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

Maria Paula Camargo Rincon*Laura Katherine Martínez Castiblanco†

Contents

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

*mcamargori@unal.edu.co
†laumartinezca@unal.edu.co

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:

Variable	Descripcion
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

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

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

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

Tabla 2. Primeros 6 datos de la base de datos original.

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 de la base de datos con tal de identificar relaciones entre las variables y establecer la estructura inicial del modelo de regresión lineal múltiple.

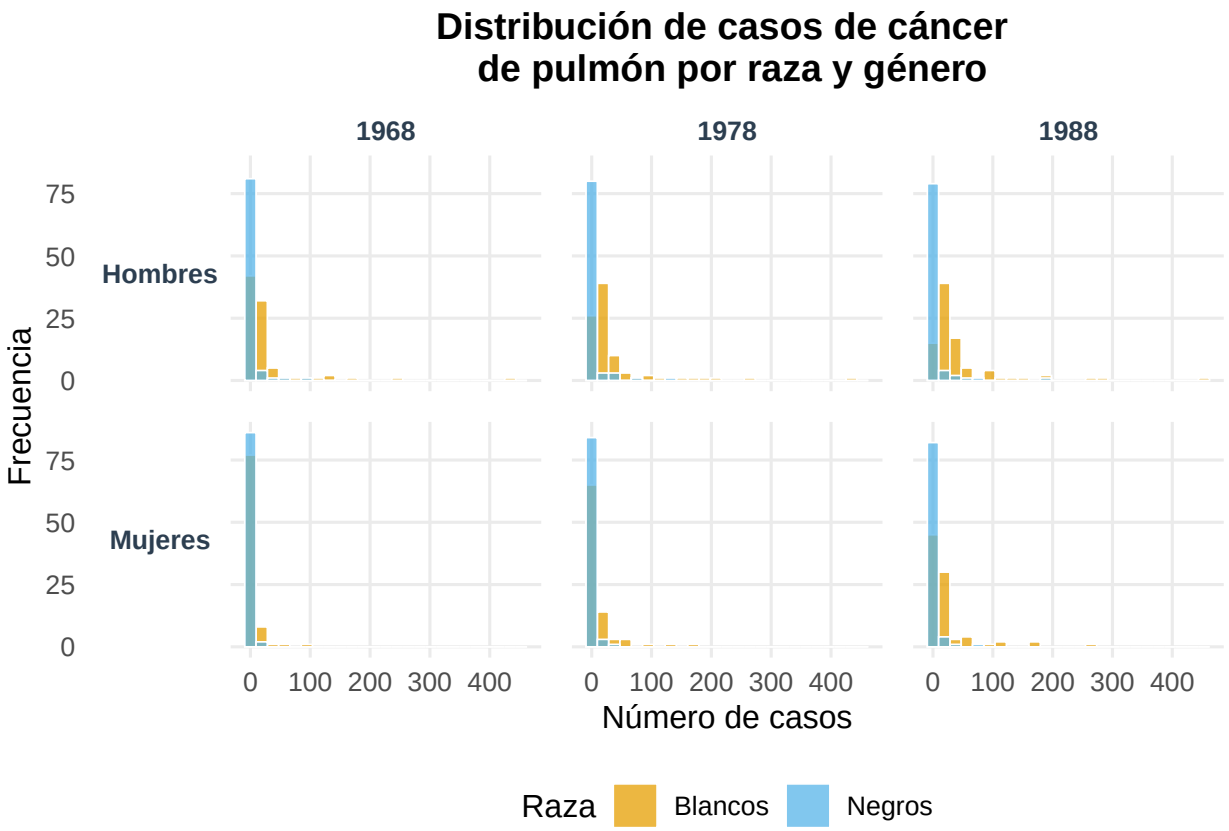
La Tabla 3 se presenta los principales estadísticos descriptivos de los casos de mortalidad por cáncer de pulmón y las poblaciones en riesgo. Se observa una alta variabilidad en el número de casos, con desviaciones estándar frecuentemente superiores a las medias. Los promedios de casos son consistentemente más altos en hombres que en mujeres y en la población blanca respecto a la negra. Además, se evidencia un incremento temporal en el número promedio de casos entre 1968 y 1988.

