Tasa de criminalidad en USA

Modelos Líneales

Trabajo final

Docentes : Laura Nalbarte y Fernando Massa

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:

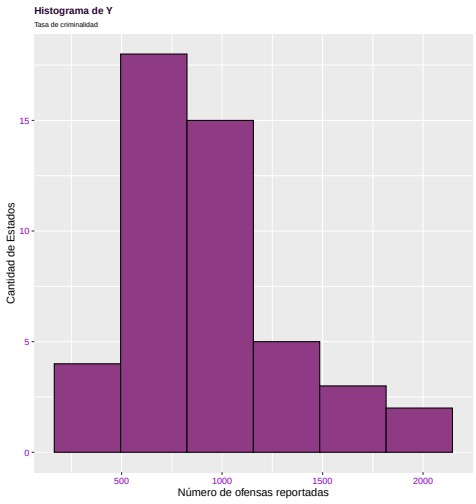


Figura 1: Histograma de la Tasa de Criminalidad

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

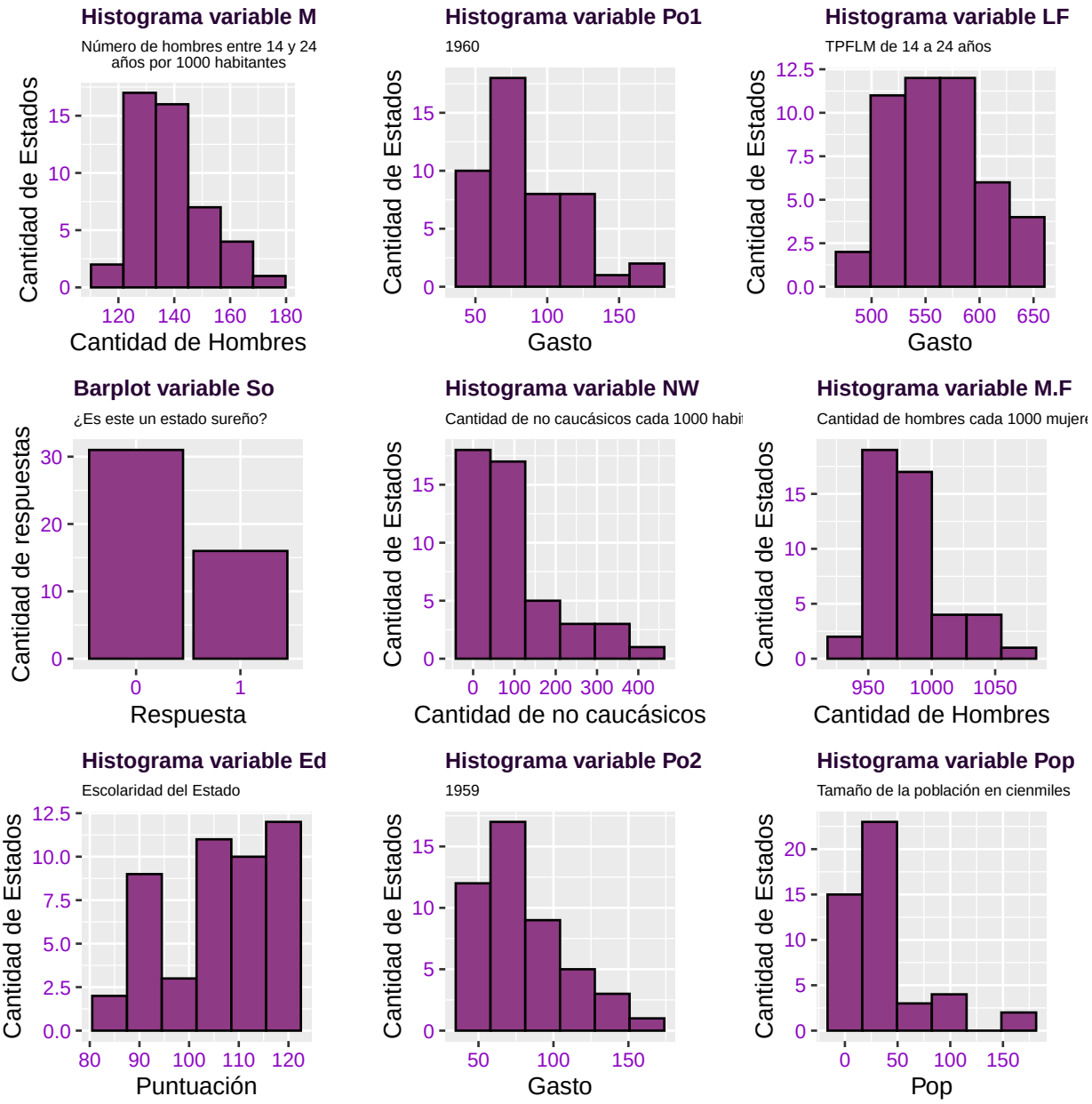


Figura 2: Histogramas (1)

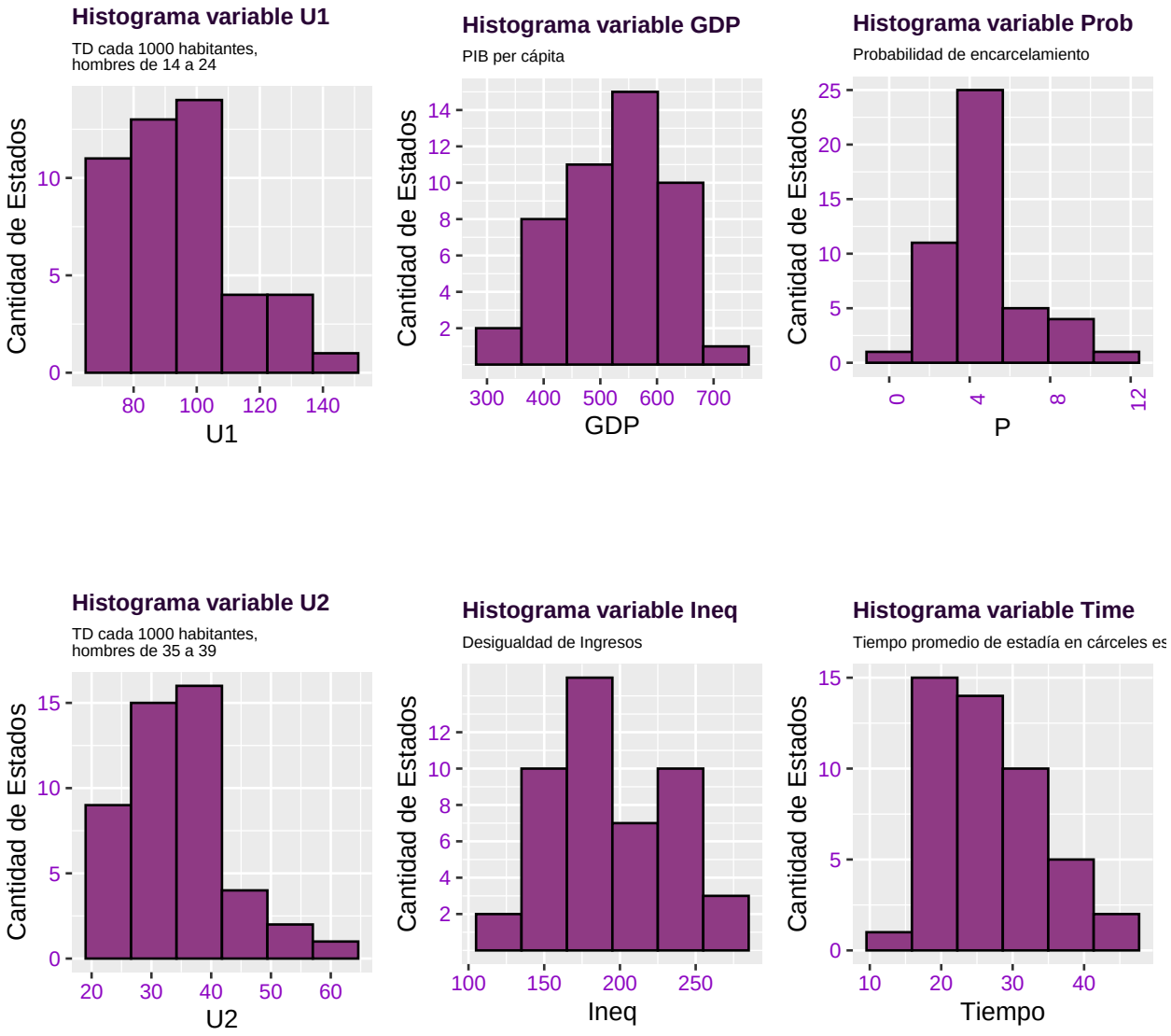


Figura 3: Histogramas (2)

