Tasa de criminalidad en USA

Modelos Líneales

Trabajo final

Ignacio Acosta - Sofía Itté - Mauro Loprete

1er semestre 2021

Índice

Índice de figuras

1. Introducción

El objetivo de este informe es la construcción de un modelo de regresión lineal múltiple que explique la tasa de criminalidad en USA (número se ofensas reportadas a la policia por habitante).

Para ello se hará uso de una base de datos con un conjunto de variables que en principio se encuentran relacionadas con la variable a explicar.

Haciendo uso de las distintas técnicas estadíticas aprendidas en el curso se buscará descartar variables cuyo aporte no sea suficientemente significativo. Esto busca llegar a un modelo final eficiente (es decír, que explique la tasa de criminalidad de manera acertada haciendo uso de la menor cantidad de variables posibles).

En el transcurso del texto se pondrán a prueba las distintas hipótesis centrales del modelo, tales como la normalidad de los errores y la heterosedasticidad.

También se trabajará con las observaciones y la existencia de algunas que aporten el mismo nivel de información.

El herramental gráfico juega un rol fundamental al momento de transmitir la información de manera concisa y entendible. El mismo se encuentra respaldado por tablas que resumen la información de manera más detallada.

Cabe destacar que las observaciones se corresponden con los distintos estados de Estado Unidos.

De manera general se muestra el comportamiento de **y** de manera resumida:

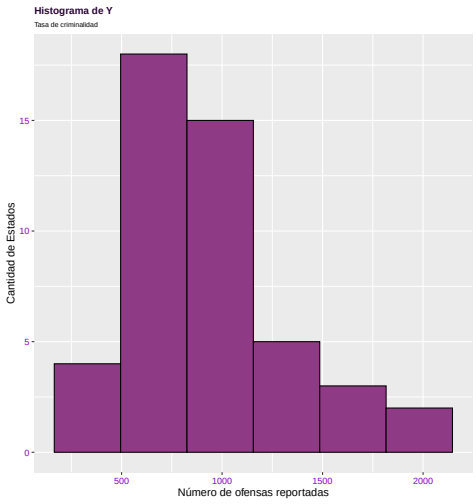


Figura 1: Histograma de la Tasa de Criminalidad

Es claro que **y** cuenta con una distribución medianamente asimétrica, tiene un intervalo modal entre los valores de 500 y 800 ofensas (13 estados presentan una tasa de criminalidad comprendida entre esos valores).

2. Análisis Exploratorio de datos

El objetivo de esta sección es presentar las variables a estudiar y como las mismas se relacionan entre sí.

Para ello se hará uso de distintas medidas de resumen univariadas y bivariadas, así como también un herramental gráfico variado que simplificará el entendimiento de las mismas.

Es esta sección fundamental al momento de discutir el modelo final y como a partir de distintas técnicas estadísticas aprendidas en el curso se puede simplificar el *modelo completo* que se presentará en la sección siguiente.

2.1. Análisis Univariado

En esta primer sección se hará especial enfásis en las variables por sí mismas.

Se estudiarán medidas de resumen y a partir de histogramas tendremos un primer acercamiento a la distribución de las mismas y su comportamiento.

Nombre	Descripción	Clasificación
Y	Tasa de criminalidad, número de ofensas reportadas a la policía por habitante	Cuantitativa
M	Número de hombres entre 14 y 24 años cada 1000 habitantes	Cuantitativa
So	Variables indicadora de los estados del sur (0=No, 1=Si)	Cualitativa
Ed	Indice que refeleja la escolaridad del estado	Cuantitativa
Po1	Gasto per cápita en policía realizado por el gobierno estatal o local en 1960	Cuantitativa
Po2	Gasto per cápita en policía realizado por el gobierno estatal o local en 1959	Cuantitativa
LF	Tasa de participación en la fuerza laboral civil de sexo masculino entre 14 y 24 años, cada 1000 habitantes	Cuantitativa
M.F	Número de hombres por cada 1000 mujeres	Cuantitativa
Pop	Tamaño de la población del estado cada 100000 habitantes	Cuantitativa
NW	Número de no caucásicos cada 1000 habitantes	Cuantitativa
U1	Tasa de desempleo urbana de hombres entre 14 y 24 años por 1000 habitantes	Cuantitativa
U2	Tasa de desempleo urbana de hombres entre 35 y 39 años por 1000 habitantes	Cuantitativa
GDP	Producto bruto interno per cápita	Cuantitativa
Ineq	Desigualdad del ingreso	Cuantitativa
Prob	Probabilidad de encarcelamiento	Cuantitativa
Time	Tiempo promedio de estadía en cárceles estatales	Cuantitativa