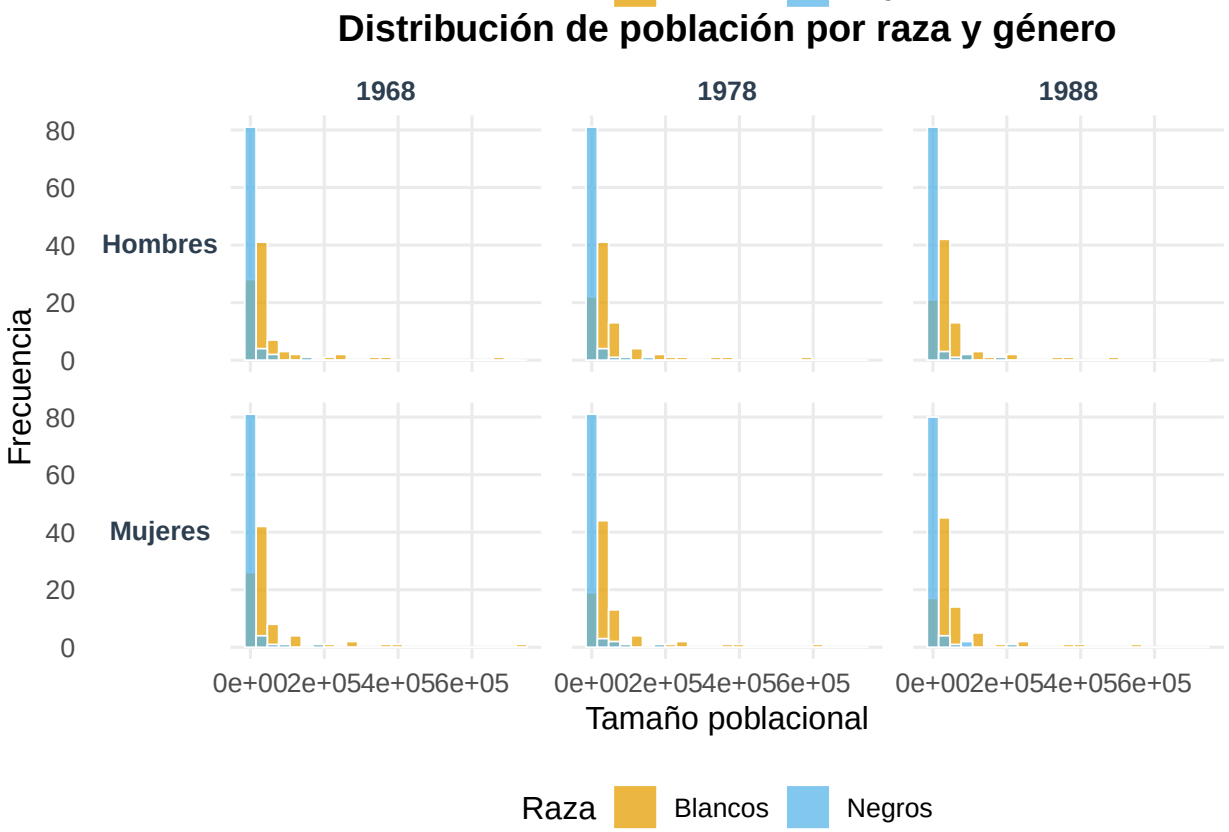
Variable	Valor	Variable	Valor
LMW68 Media	26.88	LFB78 Min	0.00
LMW68 SD	58.57	LFB78 Max	36.00
LMW68 Min	1.00	LF78 Media	13.78
LMW68 Max	435.00	LF78 SD	29.42
LMB68 Media	3.26	LF78 Min	0.00
LMB68 SD	11.75	LF78 Max	201.00
LMB68 Min	0.00	LFW88 Media	22.33
LMB68 Max	87.00	LFW88 SD	41.47
LM68 Media	30.14	LFW88 Min	0.00
LM68 SD	70.12	LFW88 Max	269.00