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

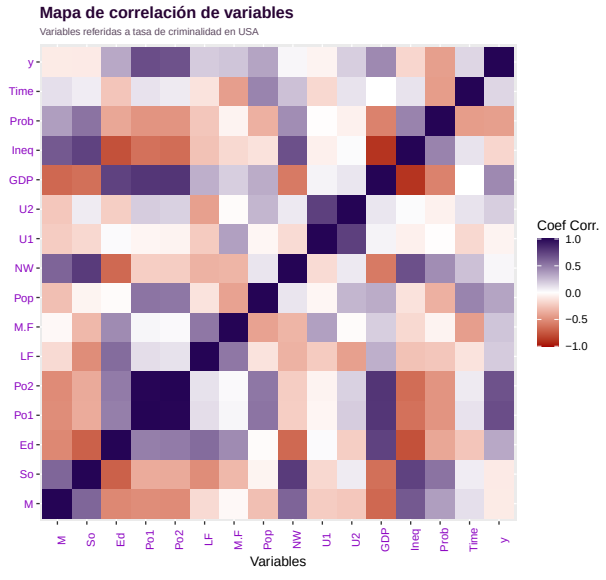


Figura 4: Mapa de correlación de variables incluidas

3. Introducción al Modelo Completo

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).

De manera resumida podría decirse que este informe seguirá fielmente el principio de *parsimonia*[1]. Respetando en todo momento el objetivo principal, la creación de un modelo lo más certero posible.

En una primer instancia se planteará el **Modelo completo** constituído por la totalidad de las variables de las cuales se poseen datos.

Claro está, se podría haber planteado en primera instancia un *modelo inicial* (es decír, que contenga parte de las variables). No es esto errado, sin embargo es este un procedimiento que requiere de cierta experiencia en el tema de criminología.

Como ya es sabido, variables que presentan una correlación muy alta no son (en general) 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.

Al comienzo de este trabajo jugó este principio de la correlación un rol monumental. Basándose en la alta correlación de variables 2 a 2, se procedió a retirar de la dupla a aquella que menor correlación mantenía con **y**.
Esto es en principio coherente, a pesar de ello y trás el análisis de disntitos panoramas, el caso partícular de **Ineq** y **GDP** dejó en claro que esta técnica fue en un principio apresurada. Si bien las variables comparten gran nivel de información, trás la aplicación de las técnicas de **Forward,Backwards,Stepwise** ambas varibles eran incluídas en el modelo en todo escenario.

Esta situación subyace de que incluso cuando ambas comparten un nivel elevado de información, el aporte único de las mismas es alto en comparación con el resto de las variables presentadas.

Lo último se vió claramente en el R_a^2 de los modelos finales.

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)

Como veremos en las siguientes secciones del documento, el caso de **Ineq** y **GDP** es partícular. Como se pensó en un principio, en el resto de las duplas muy probablemente sobreviva una sola de las varibles (si no es que ambas son descartadas).

[1] Frugalidad y moderación en los gastos.

3.1. Modelo Completo

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 4 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 **M**,**PO1**,**U2** y por último **Prob*1000**.

¿Significa esto que se debe descartar el resto de las varibles y plantear un modelo caracterízado por tan solo las 4?, 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 previamente mencionado, se analizará 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 varíen ante pequeñas perturbaciones en las observaciones, logrando así que predicciones sean menos confiables.

Estamos 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 de que en el caso de que se cumpla el fenómeno de forma exacta, la matriz no sería invertible y no existirían estimaciones únicas de los regresores (Teorema de Gauss Markov). A su vez no se obtendrían estimadores eficientes (insesgados y de mínima 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, siendo el determinante de la matriz inversa demasiado grande. Es decir para un σ^2 fijo, la incertidumbre sería demasiado alta.