Cuadro 1: Variables a trabajar

2.1.1. Histogramas y Barplots

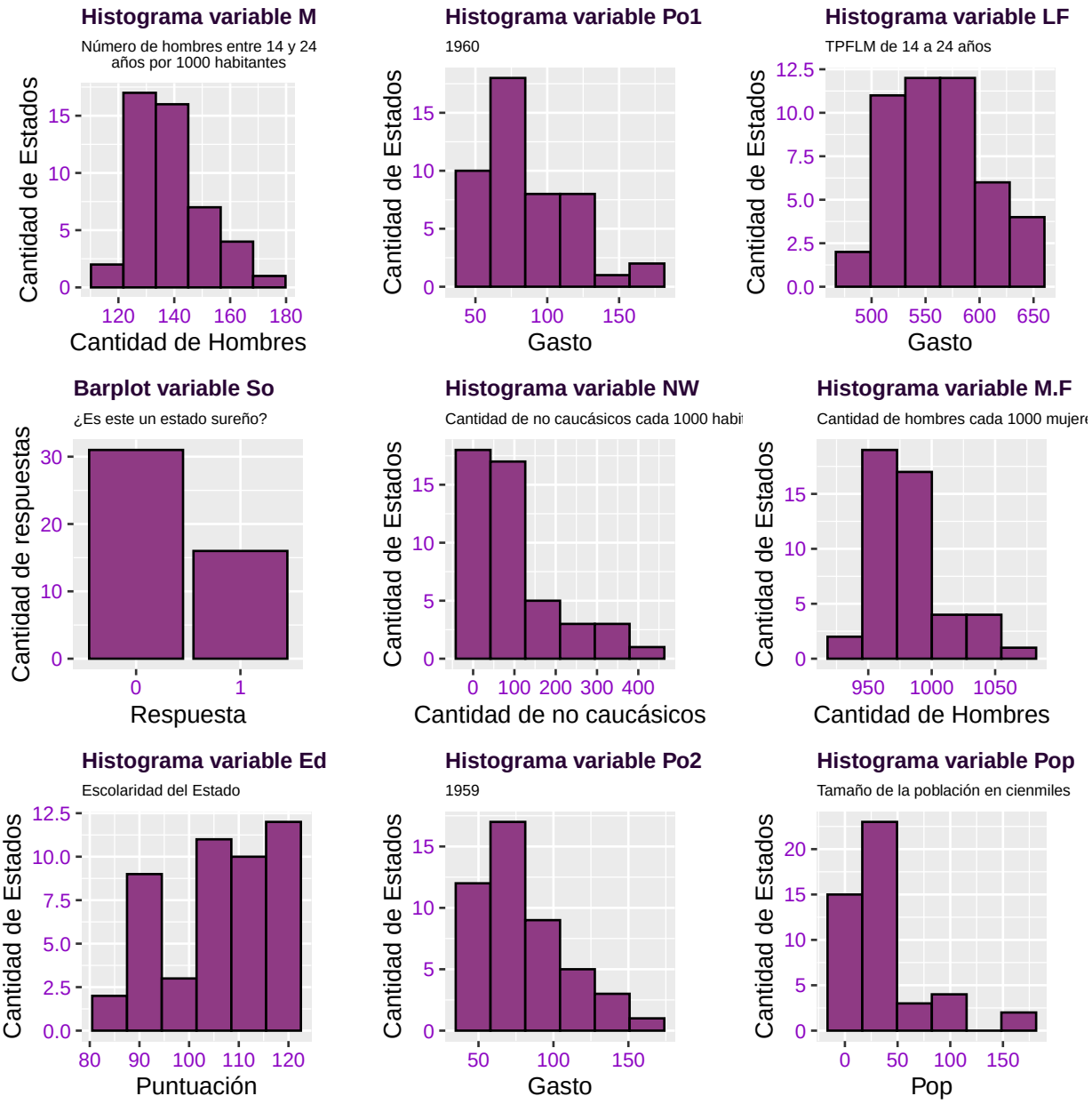


Figura 2: Histogramas (1)

