

Mortalidad por cáncer de pulmón en Ohio (décadas de 1960 a 1980)

Maria Paula Camargo Rincon*Laura Katherine Martínez Castiblanco†

Contents

| | | |
|---|------------------------------------|---|
| 1 | Intriducción | 1 |
| 2 | Análisis descriptivo | 3 |
| 3 | Modelo y verificación de supuestos | 6 |
| 4 | Conclusiones | 8 |
| 5 | Revisión bibliográfica | 8 |
| 6 | Bibliografía | 8 |

1 Intriducción

Este conjunto de datos se deriva de un estudio de mortalidad por cáncer de pulmón en Ohio realizado entre 1969 y 1971 a nivel de condado. La base de datos **ohiolung** contiene información para los 88 condados del estado de Ohio, (EE. UU.). La base de datos proviene del GeoDa (Center for Geospatial Analysis and Computation) y presenta 42 variables. En las estas se presentan datos referentes a la mortalidad por cáncer de pulmón estratificada por género (masculino/femenino) y raza (blanca/negra), junto variables geográficas básicas.

La siguiente tabla resume las variables principales de la base de datos:

*mcamargori@unal.edu.co

†laumartinezca@unal.edu.co

Table 1: Definición de variables principales en la base de datos ohlung.shp

| Variable | Descripción |
|------------------|---|
| CountyID | ID secuencial del condado (orden alfabético) |
| NAME | Nombre del condado |
| FIPSNO | Código FIPS del condado (numérico) |
| AREA | Área del polígono (condado) |
| PERIMETER | Perímetro del polígono (condado) |
| RECORD_ID | ID único del registro |
| COUNTYID | ID del condado |
| LGRyy | Casos de cáncer de pulmón para género G (M/F) y raza R (W/B) en año yy (1968, 1978, 1988) |
| POPGRyy | Población en riesgo para género G y raza R en año yy |
| LGyy | Total de casos de cáncer de pulmón por género G en año yy |
| POPGyy | Población total en riesgo por género G en año yy |

Hay 12 variables por año (4 combinaciones de género-raza \times 3 tipos: casos detallados, población detallada, totales por género)

Table 2: Primeros 6 datos

| COUNTYID | NAME | FIPSNO | AREA | PERIMETER | LMW68 | POPMW68 | LMB68 |
|----------|----------|--------|------------|-----------|-------|---------|-------|
| 48 | Lucas | 39095 | 873382000 | 164533 | 128 | 205421 | 19 |
| 26 | Fulton | 39051 | 1054690000 | 134891 | 4 | 15521 | 0 |
| 28 | Geauga | 39055 | 1005750000 | 146204 | 4 | 30415 | 0 |
| 86 | Williams | 39171 | 1089180000 | 136740 | 5 | 16252 | 0 |
| 18 | Cuyahoga | 39035 | 1242480000 | 173664 | 435 | 677185 | 87 |
| 62 | Ottawa | 39123 | 634883000 | 128356 | 10 | 17638 | 0 |

| POPMB68 | LM68 | POPM68 | LFW68 | POPFW68 | LFB68 | POPFB68 | LF68 | POPF68 | LMW78 |
|---------|------|--------|-------|---------|-------|---------|------|--------|-------|
| 25956 | 147 | 231377 | 27 | 220106 | 1 | 28527 | 28 | 248633 | 149 |
| 50 | 4 | 15571 | 1 | 16358 | 0 | 47 | 1 | 16405 | 10 |
| 472 | 4 | 30887 | 3 | 31131 | 0 | 496 | 3 | 31627 | 11 |
| 24 | 5 | 16276 | 0 | 17216 | 0 | 20 | 0 | 17236 | 7 |
| 157046 | 522 | 834231 | 88 | 734014 | 23 | 176614 | 111 | 910628 | 442 |
| 164 | 10 | 17802 | 0 | 18430 | 0 | 182 | 0 | 18612 | 8 |

