Análisis de Datos Categóricos

Ayudantía 7

Felipe Olivares

Contenido

Bootstrap method

¿Que ocurre si tomamos una muestra de nuestra muestra?

Es decir, una muestra aleatoria de nuestra muestra sucesivamente y muchas veces. Lo que busca el *boostrap method* es estimar un subconjunto de nuestros datos para obtener una estimación. Luego tomamos otro subconjunto y así muchas veces vamos a repetir este ejercicio. Es decir, en vez de simular los datos (como en el caso de Montecarlo), tomemos una muestra repitiendo las estimaciones varias veces. Así tendremos un comportamiento de nuestro estimador.

Revisaremos los mismos datos de la ayudantía anterior para crear un modelo de regresión logística en base a predictores que ya hemos utilizado.

Bootstrap method

Para efecto de la regresión que modelaremos la función de un control violento de la protesta respecto de un control negociado. Por lo tanto, es importante que el control negociado de la protesta=0 y control violento de la protesta=1 (consideren lo dicho previamente respecto del sentido del efecto que queremos estimar)

Bootstrap method

Revisemos los datos nuevamente. . .

```
head(df2)
```

```
# # A tibble: 6 x 7
#   apolicial ppolicial educacion estudiantes disruptiva trabajadores participan~1
#       <dbl> <fct>         <dbl>       <dbl>      <dbl>        <dbl>         <dbl>
# 1         0 No                0           0          0            0            81
# 2         1 No                0           0          0            0            40
# 3         1 No                0           0          1            0            20
# 4         0 No                0           0          1            1          2000
# 5         1 No                0           0          0            0            60
# 6         1 No                0           0          1            1           150
# # ... with abbreviated variable name 1: participantes
```

Bootstrap method

Ajuste un modelo de regresión logística que predice el control violento de la protesta como función de distinas covariables: educación, estudiantes y tácticas disruptivas.

Formalmente:

$$\underbrace{\ln \frac{p_i}{1-p_i}}_{\text{logit}(p_i)} = \beta_0 + \beta_1 \text{demandas por educación}_i + \beta_2 \text{presencia de estudiantes}_i + \beta_3 \text{presencia de tácticas disruptivas}_i + \beta_4 \text{ N° de manifestantes}_i$$

Bootstrap method

La trasformación de probabilidades a odds es monotónica, si la probabilidad aumenta también lo hacen los odds, y viceversa. El rango de valores que pueden tomar los odds es de $(0, \infty+)$. Dado que el valor de una probabilidad está acotado entre (0,1)

donde:

- e^{β_k} está restringido al rango $[0, \infty+)$. Es una constante que "comprime" o amplifica las odds de éxito.
- Si $\beta_k < 0 \rightarrow (0 < e^{\beta_k} < 1)$. Es decir, un aumento en x_k está asociado con una reducción (multiplicativa) de las odds de éxito.
- Si $\beta_k = 0 \rightarrow (e^{\beta_k} = 1)$. Es decir, un cambio en x_k está asociado a un cambio nulo en las odds de éxito.
- Si $\beta_k > 0 \rightarrow (e^{\beta_k} > 1)$. Es decir, un aumento en x_k está asociado a aumento (multiplicativo) en de las odds de éxito.

Bootstrap method

```r
logit1 <- glm(apolicial ~ educacion + estudiantes + disruptiva + participantes, family = binomial(link=logit), data=df2)
summary(logit1)
```

```
#
# Call:
# glm(formula = apolicial ~ educacion + estudiantes + disruptiva +
#     participantes, family = binomial(link = logit), data = df2)
#
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -2.5818  -1.3051   0.7954   1.0546   1.2964
#
# Coefficients:
#                 Estimate Std. Error z value Pr(>|z|)
# (Intercept)   -2.754e-01  6.664e-02  -4.133 3.58e-05 ***
# educacion      1.899e-01  1.332e-01   1.426   0.1539
# estudiantes    5.029e-01  1.289e-01   3.901 9.58e-05 ***
# disruptiva     5.707e-01  7.964e-02   7.166 7.70e-13 ***
# participantes  5.007e-06  2.465e-06   2.031   0.0422 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
#     Null deviance: 3736.6  on 2732  degrees of freedom
# Residual deviance: 3621.2  on 2728  degrees of freedom
# AIC: 3631.2
#
# Number of Fisher Scoring iterations: 4
```

```r
#Estimación de GLM (Regresión logística)
 exp(summary(logit1)$coefficients[,1]) # en odds
```

```
#   (Intercept)     educacion   estudiantes    disruptiva participantes
#      0.759234      1.209140      1.653472      1.769523      1.000005
```

Bootstrap method

A partir del modelo estimado:

$$\frac{p_i}{1-p_i} = \beta_0 + \beta_{Educación} + \beta_{estudiantes} + \beta_{disruptiva} + \beta_{participantes} \ )$$

calcule un IC al 95% para efecto de estudiantes como "odds ratio", utilizando el Bootstrap Method.