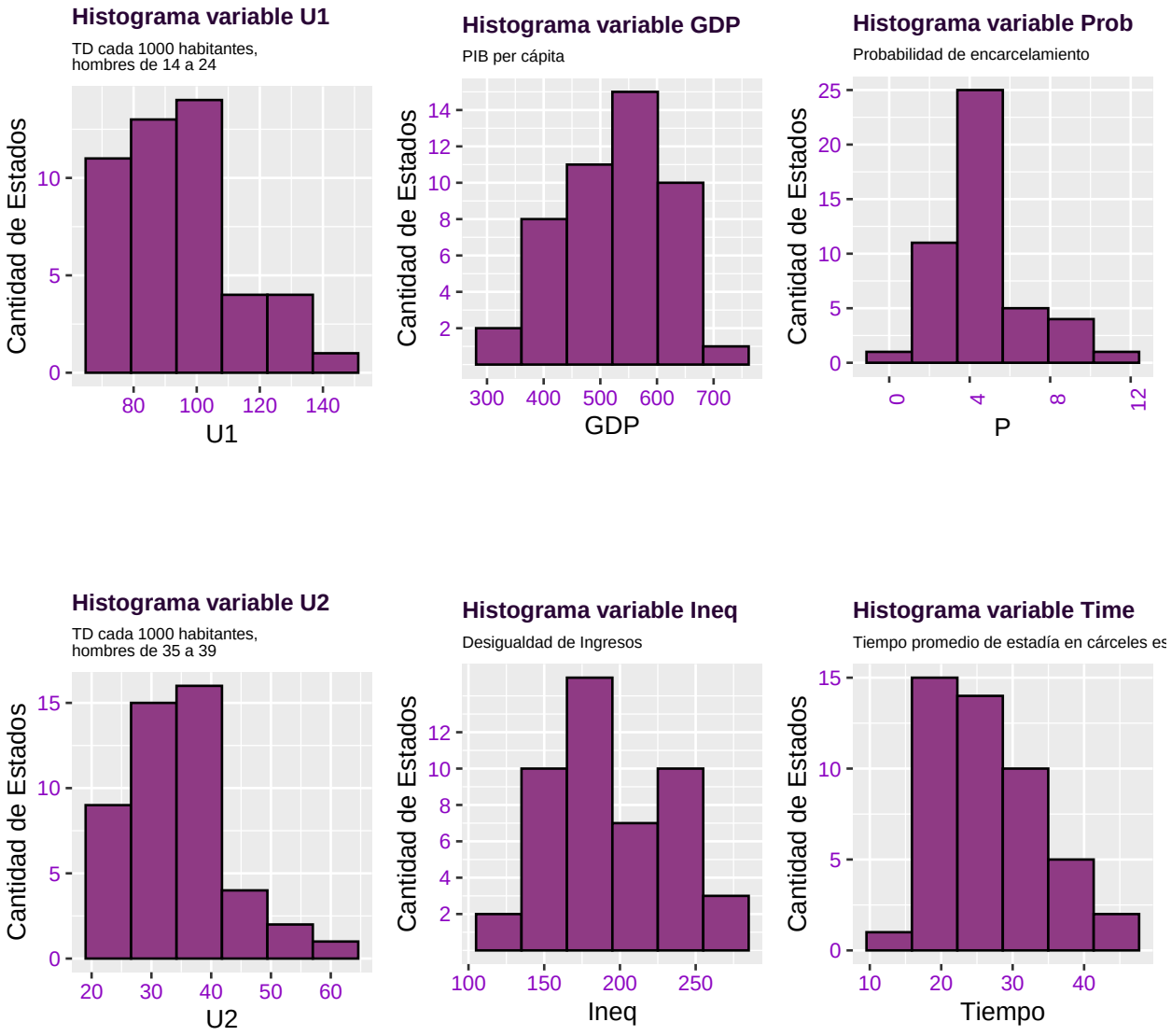


Figura 3: Histogramas (2)

Como se verá en los histogramas presentados a continuación y haciendo uso de la tabla (más precisamente del **CV**) es claro que las variables, de manera generalizada, presentan una variabilidad baja.

De manera más específica, los histogramas de las variables M, Po1, Nw, Po2, M.F, Pop, U1, U2, Prob y Time cuentan con una distribución asimétrica. La variabilidad entre los valores comprendidos hasta la mediana (aunque baja, como ya se mencionó) es menor que en el resto de las observaciones.

En el caso de la variable GDP y LF, la distribución a diferencia del resto es apróximadamente simétrica. La mediana y la media difieren en un número despreciable.

La variable Ineq también cuenta con una distribución asimétrica pero a diferencia de las demás,cuenta con menor variabilidad entre las observaciones en el tramo central (primer cuartil a tercer cuartil).

2.1.2. Medidas de resumen

Se presenta en forma de tabla el resumen de las variables númericas. En el mismo se presenta el valor mínimo y máximo de cada variable, medidas de tendencia central tales como lo son el primer y tercer cuartil, junto a la mediana.

A su vez, para estudiar la dispersión se incluye la media aritmética y una medida de variabilidad de la misma, el coeficiente de variación.

Cuadro 2: Medidas descriptivas para variables númericas

Variable	Min	1er Qu.	Mediana	3er Qu.	Max	Media	CV*100
Número de Hombres 14-24 / 1.000	119.0	130.0	136.0	146.0	177.0	138.6	9.1
Indice Escolaridad	87	98	108	114	122	106	11
Gasto per cápita 1.960	45	62	78	104	166	85	35
Gasto per cápita 1.959	41	58	73	97	157	80	35
Tasa participación masculina 14-24 por 1.000	480.0	530.5	560.0	593.0	641.0	561.2	7.2
Hombres cada 1.000 mujeres	934	964	977	992	1071	983	3
Población cada 100.000	3	10	25	42	168	37	104
Número de no caucásicos cada 1.000 habitantes	2	24	76	132	423	101	102
Tasa desempleo urbana Hombres 14-24 por 1.000	70	80	92	104	142	95	19
Tasa desempleo urbana Hombres 35-39 por 1.000	20	28	34	38	58	34	25
Producto bruto interno per cápita	288	460	537	592	689	525	18
Desigualdad ingreso	126	166	176	228	276	194	21
Probabilidad Encarcelamiento	0.69	3.27	4.21	5.45	11.98	4.71	48.28
Tiempo de estadía en carceles	12	22	26	30	44	27	27
Tasa de criminalidad	342	658	831	1058	1993	905	43

2.1.3. Correlación entre variables

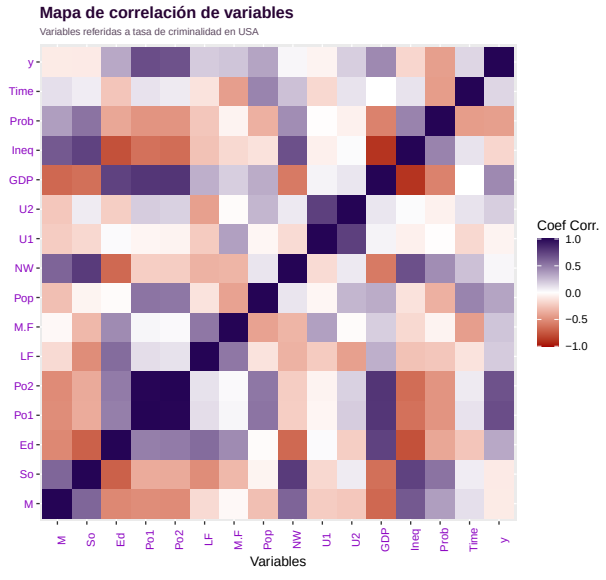


Figura 4: Mapa de correlación de variables incluidas

3. Especificación y selección de modelos inicial

El objetivo de esta sección es la aplicación de las distintas técnicas estadísticas impartidas en el curso para así llegar a un modelo final que no solo sea significativo al momento de estimar a **y**, sino que también se adecúe a los supuestos y propiedades deseadas (ganando de esta manera fidelidad).

En una primer instancia se planteará un *modelo inicial* constituído por parte de las variables de las cuales se poseen datos.

Claro está, se podría haber planteado en primera instancia un *modelo completo* (es decír, que contenga absolutamente a todas las variables). No es esto errado, pero si desconsiderado con el extenso análisis descriptivo planteado con anterioridad.

Como ya es sabido, variables que presentan una correlación muy alta no son marginalmente significativas al momento de definir la variable de respuesta.