| POPMW78 | LMB78 | POPMB78 | LM78 | POPM78 | LFW78 | POPFW78 | LFB78 | POPFB78 |
|---------|-------|---------|------|--------|-------|---------|-------|---------|
| 194976 | 34 | 30252 | 183 | 225228 | 48 | 210307 | 4 | 33790 |
| 18268 | 0 | 75 | 10 | 18343 | 3 | 19080 | 0 | 70 |
| 35839 | 1 | 604 | 12 | 36443 | 3 | 36148 | 0 | 619 |
| 17678 | 1 | 44 | 8 | 17722 | 0 | 18126 | 0 | 70 |
| 567350 | 138 | 161461 | 580 | 728811 | 165 | 623107 | 36 | 186806 |
| 19756 | 0 | 170 | 8 | 19926 | 6 | 20388 | 0 | 188 |

| LF78 | POPF78 | LMW88 | POPMW88 | LMB88 | POPMB88 | LM88 | POPM88 | LFW88 |
|------|--------|-------|---------|-------|---------|------|--------|-------|
| 52 | 244097 | 150 | 185367 | 39 | 36104 | 189 | 221471 | 108 |
| 3 | 19150 | 8 | 18600 | 1 | 62 | 9 | 18662 | 2 |
| 3 | 36767 | 17 | 37759 | 1 | 1051 | 18 | 38810 | 12 |
| 0 | 18196 | 10 | 17822 | 0 | 61 | 10 | 17883 | 7 |
| 201 | 809913 | 453 | 499655 | 188 | 178432 | 641 | 678087 | 269 |
| 6 | 20576 | 16 | 19164 | 0 | 205 | 16 | 19369 | 9 |

| POPFW88 | LFB88 | POPFB88 | LF88 | POPF88 |
|---------|-------|---------|------|--------|
| 200458 | 8 | 41115 | 116 | 241573 |
| 19444 | 0 | 65 | 2 | 19509 |
| 37973 | 0 | 1040 | 12 | 39013 |
| 18506 | 0 | 110 | 7 | 18616 |
| 552388 | 83 | 207628 | 352 | 760016 |
| 20063 | 1 | 221 | 10 | 20284 |

2 Análisis descriptivo

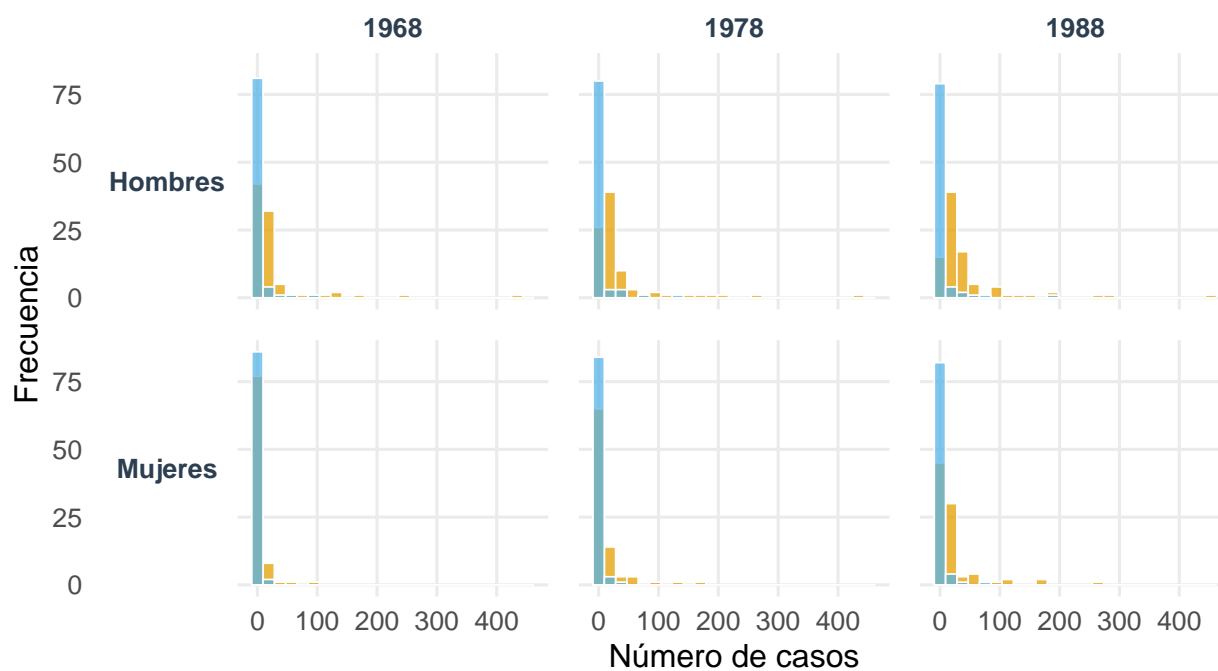
En esta sección se realiza una exploración exhaustiva de la base de datos con el fin de identificar relaciones entre las variables, detectar posibles transformaciones y establecer la estructura inicial del modelo de regresión lineal múltiple.

