SOC3070 Análisis de Datos Categóricos

Tarea corta 4

Ponderación: 6% de la nota final del curso.

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 dos tablas de contingencia que clasifican a padres e hijos según su clase social en USA y UK (tablas `ctable_usa` y `ctable_uk`, respectivamente), 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).

```r
# Escribir install.packages("tinytex") en la consola para instalar "tinytex"
# Carga "tinytex" para compilar PDF
library("tinytex")
library("tidyverse")
library("vcd")
library("vcdExtra")

data("Yamaguchi87")
data <- Yamaguchi87
ctable <- xtabs(Freq ~ Son + Father+Country, data=Yamaguchi87)
ctable <- ctable[c(1,4),c(1,4),c(1,2)]

ctable_usa <- ctable[,,1]
ctable_uk  <- ctable[,,2]
```

Como se observa, cada tabla tiene dos dimensiones: ocupación del hijo (filas) y ocupación del padre (columnas).

```r
#USA
print(ctable_usa)
```

```
##       Father
## Son    UpNM  LoM
##   UpNM 1275 1159
##   LoM   272 2046
```

```r
#UK
print(ctable_uk)
```

```
##       Father
## Son    UpNM  LoM
##   UpNM  474  601
##   LoM   124 1789
```

Problemas:

1). Calcula manualmente el estadístico χ^2 para ambas tablas de contingencia y evalua si la clase de origen es o no independiente de la clase de destino en USA y UK. Puedes escribir tu propia función en `R` para facilitar el trabajo. Compara tus resultados con lo obtenido usando el comando `chisq.test( tutabla, correct = FALSE)`.

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

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. Pista: Si $\hat{\beta}_{USA}$ y $\hat{\beta}_{UK}$ son los estadísticos estimados para cada país (β es un signo arbitrario para denotar el estadístico que elijas), entonces

$$SE(\hat{\beta}_{USA} - \hat{\beta}_{UK}) = \sqrt{\mathbb{V}\text{ar}(\hat{\beta}_{USA} - \hat{\beta}_{UK})} = \sqrt{\mathbb{V}\text{ar}(\hat{\beta}_{USA}) + \mathbb{V}\text{ar}(\hat{\beta}_{UK})}$$

Respuestas:

1)

```
mi_chi2 <- function(tabla) {

  n = sum(tabla)
  marginal_fila    <- apply(tabla, 1, sum)/n
  marginal_columna <- apply(tabla, 2, sum)/n

  frec_esperadas <- (marginal_fila %*% t(marginal_columna))*n
  mi_chi2 <- sum(((tabla - frec_esperadas)^2)/ frec_esperadas)
  mi_df <- (length(marginal_fila)-1)*(length(marginal_columna)-1)
  mi_pvalue <- 1 - pchisq(mi_chi2, df=1)

  print(list(chi2 = mi_chi2, pvalue = mi_pvalue))

}

mi_chi2(ctable_usa)
```

```
## $chi2
## [1] 893.4792
##
## $pvalue
## [1] 0
```

```
chisq.test(ctable_usa, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ctable_usa
## X-squared = 893.48, df = 1, p-value < 2.2e-16
```

```
mi_chi2(ctable_uk)
```

```
## $chi2
## [1] 608.1855
##
## $pvalue
## [1] 0
```

```
chisq.test(ctable_uk, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ctable_uk
## X-squared = 608.19, df = 1, p-value < 2.2e-16
```

Los resultados indican que ambos casos existe evidence para rechazar la hipótesis (nula) de que la clase de origen no está asociada a la clase de destino.

2) 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 cambios en la estructura de clases entre ambas generaciones. En este sentido una mayor odds ratio indica menor movilidad social relativa.

En R:

```
# Función que calcula odds ratio
or <- function(table) {
  theta_hat <- (table[1,1]*table[2,2])/(table[1,2]*table[2,1])
  return(theta_hat)
}

or_usa_hat <- or(ctable_usa); or_usa_hat
```

```
## [1] 8.274914
```

```
or_uk_hat  <- or(ctable_uk); or_uk_hat
```

```
## [1] 11.3787
```

3) Nuestro estadístico de interés es la diferencia entre los log odds ratios para cada país (llamémoslo Δ por conveniencia): $\Delta = \ln \hat{\theta}_{USA} - \ln \hat{\theta}_{UK}$

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

$$95\%\text{CI}_{\hat{\Delta}} : \hat{\Delta} + 2 \cdot SE_{\hat{\Delta}}$$

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

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

Podemos implementar este calculo en R del siguiente modo:

```
# Funciones para estimar la log odds ratio y su respectiva desviación estándar (SE)
log_or <- function(table) {
  theta_hat <- (table[1,1]*table[2,2])/(table[1,2]*table[2,1])
  log_theta_hat <- log(theta_hat)
  return(log_theta_hat )
}

sd_log_odd_ratio <- function(table) {
  sd_log_theta_hat  <- sqrt(sum(1/table))
  return(sd_log_theta_hat)
```

```r
}

# Aplica función a cada tabla para obtener resultados en términos de log odds ratio, luego calcula stan

log_or_usa_hat <-  log_or(ctable_usa)
sd_log_or_usa_hat <- sd_log_odd_ratio(ctable_usa)

log_or_uk_hat <-  log_or(ctable_uk)
sd_log_or_uk_hat <- sd_log_odd_ratio(ctable_uk)


delta_hat <- log_or_usa_hat - log_or_uk_hat

var_log_or_usa_hat <- sd_log_or_usa_hat^2
var_log_or_uk_hat  <- sd_log_or_uk_hat^2

var_delta_hat <- var_log_or_usa_hat + var_log_or_uk_hat
sd_delta_hat  <- sqrt(var_delta_hat)

ci_delta_hat <- delta_hat + c(-2,2)*sd_delta_hat


print(paste0("Diferencia log OR  (movilidad relativa) USA-UK:",  round(delta_hat,2) ))
```

```
## [1] "Diferencia log OR  (movilidad relativa) USA-UK:-0.32"
```

```r
print(paste0("SE diferencia log OR  (movilidad relativa) USA-UK:",  round(sd_delta_hat,2) ))
```

```
## [1] "SE diferencia log OR  (movilidad relativa) USA-UK:0.13"
```

```r
print(paste0("95% CI diferencia log OR  (movilidad relativa) USA-UK: [", round(ci_delta_hat[1],2),",",r
```

```
## [1] "95% CI diferencia log OR  (movilidad relativa) USA-UK: [-0.59,-0.05]"
```

Este resultado indica que USA presenta un nivel ligeramente más alto de movilidad social relativa respecto a UK.