Esto se evidencia en los tests de hipótesis en donde se analiza el aporte de cada variable dada las demás variables. Una correlación alta entre variables, podría indicar que parte de la información que aportan una de ellas está también presente en otra y esa cantidad de información se vió cuantificada de manera previa. Lo que tarde o temprano llevaría a descartar alguna de ellas.
En principio, a partir del "arsenal" descriptivo es claro que:

- **Ineq** y **GDP** tienen una correlación negativa altísima (-0.884).
- **P01** y **P02** poseen una correlación negativa casi perfecta (0.994)
- **Ineq** y **Ed** tienen también una correlación negativa bastante alta (-0,794)
- **SO** y **NW** mantienen una correlación positiva y de nivel alto (0,767)

A partir de lo afirmado, se procederá a "descartar.ª alguna de las variables que constituye cada dupla respaldándose en el valor del coeficiente de correlación existente entre las variables explicativas y **y**.

$\rho_{y,Ineq} = -0,179$ $\rho_{y,P01} = 0,688$ $\rho_{y,Ed} = 0,323$ $\rho_{y,NW} = 0,03$

$\rho_{y,GDP} = 0,441$ $\rho_{y,P02} = 0,667$ $\rho_{y,S0} = -0,09$

Se elige aquella variable cuyo coeficiente de correlación con **y** en valor absoluto sea mayor.

Trás esto, se decide que **Ineq** , **P02** y **NW** no sean incluídos en el modelo inicial ya que se entiende que las mismas no tendrán un aporte significativo en presencia de sus pares.

A manera de resumen podría decirse que este primer acercamiento al modelo sigue fielmente el principio de *parsimonia*[1].

Quedan entonces determinadas las variables a conformar el modelo inicial, que será analizado con detenimiento en la sección siguiente.

[1] Frugalidad y moderación en los gastos.

3.1. Modelo Inicial

Como una primera aproximación, se construye un modelo donde se incluyen todas las variables de la tabla de datos, en concreto el siguiente modelo de regresión:

$$\hat{y} = \beta_0 + \beta_1 Time + \beta_2 Prob + \cdots \beta_M M$$

Cuadro 3: Test sobre el modelo completo

$R^2.adj$	RSE	F Obs.	P-valor*100	Regresión.gl	Residuos.gl
70.781	209.064	8.429	0	15	31

Recordando que el R_a^2 hace referencia al porcentaje de variabilidad de **y** que es explicada con el modelo estimado, se considera al mismo como *aceptable*. Por otro lado, haciendo referecia a la significación del modelo, se consideranda el siguiente test de hipótesis y el estadístico F :

$$H_0) \beta_1 = \beta 2 = \cdots = \beta_k = 0$$

$$H_1) \text{No } H_0$$

$$F_{obs} = \frac{SCE/Regresion.gl}{RSE^2} = \frac{SCE/Regresion.gl}{SCR/Residuos.gl} = \frac{\sum (\hat{y}_i - \bar{y})^2 / Regresion.gl}{\sum (y_i - \hat{y}_i)^2 / Residuos.gl}$$

Siendo SCE la suma de cuadrados explicados por la regresión y RSE^2 el cuadrado del error estandar de los residuos, resulta para este caso partícular $SCE = 5.525.982$ y $RSE^2 = 43707,766$, de esta manera se obtiene el F_{obs} que permite rechazar H_0 y así afirmar que el modelo es estadísticamente significativo para explicar a **y**.

A continuación se testea la siginificación de cada variable en forma independiente, los resultados se muestran en el siguiente cuadro :

Cuadro 4: Estimación,error estandar y test individual del modelo completo

Variable	Estimación	Error estandar	Estadístico F	P valor	$\left(H_0^{\alpha=0.05}\right) \beta_i = 0$
Intercepto	-5984.288	1628.318	-3.675	0.001	Se rechaza H0
Número de Hombres 14-24 / 1.000	8.783	4.171	2.106	0.043	Se rechaza H0
Indicadora Estado Sur	-3.803	148.755	-0.026	0.980	No se rechaza H0
Indice Escolaridad	18.832	6.209	3.033	0.005	Se rechaza H0
Gasto per cápita 1.960	19.280	10.611	1.817	0.079	No se rechaza H0
Gasto per cápita 1.959	-10.942	11.748	-0.931	0.359	No se rechaza H0
Tasa participación masculina 14-24 por 1.000	-0.664	1.470	-0.452	0.655	No se rechaza H0
Hombres cada 1.000 mujeres	1.741	2.035	0.855	0.399	No se rechaza H0
Población cada 100.000	-0.733	1.290	-0.568	0.574	No se rechaza H0
Número de no caucásicos cada 1.000 habitantes	0.420	0.648	0.649	0.521	No se rechaza H0
Tasa desempleo urbana Hombres 14-24 por 1.000	-5.827	4.210	-1.384	0.176	No se rechaza H0
Tasa desempleo urbana Hombres 35-39 por 1.000	16.780	8.234	2.038	0.050	No se rechaza H0
Producto bruto interno per cápita	0.962	1.037	0.928	0.361	No se rechaza H0
Desigualdad ingreso	7.067	2.272	3.111	0.004	Se rechaza H0
Probabilidad Encarcelamiento	-48.553	22.724	-2.137	0.041	Se rechaza H0
Tiempo de estadía en carceles	-3.479	7.165	-0.486	0.631	No se rechaza H0

A partir del cuadro presentado y conforme a los tests realizados, se ve claramente que son tan solo 3 son las variables que de manera independiente (**y muy importantemente, en presencia de todas las demás**) logran un aporte significativo al momento de explicar el comportamiento de la tasa de criminalidad.

Ellas son **PO1**: Gasto pér capita en policía en 1960, **U2**: Tasa de desempleo Urbana de hombres entre 35 y 39 años por 1000 habitantes y por último **Prob*1000** Probabilidad de encarcelamiento cada 1000 habitantes.