Table 7: Estadísticas descriptivas de casos y población

| Variable | Valor | Variable | Valor |
|-------------|--------|---------------|----------|
| LMW68 Media | 26.88 | LFB78 Min | 0.00 |
| LMW68 SD | 58.57 | LFB78 Max | 36.00 |
| LMW68 Min | 1.00 | LF78 Media | 13.78 |
| LMW68 Max | 435.00 | LF78 SD | 29.42 |
| LMB68 Media | 3.26 | LF78 Min | 0.00 |
| LMB68 SD | 11.75 | LF78 Max | 201.00 |
| LMB68 Min | 0.00 | LFW88 Media | 22.33 |
| LMB68 Max | 87.00 | LFW88 SD | 41.47 |
| LM68 Media | 30.14 | LFW88 Min | 0.00 |
| LM68 SD | 70.12 | LFW88 Max | 269.00 |
| LM68 Min | 1.00 | LFB88 Media | 2.60 |
| LM68 Max | 522.00 | LFB88 SD | 10.49 |
| LMW78 Media | 35.82 | LFB88 Min | 0.00 |
| LMW78 SD | 63.42 | LFB88 Max | 83.00 |
| LMW78 Min | 3.00 | LF88 Media | 24.93 |
| LMW78 Max | 442.00 | LF88 SD | 51.23 |
| LMB78 Media | 5.01 | LF88 Min | 1.00 |
| LMB78 SD | 17.79 | LF88 Max | 352.00 |
| LMB78 Min | 0.00 | POPMW68 Media | 52983.82 |

| | | | |
|-------------|--------|---------------|-----------|
| LMB78 Max | 138.00 | POPMW68 SD | 95322.08 |
| LM78 Media | 40.83 | POPMB68 Media | 5329.95 |
| LM78 SD | 80.55 | POPMB68 SD | 19460.53 |
| LM78 Min | 3.00 | POPM68 Media | 58313.77 |
| LM78 Max | 580.00 | POPM68 SD | 114110.27 |
| LMW88 Media | 43.01 | POPMW78 Media | 53411.91 |
| LMW88 SD | 68.72 | POPMW78 SD | 85016.85 |
| LMW88 Min | 2.00 | POPMB78 Media | 5915.57 |
| LMW88 Max | 453.00 | POPMB78 SD | 20732.39 |
| LMB88 Media | 6.22 | POPM78 Media | 59327.48 |
| LMB88 SD | 23.11 | POPM78 SD | 104851.91 |
| LMB88 Min | 0.00 | POPMW88 Media | 52407.83 |
| LMB88 Max | 188.00 | POPMW88 SD | 79338.55 |
| LM88 Media | 49.23 | POPMB88 Media | 6799.70 |
| LM88 SD | 90.40 | POPMB88 SD | 23382.85 |
| LM88 Min | 2.00 | POPM88 Media | 59207.53 |
| LM88 Max | 641.00 | POPM88 SD | 101568.08 |
| LFW68 Media | 6.12 | POPFW68 Media | 56094.52 |
| LFW68 SD | 12.68 | POPFW68 SD | 102927.37 |
| LFW68 Min | 0.00 | POPFB68 Media | 5862.83 |
| LFW68 Max | 88.00 | POPFB68 SD | 21878.07 |
| LFB68 Media | 0.62 | POPF68 Media | 61957.35 |
| LFB68 SD | 2.82 | POPF68 SD | 124072.32 |
| LFB68 Min | 0.00 | POPFW78 Media | 56682.57 |
| LFB68 Max | 23.00 | POPFW78 SD | 92507.22 |
| LF68 Media | 6.75 | POPFB78 Media | 6638.98 |
| LF68 SD | 15.32 | POPFB78 SD | 23892.41 |
| LF68 Min | 0.00 | POPF78 Media | 63321.55 |
| LF68 Max | 111.00 | POPF78 SD | 115408.59 |
| LFW78 Media | 12.45 | POPFW88 Media | 55763.84 |
| LFW78 SD | 24.74 | POPFW88 SD | 86678.27 |
| LFW78 Min | 0.00 | POPFB88 Media | 7650.48 |
| LFW78 Max | 165.00 | POPFB88 SD | 27083.88 |
| LFB78 Media | 1.33 | POPF88 Media | 63414.32 |
| LFB78 SD | 4.85 | POPF88 SD | 112456.93 |