Recuerde que el Bootstrap Method:

1. Asume que la distribución empírica del los datos refleja la distribución de probabilidad de las variables de interés.
2. A partir de la muestra obtenén una muestra aleatoria del mismo tamaño que la muestra original (N), con reemplazo: (y_b, X_b)
3. Regresiona y_b y X_b para obtener el estimate $\hat{\theta}_b$
4. Repite los pasos 2 y 3 un gran número de veces B.
5. El conjunto de B resultados obtenidos corresponde a la "Bootstrap distribution" del estimate.
6. Evalúa la distribución del estimate (SE,CI, etc) o de cualquier cantidad derivada de éste.

Bootstrap method

$$\frac{p_i}{1-p_i} = \beta_0 + \beta_{Educación} + \beta_{estudiantes} + \beta_{disruptiva} + \beta_{participantes} \ )$$

```
# Función de resampling y estimación de modelo
bs_expedad  <- function(x) {
  data_bs  <- sample_n(df2,size=nrow(df2),replace=TRUE) # tomar el tamaño original de la muestra que tenemos
  rl_bs <- glm(apolicial ~ educacion + estudiantes + disruptiva + participantes, family = binomial(link=logit), data=data_bs) #
  expbeta_bs <- exp(rl_bs$coefficients[3])#coeficientes de interés que queremos resamplear
  return(expbeta_bs)}

# Iterar función y almacenar resultados
nreps = 400 # cantidad de veces que queremos iterar
expbetas_bs <- replicate(nreps,bs_expedad()); head(expbetas_bs)
```

```
# estudiantes estudiantes estudiantes estudiantes estudiantes estudiantes
#    1.693759    1.402892    1.790975    1.818062    1.312842    2.000026
```

Bootstrap method

A partir del modelo estimado:

$$\frac{p_i}{1-p_i} = \beta_0 + \beta_{Educación} + \beta_{estudiantes} + \beta_{disruptiva} + \beta_{participantes}$$

calcular un IC al 95% para efecto de estudiantes como "odds ratio", utilizando el Bootstrap Method.

```
# Sintaxis
# Cálculo de Standard Errors en base a Bootstrap Distribution
se_expbeta_bs <- sd(expbetas_bs)
se_expbeta_bs
```

```
# [1] 0.2035926
```

```
# Sintaxis
# Cálculo de Intervalos de Confianza
ci_expbeta_bs <-
  quantile(expbetas_bs, p=c(0.025,0.975))
ci_expbeta_bs
```

```
#     2.5%    97.5%
# 1.296408 2.104390
```

Bootstrap method

Calcule un IC al 95% para el Average Marginal Effect de edad sobre la probabilidad de un manejo violento de la protesta, utilizando el Bootstrap Method.

Recordar...

El efecto marginal en la media (AME) se calcula ajustando los valores de todas las covariables a sus medias dentro de la muestra. Es decir, el AME es el efecto parcial de la variable dependiente (Y) condicionado a un regresor (X) después de establecer todas las demás covariables en sus medias. En otras palabras, el MEM es la diferencia en el efecto de X sobre Y cuando todas las demás covariables están en su media.

```
# Sintaxis
# Función de resampling y estimación de modelo
bs_ame_estudiantes  <- function(x) {
  data_bs  <- sample_n(df2,size=nrow(df2),replace=TRUE)
  rl_bs <- glm(apolicial ~ educacion + estudiantes + disruptiva + participantes, family = binomial(link=logit), data=data_bs)
  beta_bs <- rl_bs$coefficients[3]
  p_hat_b  <- predict(rl_bs, type = "response") #los valores predichos del modelo
  me_estudiantes_b   <- beta_bs*p_hat_b*(1-p_hat_b) # p * 1-p
  return(ame_estudiantes_b = mean(me_estudiantes_b))
}

# Iterar función y almacenar resultados
nreps = 400
ame_estudiantes_bs <- replicate(nreps,bs_ame_estudiantes()); head(ame_estudiantes_bs) #calcula, remuestrea y guarda cada vez
```

```
# [1] 0.05654276 0.15514096 0.13268682 0.13038410 0.08855421 0.13132620
```

Bootstrap method

Calcule un IC al 95% para el Average Marginal Effect de edad sobre la probabilidad de un manejo violento de la protesta por parte de las policías, utilizando el Bootstrap Method.**

```
# Sintaxis
# Cálculo de Standard Errors en base a Bootstrap Distribution
se_ame_estudiantes_bs <- sd(ame_estudiantes_bs)
se_ame_estudiantes_bs
```

```
# [1] 0.02833678
```

```
# Sintaxis
# Cálculo de Intervalos de Confianza
ci_ame_estudiantes_bs <-
  quantile(ame_estudiantes_bs, p=c(0.025,0.975))
  ci_ame_estudiantes_bs
```

```
#       2.5%      97.5%
# 0.06577821 0.17505938
```