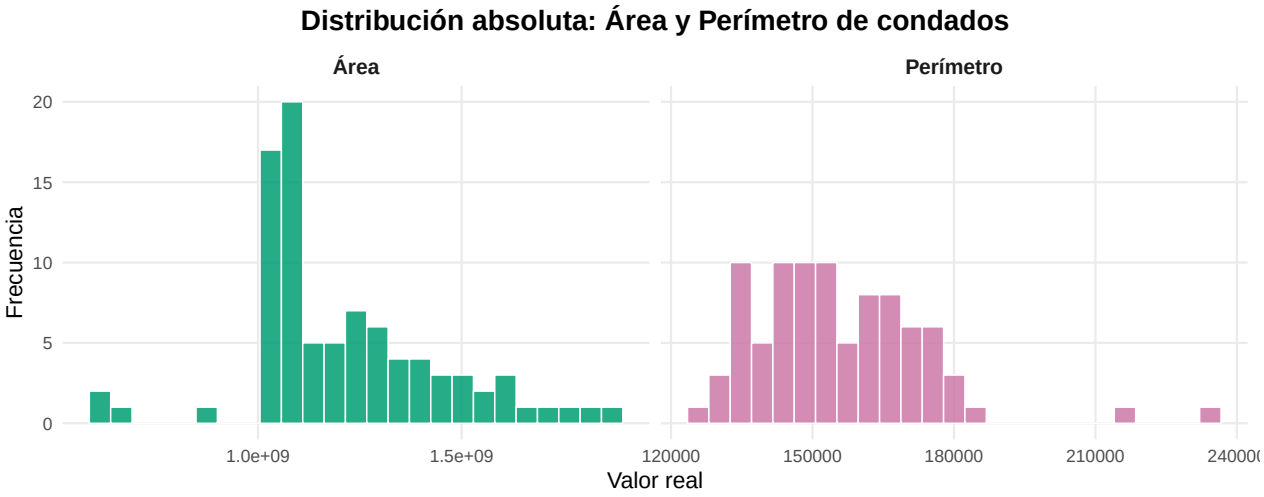
LM68 Min	1.00	LFB88 Media	2.60
LM68 Max	522.00	LFB88 SD	10.49
LMW78 Media	35.82	LFB88 Min	0.00
LMW78 SD	63.42	LFB88 Max	83.00
LMW78 Min	3.00	LF88 Media	24.93
LMW78 Max	442.00	LF88 SD	51.23
LMB78 Media	5.01	LF88 Min	1.00
LMB78 SD	17.79	LF88 Max	352.00
LMB78 Min	0.00	POPMW68 Media	52983.82
LMB78 Max	138.00	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

Tabla 3. Estadísticas descriptivas de casos y población.

Evaluando las distribuciones, se observa una asimetrica a la derecha en cada grupo de la problación evaluada, dicionalmente se evidencia que los condados más poblados tienden a presentar menos riesgo de cáncer de pulmón, muy seguramente por la separacion de zonas urbanas y rurales.


Distribución de casos de cáncer
de pulmón por raza y género
1968
1978
1988
Hombres
Mujeres
Frecuencia
75
50
25
0
0
100
200
300
400
Número de casos
Raza
Blancos
Negros

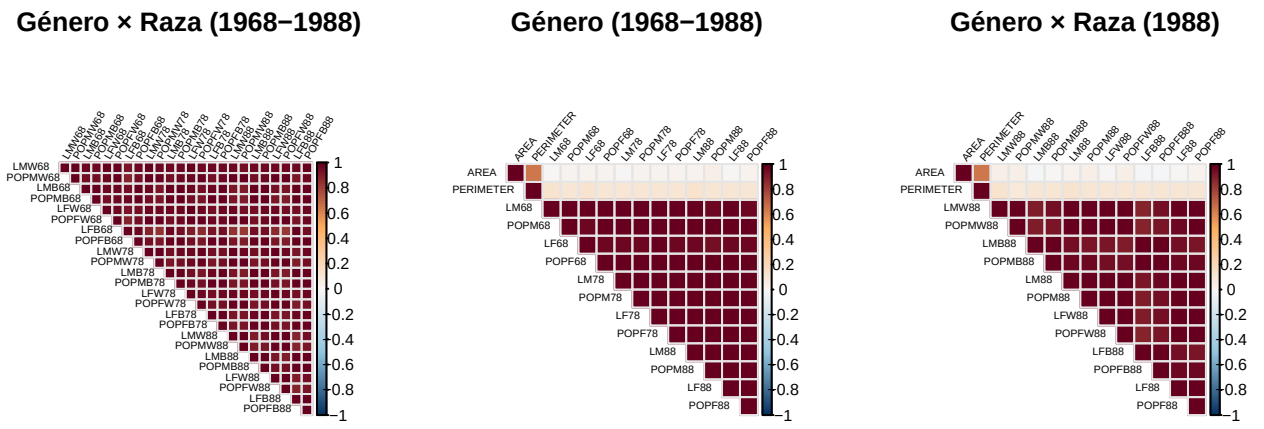

Distribución de población por raza y género
1968
1978
1988
Hombres
Mujeres
Frecuencia
80
60
40
20
0
0e+00
2e+05
4e+05
6e+05
Tamaño poblacional
Raza
Blancos
Negros



Las matrices de correlación (presentadas como las combinaciones de : género y raza en los 3 años, genero femenino en los 3 año y género y raza únicamente en el último año) muestran una fuerte correlación positiva entre el número de casos y la población en riesgo correspondiente, tanto en hombres como en mujeres. Asimismo, se observa una correlación leve positiva entre los casos y las variables geográficas `AREA` y `PERIMETER`, aunque estas dos últimas están también correlacionadas entre sí.

A su vez se observa que las variables geográficas aportan información limitada a la hora de explicar el comportamiento de cualquera de las variables aya sea de numero de casos o de población en riesgo de cáncer pulmonar.