Distribución de casos de cáncer de pulmón por raza y género



Distribución de población por raza y género

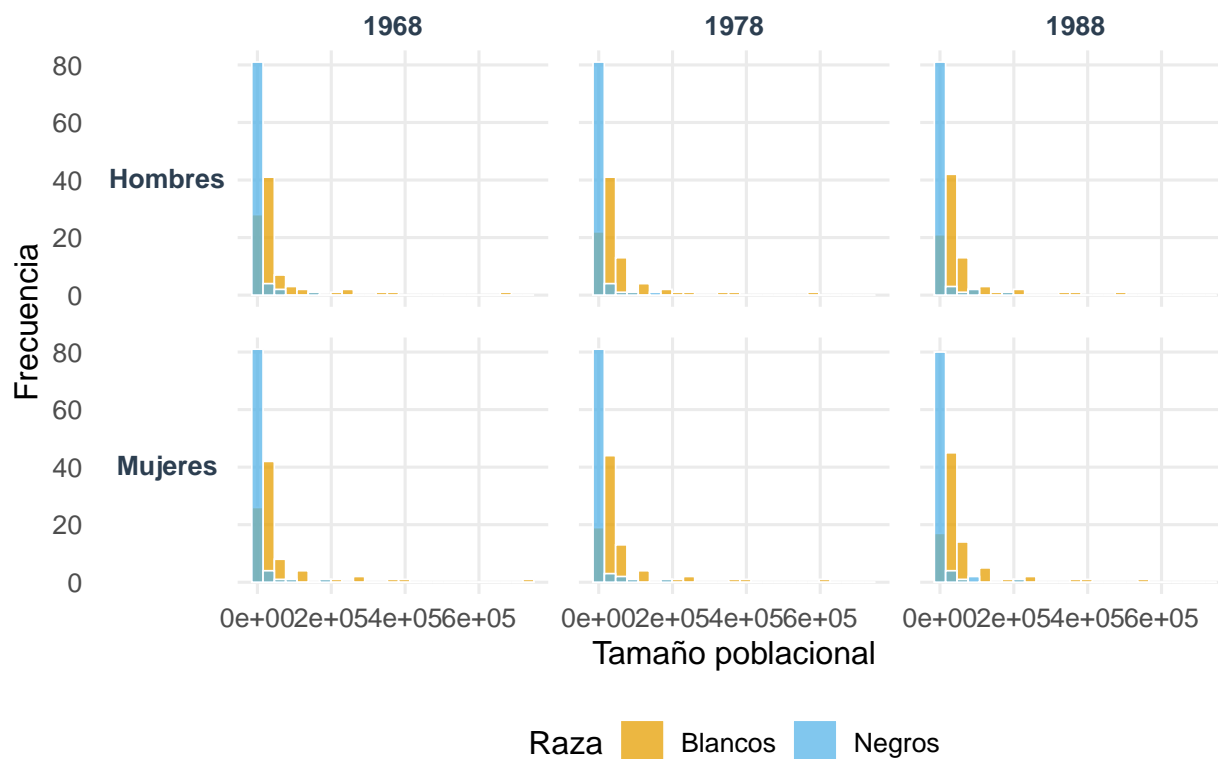
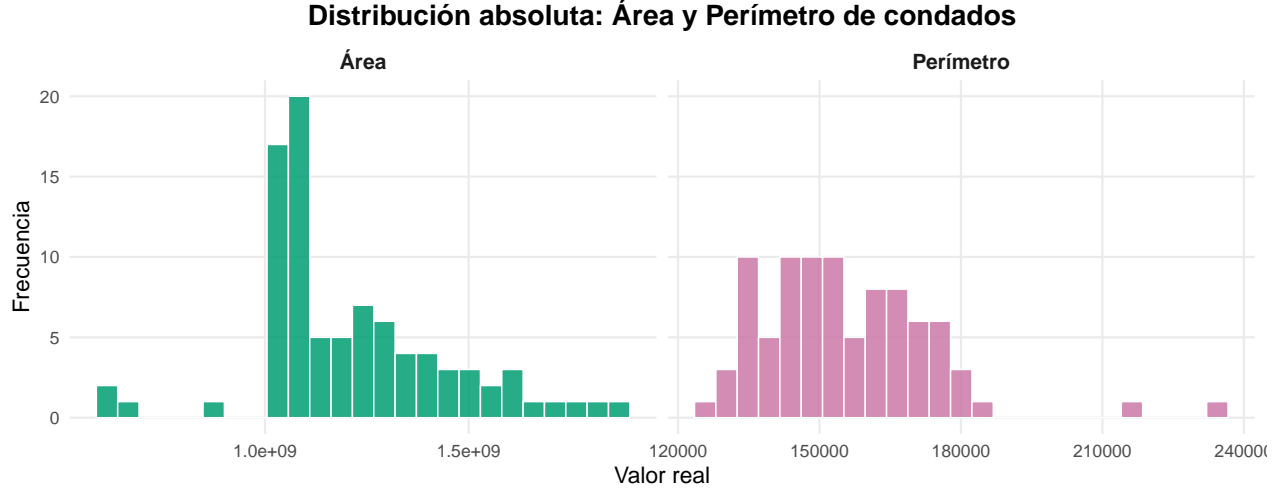


Table 8: Correlaciones parciales (controlando población masculina)

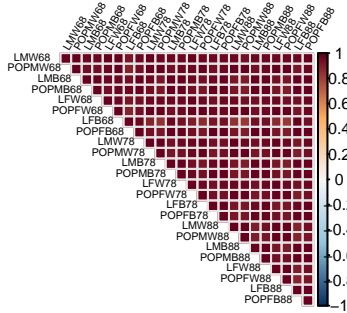
| Relación | Correlación_parcial |
|---------------------------|---------------------|
| LM68 ~ AREA POPM68 | -0.151 |
| LM68 ~ PERIMETER POPM68 | 0.143 |



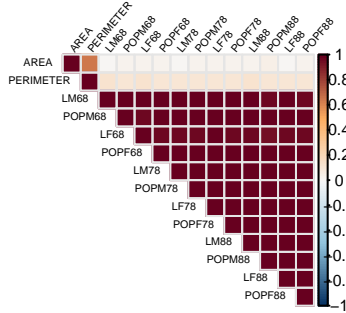
Los histogramas muestran que las tasas tienen distribución asimétrica positiva (típica de conteos).

