04,0.08]"
```

USA presenta mayores niveles de movilidad social absoluta que UK. Sin embargo, nos interesa entender si tales diferencias reflejan, al menos parcialmente, mayores niveles de **"movilidad relativa"** en USA comparado con UK.

Preguntas:

1) Elije una medida de asociación que, siguiendo la definición de Goldthorpe, capture adecuadamente los niveles de **"movilidad relativa"** en cada país. Justifica tu decisión.

2) Calcula un intervalo de confianza al 95% para el estadístico correspondiente a cada país (o el log de éste) .

3) Calcula un intervalo de confianza al 95% para la diferencia entre ambos estadísticos (o la diferencia del log de éstos). Comenta brevemente las implicancias sustantivas de este resultado.

Respuestas:

1) La medida apropiada es la Odds ratio porque es una medida "margins-free", es decir, captura la asociación neta entre dos variables categóricas sin verse afectada por la distribución marginal de éstas. Como medida de movilidad relativa las odds ratio capturan el grado con que la clase de origen afecta las chances de alcanzar diferentes clases de destino, independiente de potenciales cambios en la estructura de clases entre ambas generaciones.

2) Podemos obtener un intervalo al 95% de confianza para la log odds ratio usandos la siguiente formula:

$$95\%CI : \ln \hat{\theta} + 1.956 \cdot SE_{\ln \hat{\theta}}$$

donde SE es la desviación estándar de la "sampling distribution" de la log odds ratio. Formalmente:

$$SE_{\ln \hat{\theta}} = \sqrt{\mathbb{V}\text{ar}(\ln \hat{\theta})} = \sqrt{\sum_{ij} 1/n_{ij}}$$

En R:

```
# Función que calcula log odds ratio y respectivo intervalo de confianza al 95%
ci95_logOR <- function(table) {
  lor <- log((table[1,1]*table[2,2])/(table[1,2]*table[2,1]))
  se_lor  <- sqrt(sum(1/table))
  ci <- lor + c(-1.96,1.96)*se_lor
  return(c(logOR=lor, ci_ll=ci[1],ci_ul=ci[2]))
}

# Aplica función a cada "layer" de la tabla 3-way y para obtener resultados en términos de log odds rat

apply(ctable,3,ci95_logOR)
```

```
##        Country
##                US       UK
##   logOR 2.113228 2.431743
##   ci_ll 1.963800 2.213512
##   ci_ul 2.262657 2.649974
```

```
# Aplica función a cada "layer" de la tabla 3-way y exponencia para obtener resultados en términos de o

exp(apply(ctable,3,ci95_logOR))
```

```
##         Country
##                US        UK
##    logOR 8.274914 11.378697
##    ci_ll 7.126359  9.147789
##    ci_ul 9.608581 14.153665
```

```
# Versión automática usando funciones del paquete "vcd"

oddsratio(ctable, log = FALSE)
```

```
##  odds ratios for Son and Father by Country
##
##        US        UK
##  8.274914 11.378697
```

```
oddsratio(ctable, log = FALSE)  %>% confint(., level=0.95)
```

```
##       2.5 %    97.5 %
## US 7.126378  9.608555
## UK 9.147825 14.153609
```

Transformando los resultados a odds ratios observamos que la odds ratio para USA es más baja que la correspondiente a UK, lo que indica en USA existe un menor grado de asociación entre la clase de origen y de destino que en UK. Sin embargo, los intervalos de confianza de ambas odds ratios se superponen, *sugiriendo* que no existe una diferencia estadísticamente significativa en los niveles de movilidad relativa de ambos paises.

3) Definamos la diferencia de estimada de las log odds ratios como $\hat{\Delta} = \ln \hat{\theta}_{USA} - \ln \hat{\theta}_{UK}$

Podemos obtener un intervalo al 95% de confianza para la diferencia en los log odds ratio usando la siguiente formula:

$$95\% CI_{\hat{\Delta}} : \hat{\Delta} + 1.956 \cdot SE_{\hat{\Delta}}$$

donde $SE_{\hat{\Delta}}$ es la desviación estándar de la "sampling distribution" de la diferencia entre las log odds ratios. Formalmente:

$$SE = \sqrt{\mathbb{V}\text{ar}(\ln \hat{\theta}_{UK} - \ln \hat{\theta}_{USA})} = \sqrt{\mathbb{V}\text{ar}(\ln \hat{\theta}_{UK}) + \mathbb{V}\text{ar}(\ln \hat{\theta}_{USA})} = \sqrt{\sum_{ij} 1/n_{ij}^{UK} + \sum_{ij} 1/n_{ij}^{USA}}$$

Podemos implementar este calculo en `R` del siguiente modo:

```
lors <- oddsratio(ctable, log = TRUE)

lor_usa <- lors$coefficients[1]
lor_uk  <- lors$coefficients[2]
diff_lor <- (lor_uk - lor_usa)

var_lor_usa <- sum(1/ctable[,,1])
var_lor_uk <-  sum(1/ctable[,,2])
se_diff_lor <- sqrt(var_lor_uk + var_lor_usa)
```

```r
ci_diff_lor <- round( diff_lor + c(-1.96,1.96)*se_diff_lor,3)

print(lors)
```

```
## log odds ratios for Son and Father by Country
##
##       US       UK
## 2.113228 2.431743
```

```r
print(paste0("Diferencia log OR  (movilidad relativa) USA-UK:",  round(diff_lor,3) ))
```

```
## [1] "Diferencia log OR  (movilidad relativa) USA-UK:0.319"
```

```r
print(paste0("SE diferencia log OR  (movilidad relativa) USA-UK:",  round(se_diff_lor,3) ))
```

```
## [1] "SE diferencia log OR  (movilidad relativa) USA-UK:0.135"
```

```r
print(paste0("95% CI diferencia log OR  (movilidad relativa) USA-UK: [", ci_diff_lor[1],",",ci_diff_lor
```

```
## [1] "95% CI diferencia log OR  (movilidad relativa) USA-UK: [0.054,0.583]"
```

Este resultado indica que, a diferencia de lo sugerido por los intervalos de confianza de cada odds ratio, USA efectivamente presenta un nivel ligeramente más alto de movilidad social relativa respecto a UK.

Nota: puedes chequear los resultados usando una regresión logistica y comparando las cantidades de interés.

```r
subdata <- data %>% filter(Country!="Japan") %>% filter(Son=="UpNM" | Son=="LoM") %>% filter( Father=="
model_usauk <- glm(Son ~ Father*Country, weights = Freq, family = "binomial", data=subdata )
summary(model_usauk)
```

```
##
## Call:
## glm(formula = Son ~ Father * Country, family = "binomial", data = subdata,
##     weights = Freq)
##
## Deviance Residuals:
##      1       2       3       4       5       6       7       8
## -22.21   30.75  -48.56   42.85  -14.84   19.75  -40.73   32.19
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.54490    0.06679 -23.131   <2e-16 ***
## FatherLoM            2.11323    0.07624  27.719   <2e-16 ***
## CountryUK            0.20397    0.12098   1.686   0.0918 .
## FatherLoM:CountryUK  0.31851    0.13494   2.360   0.0183 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 10662.5  on 7  degrees of freedom
## Residual deviance:  8939.2  on 4  degrees of freedom
## AIC: 8947.2
##
## Number of Fisher Scoring iterations: 5
```

```
confint(model_usauk)
```

```
## Waiting for profiling to be done...
```

```
##                            2.5 %     97.5 %
## (Intercept)         -1.67771234 -1.4157971
## FatherLoM            1.96524014  2.2641769
## CountryUK           -0.03539819  0.4391582
## FatherLoM:CountryUK  0.05565564  0.5848764
```