Matrices de correlación



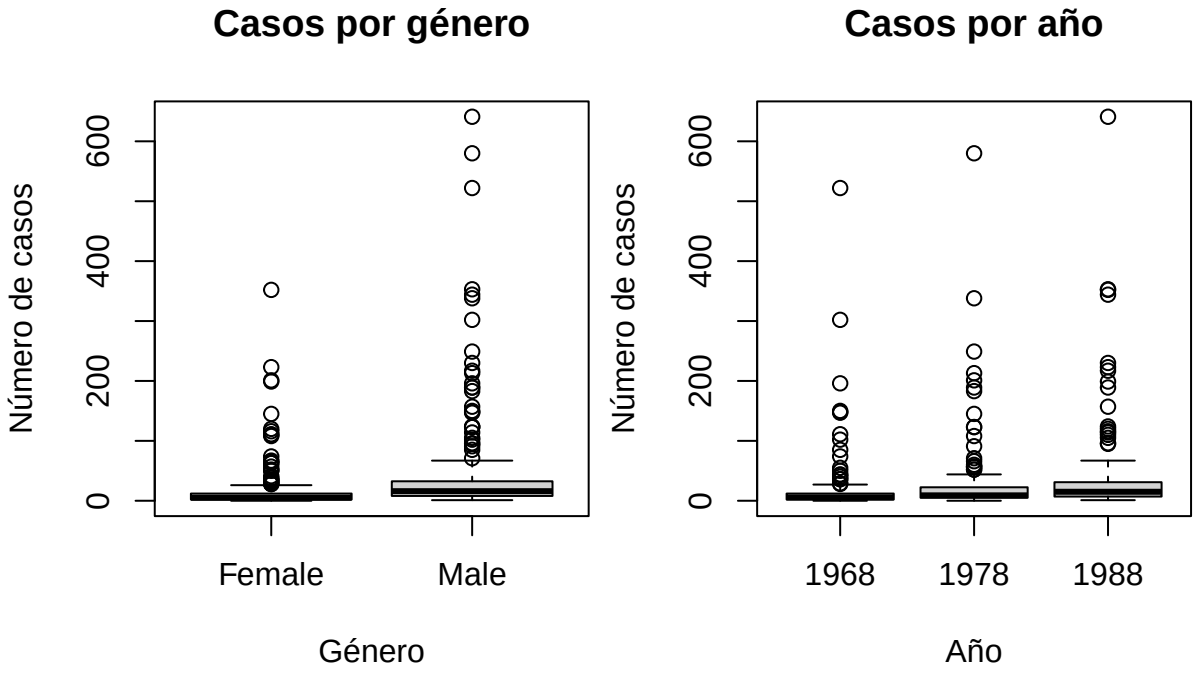
Por otro lado es de destacar que se observa una alta correlacion lineal en el tiempo, luego es recomendable añadir un componente que explique esta temporalidad al modelo, o evaluar la base de datos para tratar el año como una variable descriptiva para un modelo lineal multiple.

Para cumplir con el requisito de incluir al menos una variable cualitativa, y teniendo en cuenta las observaciones anteriores, se reestructuraron los datos al formato largo utilizando exclusivamente las variables originales de la base de datos (es decir se reorganizó la base de datos, sin alterar los datos originales). De esta forma se generaron las variables categóricas `gender` (Masculino/Femenino) y `year` (1968, 1978, 1988).