¿Significa esto que se debe descartar el resto de las varibles y plantear un modelo caracterízado por tan solo las 3?, la respuesta es **no**.

Como bien se menciona anteriormente, los tests analizan el aporte dada las demás variables. Una correlación alta entre variables, podría indicar que parte de la información que aportan una de ellas está también presente en otra y esa cantidad de información se vió cuantificada de manera previa.

Con base en esta última afirmación es que se promueve el uso de distintas técnicas que nos permitirán elegír las variables de manera más acertada (y teniendo en cuenta este panorama).

4. Diagnóstico

4.1. Análisis de multicolinealidad

Consierando el problema oreviamente mencionado, se analizara la multicolinealidad (aproximada) de las variables independientes del modelo planteado, esto es relevante ya que si existe una relación lineal en la matriz de diseño, esto impactaría directamente en la varianza de los regresores $\beta_k = \left(X^T X\right)^{-1} X^T$, haciendo que las estimaciones varien ante pequeñas variaciones en nuestras observaciones y las predicciones serian menos confiables.

Estamos en frente a un problema de multicolinealidad aproximada cuando es posible afirmar que existe una relación lineal entre las variables explicativas. El término aproximado refiere al hecho que en el caso que se cumpla el fenomeno de forma exacta, la matriz no sería invertible y no existirian estimaciones únicas de los regresores únicas (Teorema de Gauss Markov) y no estaríamos frente a estimadores eficientes (insesgados y de minima varianza)
Recordando que :

$$\hat{\beta_k} \sim N\left(\beta, \sigma^2 \left(X^T X\right)^{-1}\right)$$

Como se menciono anteriormente, ante una posible relación lineal el determinante de la matriz $X^T X$ sería proximo a cero, obteniendo un determinante de la matriz inversa demasiado grande. Es decir para un σ^2 fijo la incertidumbre sería demasiado alta, considerando el hecho que en nuestra primera aproximación a un modelo de regresión es globalmente significativo pero salvo en una cantidad demasiado pequeña se puede afirmar que, de forma independiente existe una relación lineal con la tasa de criminalidad, es por esto que se cuantificara la intensidad de la multicolinealidad con el **Factor de inflación de varianza**.
El **VIF** nos indica en cuantas unidades se incrementa la varianza del estimador ante presencia de colinealidad y se define como :

$$VIF_j = \frac{1}{1 - R_j^2}$$

Donde R_j^2 hace referecia al coeficiente de determinación de una regresión que intenta establecer una relación lineal de X_j con las demás variables explicativas.

Pondremos a prueba las variables explicativas del modelo anteriormente mencionado y diremos que estamos frente a problemas de colinealidad con un $VIF \geq 10$, los resultados se muestran en el cuadro a continuación.

Cuadro 5: Prueba de multicolinealidad : Factor de incremento de Varianza **VIF**

Variable	VIF	Prueba
Número de Hombres 14-24 / 1.000	2.892	No hay problema de colinealidad
Indicadora Estado Sur	5.343	No hay problema de colinealidad
Indice Escolaridad	5.077	No hay problema de colinealidad
Gasto per cápita 1.960	104.659	Problema de colinealidad
Gasto per cápita 1.959	113.559	Problema de colinealidad
Tasa participación masculina 14-24 por 1.000	3.713	No hay problema de colinealidad
Hombres cada 1.000 mujeres	3.786	No hay problema de colinealidad
Población cada 100.000	2.537	No hay problema de colinealidad
Número de no caucásicos cada 1.000 habitantes	4.674	No hay problema de colinealidad
Tasa desempleo urbana Hombres 14-24 por 1.000	6.064	No hay problema de colinealidad
Tasa desempleo urbana Hombres 35-39 por 1.000	5.089	No hay problema de colinealidad
Producto bruto interno per cápita	10.530	Problema de colinealidad
Desigualdad ingreso	8.645	No hay problema de colinealidad
Probabilidad Encarcelamiento	2.809	No hay problema de colinealidad
Tiempo de estadía en carceles	2.714	No hay problema de colinealidad

En base a esto, podemos afirmar que nuestro modelo presenta problemas con la colinealidad y es por esto que continuaremos con la selección a pasos por el método Stepwise, también que hay que recordar que en el caso de seleccionar variables por la correlación dos a dos entre ellas y quedarse con aquellas que tienen una correlación mas alta con y es un grave error ya que nos estamos olvidando del efecto que puede tener con las demás variables.

4.2. Método Stepwise

El método de Stepwise (basado en el *F-Test*) comienza seleccionado aquella que tiene una mayor correrlación con la variable **y**, la segunda es aquel modelo que con la variable que se incluyó en el paso anterior maximiza el coeficiente de determinación siguiendo iterando hasta que no se cumpla con el criterio de entrada.

Cuadro 6: Estimación,error estandar y test individual tras aplicar el método Stepwise