Matrices de correlación

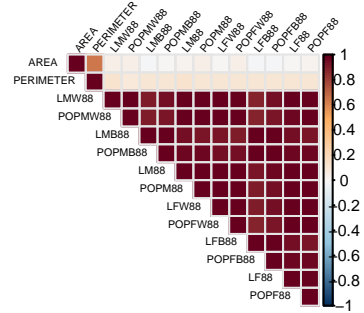
**Género x Raza
(1968–1988)**



**Género + Geografía
(1968–1988)**



**Género x Raza + Geografía
(1988)**



3 Modelo y verificación de supuestos

Para cumplir con el requisito de incluir al menos una variable cualitativa, reorganizamos los datos a formato largo con base en las variables originales. Esto genera las variables

cualitativas **gender** (Masculino/Femenino) y **year** (1968/1978/1988) a partir de los nombres de las columnas existentes.

Table 9: Vista de los datos en formato largo

| COUNTYID | NAME | AREA | PERIMETER | gender | year | cases | pop |
|----------|--------|------------|-----------|--------|------|-------|--------|
| 48 | Lucas | 873382000 | 164533 | Male | 1968 | 147 | 231377 |
| 48 | Lucas | 873382000 | 164533 | Female | 1968 | 28 | 248633 |
| 48 | Lucas | 873382000 | 164533 | Male | 1978 | 183 | 225228 |
| 48 | Lucas | 873382000 | 164533 | Female | 1978 | 52 | 244097 |
| 48 | Lucas | 873382000 | 164533 | Male | 1988 | 189 | 221471 |
| 48 | Lucas | 873382000 | 164533 | Female | 1988 | 116 | 241573 |
| 26 | Fulton | 1054690000 | 134891 | Male | 1968 | 4 | 15571 |
| 26 | Fulton | 1054690000 | 134891 | Female | 1968 | 1 | 16405 |
| 26 | Fulton | 1054690000 | 134891 | Male | 1978 | 10 | 18343 |
| 26 | Fulton | 1054690000 | 134891 | Female | 1978 | 3 | 19150 |

```
##
## Call:
## lm(formula = cases ~ AREA + PERIMETER + gender + year + pop,
##     data = ohlung_long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -302.741   -9.193   -0.937    9.265   293.559
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.308e+01  1.209e+01  -2.736  0.00643 **
## AREA        -9.100e-09  7.592e-09  -1.199  0.23124
## PERIMETER    1.312e-04  9.966e-05   1.316  0.18862
## genderMale   2.680e+01  2.726e+00   9.831 < 2e-16 ***
## year1978     8.295e+00  3.338e+00   2.485  0.01326 *
## year1988     1.807e+01  3.338e+00   5.415 9.36e-08 ***
## pop          4.781e-04  1.228e-05  38.928 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.31 on 521 degrees of freedom
## Multiple R-squared:  0.7613, Adjusted R-squared:  0.7586
## F-statistic: 277 on 6 and 521 DF, p-value: < 2.2e-16

##
## Call:
```

```
## lm(formula = cases ~ gender + year + pop, data = ohlung_long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -301.984   -9.110   -1.014    9.029   294.680
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.380e+01  2.828e+00  -8.415 3.77e-16 ***
## genderMale   2.680e+01  2.726e+00   9.834 < 2e-16 ***
## year1978     8.293e+00  3.338e+00   2.485  0.0133 *
## year1988     1.807e+01  3.338e+00   5.415 9.36e-08 ***
## pop          4.796e-04  1.220e-05  39.314 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.31 on 523 degrees of freedom
## Multiple R-squared:  0.7605, Adjusted R-squared:  0.7586
## F-statistic: 415.1 on 4 and 523 DF,  p-value: < 2.2e-16
```

4 Conclusiones

Aquí solo incluir lo que sugiere el consultante si lo hizo, de lo contrario dejar en blanco

5 Revisión bibliográfica

Fuentes propuestas por los consultantes:

- @gareth2021introduction
- @pena2013analysis
- @schoenberg1935remarks
- @corradino1990proximity

Fuentes consultadas por el grupo:

- @bib2
- @lapointe1994classification

6 Bibliografía