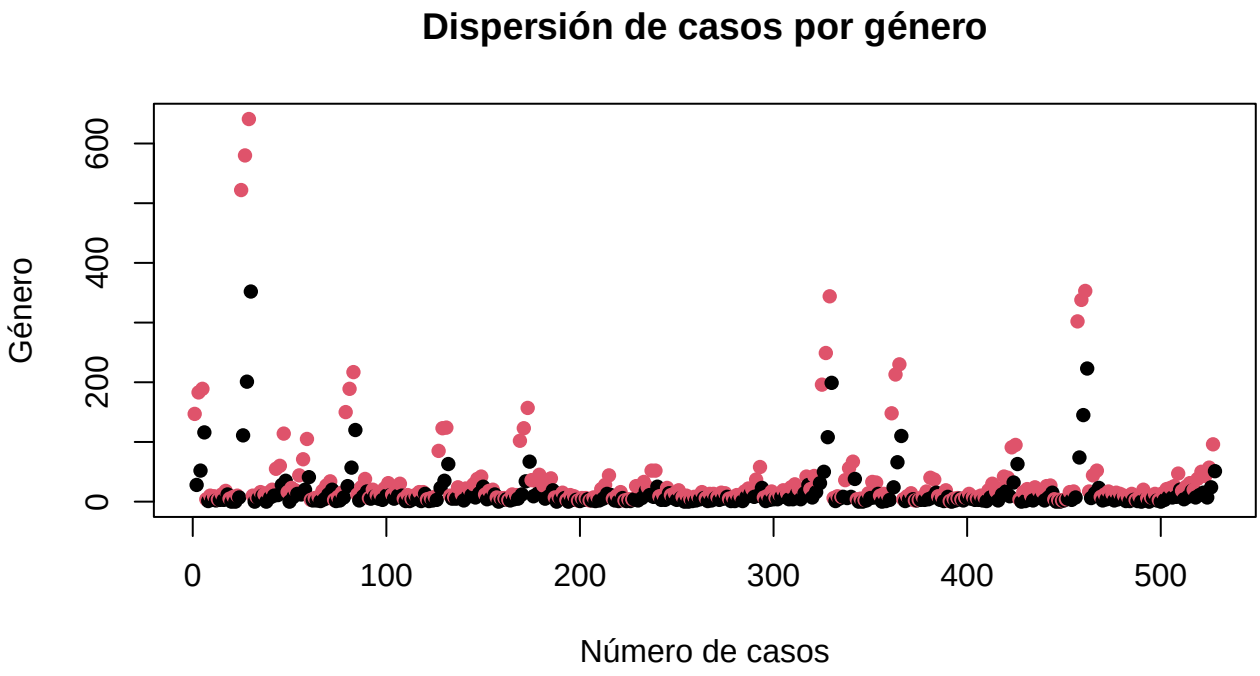
Table 8: Vista de los datos en formato largo

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

Considerando la nueva estructura de los datos se realiza un análisis adicional sobre la variable respuesta `cases` para evaluar su comportamiento en función del género y el año.



Se observa una precencia de datos atípicos por años, mas dispersión conforme pasan los años, sin embargo la media parece mantenerse.Más dispersos en hombres que mujeres pero media constante, más datos atípicos por hombres que por mujeres.



---

3 Modelo y verificación de supuestos

Considerando el aálisis anterior se plantea el siguiente modelo:

$$\text{cases} = \beta_0 + \beta_1 \text{AREA} + \beta_2 \text{PERIMETER} + \beta_3 \text{genderMale} + \beta_4 \text{year1978} + \beta_5 \text{year1978} + \beta_6 \text{pop} + \epsilon$$

donde `cases` corresponde a las variables `LMyy` o `LFyy` según el género y año, y `pop` es la población en riesgo correspondiente.

Los resultados de este modelo muestran que la población (`pop`) es altamente significativa y tiene el mayor impacto, como era esperado. El género masculino se asocia con un promedio de aproximadamente 26.8 casos adicionales respecto al femenino (controlando por las demás variables). Además, se observa un aumento significativo en el número de casos en 1978 y 1988 respecto a 1968.

```
##
## 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
```

Las variables `AREA` y `PERIMETER` no resultan estadísticamente significativas. Un modelo reducido sin estas variables geográficas presenta un R^2 prácticamente idéntico (0.7605 vs 0.7613), lo que confirma su escaso aporte y se propone en cambio el siguiente modelo reducido:

$$\text{cases} = \beta_0 + \beta_1 \text{genderMale} + \beta_2 \text{year1978} + \beta_3 \text{year1988} + \beta_4 \text{pop} + \epsilon$$

```
##
## 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
```

Además de esto, hemos realizado regresión por ridge y lasso mostradas a continuación:

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept) -1.662474e+03
## COUNTYID    -4.783533e-02
## NAME         .
## AREA        -7.848866e-09
## PERIMETER    1.378861e-04
## gender       .
## year         8.359870e-01
## pop          4.382109e-04
```

La regresión ridge mostro que las variables que tienen un impacto casi nulo en la variable respuesta son: `gender` y `NAME`.

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept) -4.988451e+02
## COUNTYID     .
## NAME         .
## AREA         .
## PERIMETER    .
## gender       .
## year         2.529125e-01
## pop          4.299486e-04
```

Por otro lado el modelo de regresión lasso indica que las variables que son influyentes son: 'year' y 'pop'. Por lo que en el modelo final se ha tomado la decisión de tener como variables explicativas: `gender`, `year` y `pop`.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon \text{cases} = \beta_0 + \beta_1 \text{genderMale} + \beta_2 \text{year1978} + \beta_3 \text{year1988} + \beta_4 \text{pop} + \epsilon$$