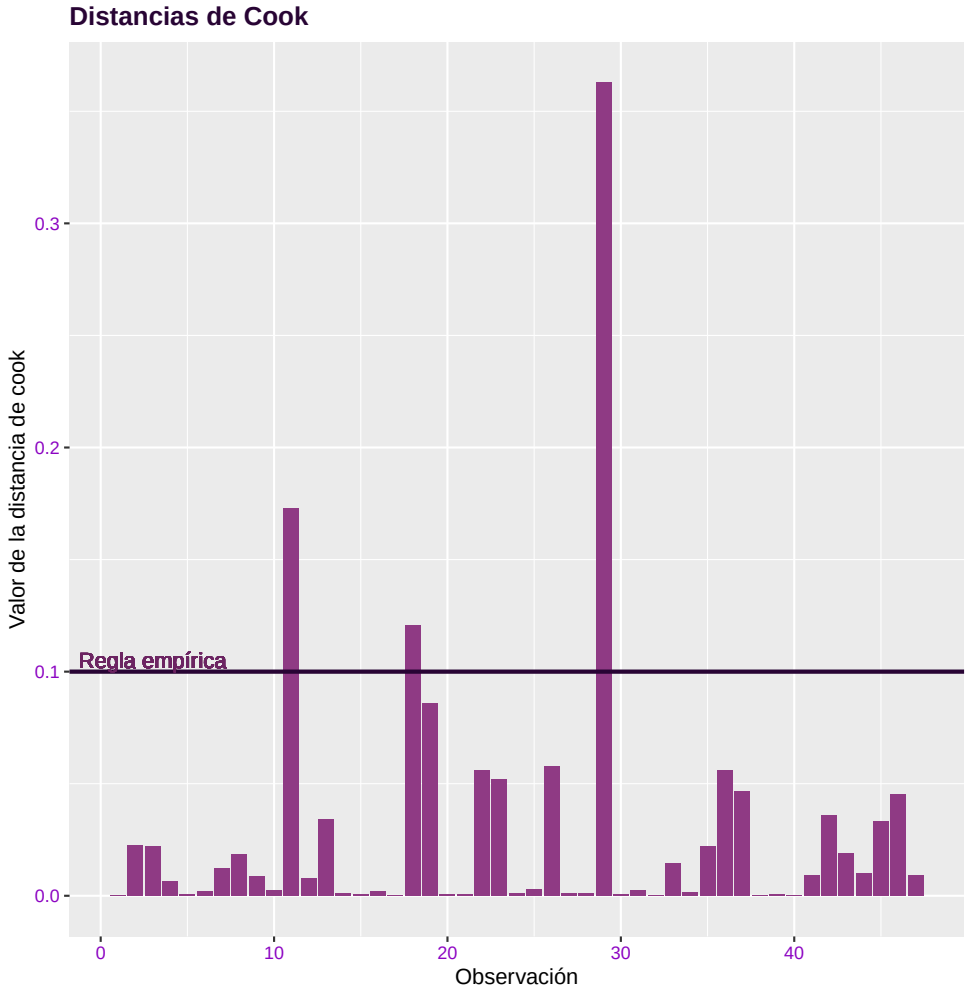
Variable	Estimación	Error estandar	Estadístico F	P valor	$\left(H_0^{\alpha=0.05}\right)\beta_i = 0$
Intercepto	-5040.505	899.843	-5.602	0.000	Se rechaza H0
Gasto per capita en policía 1960	11.502	1.375	8.363	0.000	Se rechaza H0
Desigualdad del ingreso	6.765	1.394	4.855	0.000	Se rechaza H0
Indice que refleja la escolaridad del estado	19.647	4.475	4.390	0.000	Se rechaza H0
Número de hombres entre 14 y 24 / 1000	10.502	3.330	3.154	0.003	Se rechaza H0
Probabilidad de encarcelamiento	-38.018	15.281	-2.488	0.017	Se rechaza H0
Tasa de desempleo urbana hombres 35-39 años x 1000	8.937	4.091	2.185	0.035	Se rechaza H0

4.3. Busqueda de observaciones influyentes

En esta sección se buscará estudiar cuales de las observaciones presentan valores influyentes, para ello se hará uso de la Distancia de Cook.

Esta es una medida del nivel de influencia de la observación i-ésimas sobre la estimación de $\widehat{\beta}$, es decir se busca medir si su presencia o ausencia en el modelo hace que el mismo cambie.
Una distancia de Cook elevada significa que una observación tiene mayor influencia al momento de determinar los $\widehat{\beta}$.

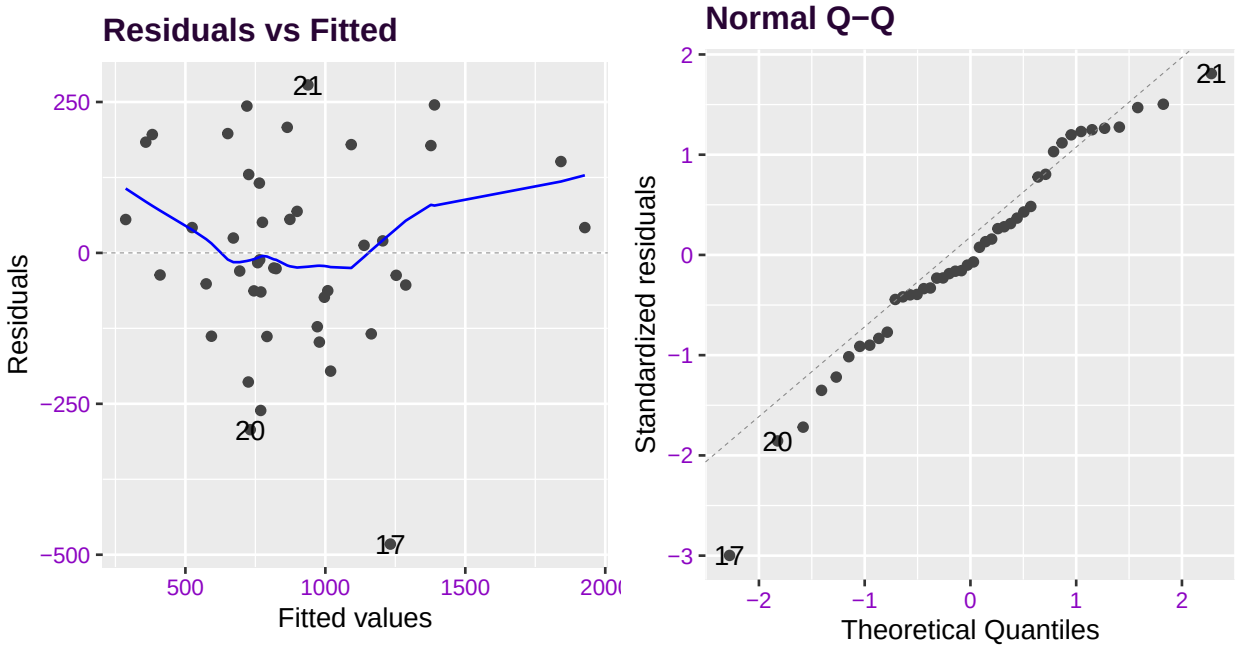
$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}(-i))' X' X (\widehat{\beta} - \widehat{\beta}(-i))}{(k+1)\widehat{\sigma^2}}$$