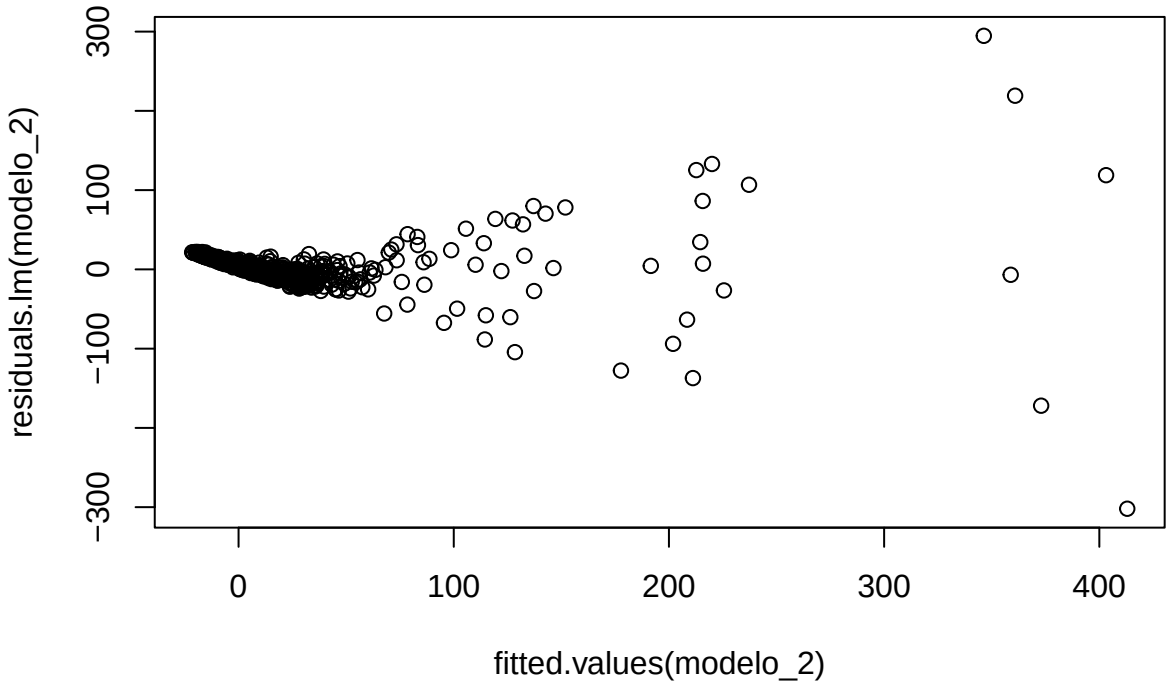
donde:

- x_{i1} es la variable dummy referente al género masculino, es decir, 1 si es masculino 0 en otro caso.
- x_{i2} es la variable dummy referente al año de 1978.
- x_{i3} es la variable dummy referente al año de 1988.
- x_{i4} es la variable númerica de la población en riesgo.

4 Verificación de supuestos

Para un primer avistamiento sobre la validación del modelo veamos el comportamiento de los residuales:

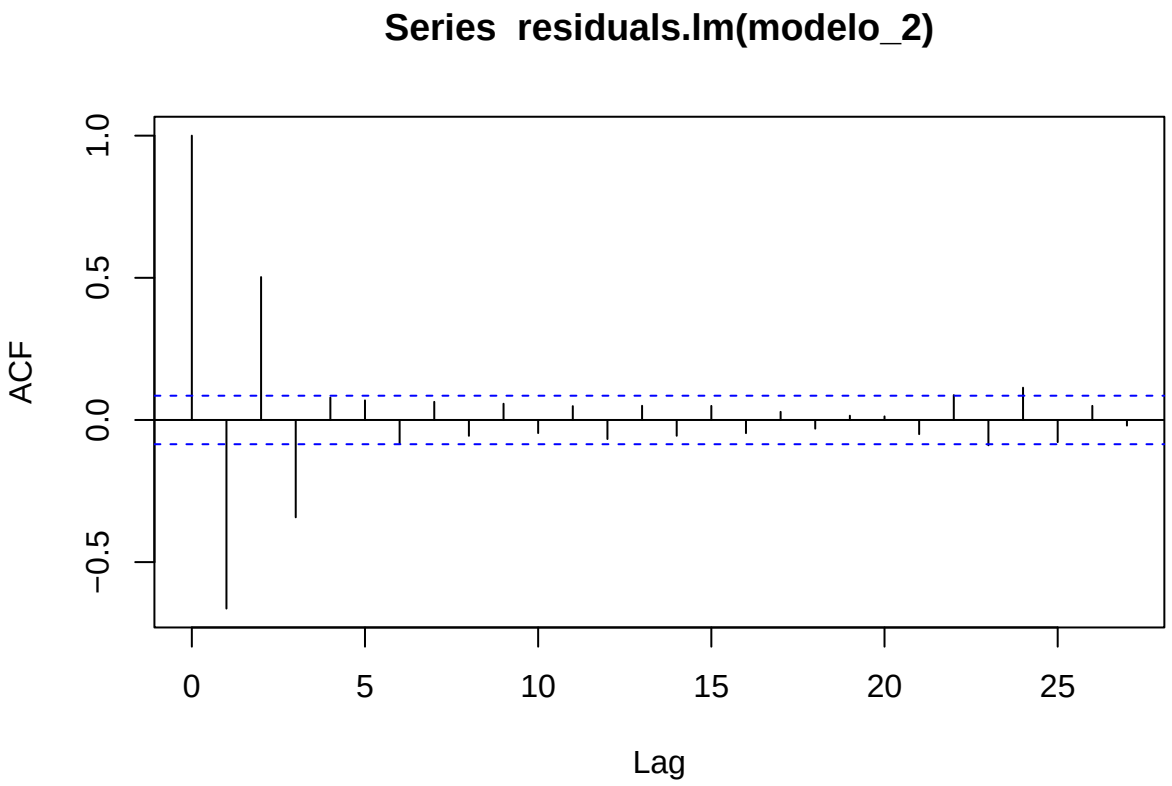
```
plot(fitted.values(modelo_2),residuals.lm(modelo_2))
```