Tomando como regla empírica el valor de $\frac{4}{n-k-1} = 4/40$ puede verse que las observaciones **11**, **29** (de manera excesiva) y **18** sobrepasan la regla estipulada.
Por ende, se decide retirarlas del modelo reducido ya que las mismas tienen una influencia preponderante en la estimación. Como se vió en clase, observaciones de este tipo pueden llevar a un modelo alejado de la realidad.

4.4. Normalidad

Uno de los supuestos que dimos acerca de nuestro modelo lineal, refiere a la distribución que le dimos a los errores, en este caso siguen una distribución normal, esto nos permite realizar las pruebas anteriormente presentadas a lo largo de este trabajo, en el caso que este supuesto no se cumpla los estdaísticos no reflejaran lo que nosotros realmente necesitamos,es por esto que es de crucial importancia chequear la distribución de los errores. En primer lugar haremos un primer acercamiento con un gráfico de QQ-Plot que muestra la similitud de los percentiles de la distribución normal téorica y la observada en nuestros errores, por último un gráfico de residudos contra los valores predichos para ver si existe algun patrón en su recorrido.



En base a los gráficos podemos ver que la esperanza de los errores se mantiene cercana a cero, a diferencia de algunas observaciones. También hay que mencionar que no encontramos un patrón en la dispersión de los errores, en cambio en la gráfica QQ-plot podemos ver que en valores centrales la distribución se asemeja a una distribución normal a exepción de la observación 17, 20 y 21.
Para poder afirmar que se cumple el supuesto de normalidad de los errores relizaremos, tres pruebas

- **Test Kolmogrov-Smirnov Lillie:** Compara la distribución teórica F^* y la distribución empírica de los errores $S(x)$ $T = sup_x|F^*(x) - S(X)|$
- **Shapiro-Wilks:** Que plantea un estadísitco que es una función de las estadísticas de órden de la distribución de los errores y se compara con un valor de tabla
- **Jarque Bera:** Se basa en los coeficientes de simetría y curtosis de la muestra, que para una nomral son 0 y 3 respectivamente y con estadístico χ^2_2

En base a los test puedo afirtmar que la distribución de los errores sigue una distribución Normal

Cuadro 7: Test de Normalidad

Test	Pvalor	Estadistico	Resultado
Lillie	0.188	0.111	No rechazo normalidad
Shapiro	0.115	0.959	No rechazo normalidad
Jarque Bera	0.070	5.310	No rechazo normalidad

4.5. Heterosedasticidad

4.5.1. Test de Breusch-Pagan

```
##
##  studentized Breusch-Pagan test
##
## data:  .
## BP = 3.4998, df = 6, p-value = 0.744
```