Podemos ver que hay un problema evidente de heteroscedastiidad que más adelante será confirmado con las respectivas pruebas.

4.1 *Independencia:*

```
acf(residuals.lm(modelo_2))
```



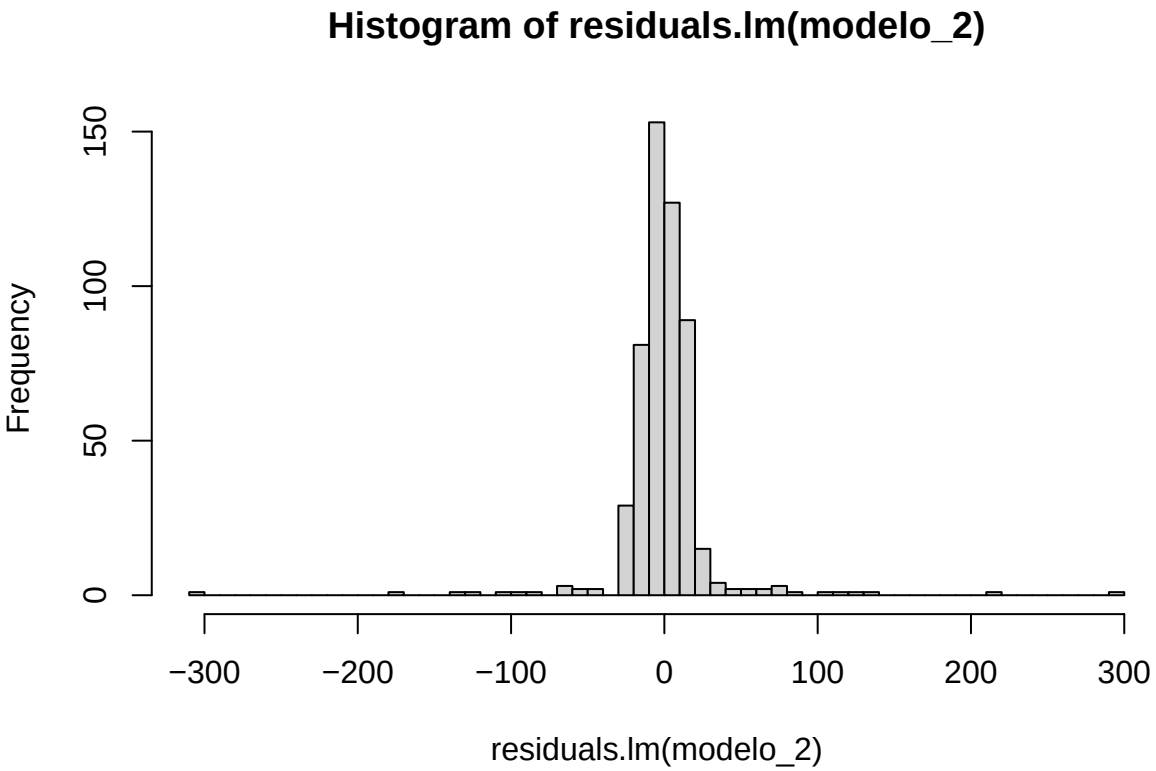
Parecen haber algunas correlaciones entre los residuales.

```
dwtest(modelo_2)
```