5. Anexo

5.1. Selección de Modelos : StepWise

```
## Stepwise regression (forward-backward), alpha-to-enter: 0.15, alpha-to-remove: 0.15
##
## Full model: y ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
##     GDP + Ineq + Prob + Time
## <environment: 0x00000000268683f0>
##
## --= Step (forward) 1 =--
## Single term additions
##
## Model:
## y ~ 1
##        Df Sum of Sq     RSS    AIC F value    Pr(>F)
## <none>              6130013 523.16
## M       1      2552 6127461 525.14  0.0175  0.895403
## So      1     32554 6097459 524.92  0.2242  0.638282
## Ed      1    761639 5368374 519.32  5.9588  0.018940 *
## Po1     1   3112662 3017351 493.97 43.3267 5.771e-08 ***
## Po2     1   2893417 3236596 497.06 37.5467 2.603e-07 ***
## LF      1    193870 5936143 523.74  1.3717  0.248124
## M.F     1    451371 5678643 521.79  3.3384  0.074792 .
## Pop     1    506531 5623482 521.36  3.7831  0.058487 .
## NW      1      1207 6128806 525.15  0.0083  0.927954
## U1      1      1981 6128032 525.14  0.0136  0.907797
## U2      1    204910 5925103 523.66  1.4525  0.234874
## GDP     1    984473 5145540 517.46  8.0357  0.007023 **
## Ineq    1    179734 5950280 523.85  1.2686  0.266415
## Prob    1    850163 5279850 518.59  6.7629  0.012795 *
## Time    1      4706 6125307 525.12  0.0323  0.858307
## ---
```


```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## --= Step (forward) 2 =--
## Single term additions
##
## Model:
## y ~ Po1
##        Df Sum of Sq     RSS    AIC F value   Pr(>F)
## <none>              3017351 493.97
## M       1    653261 2364090 485.24 11.3294 0.001667 **
## So      1    245684 2771667 492.23  3.6343 0.063626 .
## Ed      1     10063 3007288 495.82  0.1372 0.712994
## Po2     1    150511 2866840 493.72  2.1525 0.149964
## LF      1     13492 3003859 495.77  0.1842 0.670072
## M.F     1    151904 2865447 493.70  2.1735 0.148044
## Pop     1      6622 3010729 495.87  0.0902 0.765467
## NW      1    238982 2778369 492.34  3.5266 0.067516 .
```

```
## U1      1       660 3016691 495.96  0.0090 0.924997
## U2      1     14473 3002878 495.76  0.1976 0.659000
## GDP     1    413154 2604197 489.49  6.5046 0.014591 *
## Ineq    1    870164 2147187 481.00 16.6156 0.000205 ***
## Prob    1     38564 2978787 495.40  0.5308 0.470415
## Time    1     36079 2981272 495.44  0.4962 0.485168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## --= Step (forward) 3 =--
## Single term additions
##
## Model:
## y ~ Po1 + Ineq
##        Df Sum of Sq     RSS    AIC F value   Pr(>F)
## <none>              2147187 481.00
## M       1    148905 1998282 479.84  2.9807 0.091982 .
## So      1     43289 2103898 482.10  0.8230 0.369732
## Ed      1    493559 1653627 471.51 11.9388 0.001316 **
## Po2     1     20807 2126380 482.57  0.3914 0.535114
## LF      1    131062 2016124 480.23  2.6003 0.114708
## M.F     1    297914 1849272 476.43  6.4439 0.015129 *
## Pop     1     29447 2117739 482.39  0.5562 0.460155
## NW      1     55674 2091513 481.85  1.0648 0.308332
## U1      1     10382 2136804 482.79  0.1944 0.661693
## U2      1      4154 2143032 482.92  0.0775 0.782091
## GDP     1     53424 2093763 481.89  1.0206 0.318444
## Prob    1    245646 1901541 477.66  5.1673 0.028460 *
## Time    1      1810 2145377 482.96  0.0337 0.855192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## --= Step (forward) 4 =--
## Single term additions
##
## Model:
## y ~ Po1 + Ineq + Ed
##        Df Sum of Sq     RSS    AIC F value  Pr(>F)
## <none>              1653627 471.51
## M       1    227113 1426514 467.01  6.2091 0.01707 *
## So      1      2071 1651556 473.45  0.0489 0.82611
## Po2     1     39226 1614402 472.45  0.9476 0.33633
## LF      1      2659 1650968 473.44  0.0628 0.80341
## M.F     1     62483 1591144 471.81  1.5315 0.22328
## Pop     1      5096 1648532 473.37  0.1205 0.73031
## NW      1        16 1653612 473.51  0.0004 0.98477
## U1      1     22557 1631071 472.90  0.5393 0.46710
## U2      1     38456 1615172 472.47  0.9285 0.34118
## GDP     1      3002 1650626 473.43  0.0709 0.79141
```

```
## Prob    1    195586 1458041 467.97  5.2316 0.02769 *
## Time    1     45886 1607742 472.27  1.1131 0.29791
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## --= Step (forward) 5 =--
## Single term additions
##
## Model:
## y ~ Po1 + Ineq + Ed + M
##        Df Sum of Sq     RSS    AIC F value  Pr(>F)
## <none>              1426514 467.01
## So      1      3267 1423247 468.91  0.0872 0.76934
## Po2     1     50639 1375876 467.42  1.3986 0.24431
## LF      1      9288 1417226 468.72  0.2490 0.62062
## M.F     1     36742 1389772 467.86  1.0046 0.32253
## Pop     1        29 1426486 469.01  0.0008 0.97810
## NW      1     35547 1390968 467.90  0.9711 0.33064
## U1      1    104513 1322001 465.66  3.0042 0.09116 .
## U2      1    184235 1242280 462.92  5.6355 0.02276 *
## GDP     1     25397 1401117 468.22  0.6888 0.41175
## Prob    1    191810 1234705 462.65  5.9032 0.01995 *
## Time    1     16412 1410103 468.50  0.4423 0.51005
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## --= Step (forward) 6 =--
## Single term additions
##
## Model:
## y ~ Po1 + Ineq + Ed + M + Prob
##        Df Sum of Sq     RSS    AIC F value  Pr(>F)
## <none>              1234705 462.65
## So      1     15552 1219153 464.10  0.4720 0.49636
## Po2     1     37996 1196708 463.28  1.1748 0.28543
## LF      1     22172 1212533 463.86  0.6766 0.41604
## M.F     1     45341 1189364 463.01  1.4105 0.24254
## Pop     1     10363 1224342 464.28  0.3132 0.57912
## NW      1      1561 1233143 464.60  0.0469 0.82983
## U1      1     85909 1148796 461.48  2.7669 0.10468
## U2      1    156442 1078263 458.69  5.3682 0.02615 *
## GDP     1      7502 1227203 464.39  0.2262 0.63717
## Time    1     17871 1216833 464.01  0.5434 0.46567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## --= Step (forward) 7 =--
## Single term additions
```

```
##
## Model:
## y ~ Po1 + Ineq + Ed + M + Prob + U2
##        Df Sum of Sq     RSS    AIC F value Pr(>F)
## <none>              1078263 458.69
## So      1     13638 1064625 460.13  0.4612 0.5014
## Po2     1     32959 1045304 459.33  1.1351 0.2938
## LF      1       501 1077762 460.67  0.0167 0.8978
## M.F     1     12491 1065772 460.18  0.4219 0.5201
## Pop     1     21982 1056281 459.79  0.7492 0.3925
## NW      1      3615 1074648 460.55  0.1211 0.7299
## U1      1      3297 1074966 460.56  0.1104 0.7416
## GDP     1      1108 1077155 460.65  0.0370 0.8485
## Time    1     24089 1054173 459.70  0.8227 0.3704
##
## Call:
## lm(formula = y ~ Po1 + Ineq + Ed + M + Prob + U2, data = Datos)
##
## Coefficients:
## (Intercept)          Po1         Ineq           Ed            M         Prob
##   -5120.813       12.443        6.911       19.351       10.441      -32.177
##          U2
##       8.440
```

5.2. Script de R

6. Bibliografía