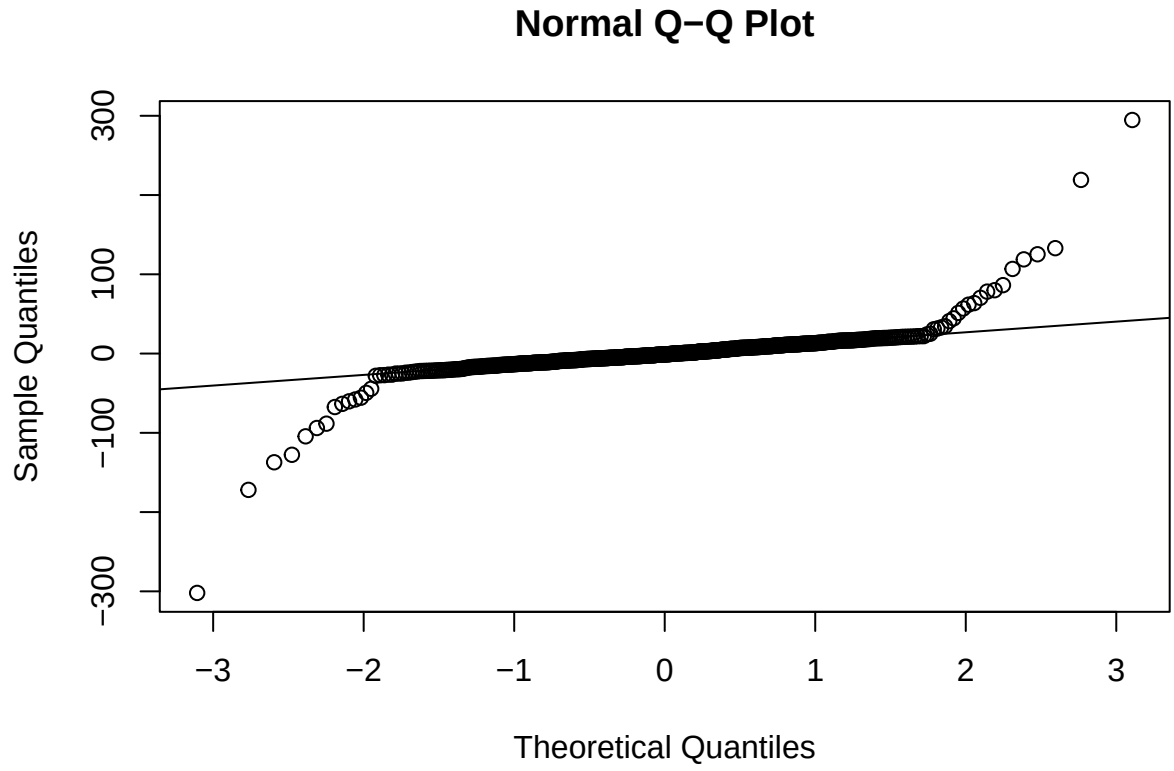
```
##
##  Durbin-Watson test
##
## data:  modelo_2
## DW = 3.3245, p-value = 1
## alternative hypothesis: true autocorrelation is greater than 0
```

Debido a que su p valor fue mayor al nivel de significancia (0.05), nos indica que no se rechaza H_0 no hay autocorrelación en los residuales, aunque el ACF sugiere algunas correlaciones en ciertos rezagos, la prueba de Durbin–Watson no encuentra evidencia estadística suficiente para rechazar la hipótesis de independencia al 5%. ## *Normalidad:*

```
hist(residuals.lm(modelo_2),breaks = 50)
```




```
qqnorm(residuals.lm(modelo_2))
qqline(residuals.lm(modelo_2))
```



Con este histograma no podemos asegurar que la distribución de los residuales sea una normal,

sin embargo se ve que es una distribución simétrica, y con media en 0. Por otro lado el qqplot nos muestra que puede presentar problemas de normalidad.

```
shapiro.test(residuals.lm(modelo_2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals.lm(modelo_2)
## W = 0.59974, p-value < 2.2e-16
```

Con un p valor menor al nivel de significancia (0.05) se rechaza la H_0 la población viene de una distribución normal. ## *Hetercedasticidad:*

Para evaluar la heteroscedasticidad formalmente, tomando en cuenta que la población no sigue o proviene de una distrincuón normal, se decidió aplicar el test de Breusch-pagan.

```
bptest(modelo_2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_2
## BP = 240.36, df = 4, p-value < 2.2e-16
```

4.2 Con un p valor menor al nivel de significancia (0.05) se rechaza la H_0 igualdad de varianzas, es algo que ya se podría detectar a simple vista en el gráfico de los residuales y los valores predichos.

5 Concluciones

- La regre

6 Revisión bibliográfica

Fuentes propuestas por los consultantes:
- @gareth2021introduction
- @pena2013analisis
- @schoenberg1935remarks
- @corradino1990proximity

Fuentes consultadas por el grupo:
- @bib2
- @lapointe1994classification

---

7 Bibliografía