Como se vió previamente, el modelo presentado hasta el momento es globalmente significativo, sin embargo tan solo 4 de las variables consideradas son significativas de manera independiente (presente todas las demás), es por esto que se cuantificaré 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

Con base en base a esto, podemos afirmar que este modelo presenta problemas de multicolinealidad y es por esto que se continuará con la selección a pasos por el método Stepwise.

4.2. Método Stepwise

El método de Stepwise (basado en el *F-Test*) inicia seleccionado aquella variable que tiene una mayor correlación con la variable **y**, la segunda en ingresar al modelo es aquella que mayor SCE(X|demás variables).

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

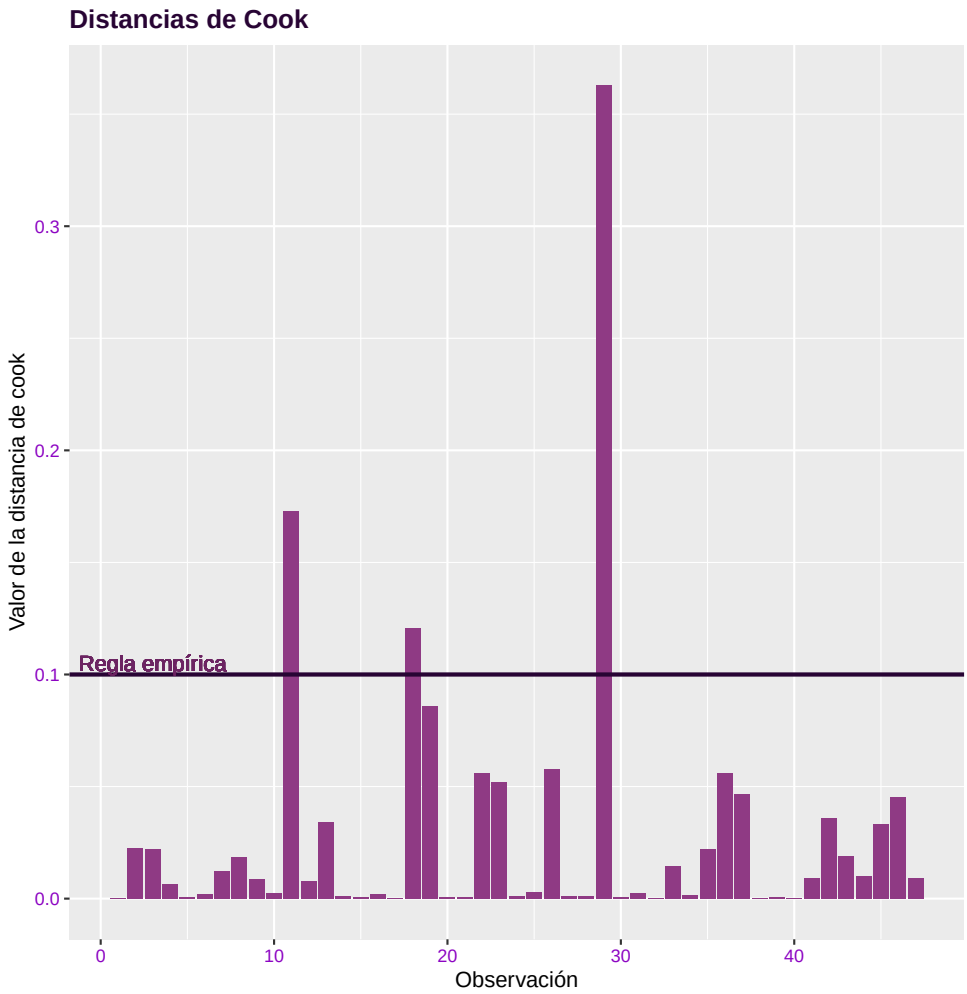
Una vez realizado dicho método[2] podemos encontrarnos con un modelo con 6 variables, todas de ellas significativas y sin problemas de multicolinealidad, además de tener un R_a^2 de 0.73 mayor que el del modelo completo 0.71.
Para continuar con el diagnostico del modelo, en la siguiente sección se hará un estudio de las observaciones atípicas y/o influyentes.

[2]Se pueden encontrar las salidas en el anexo.

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.

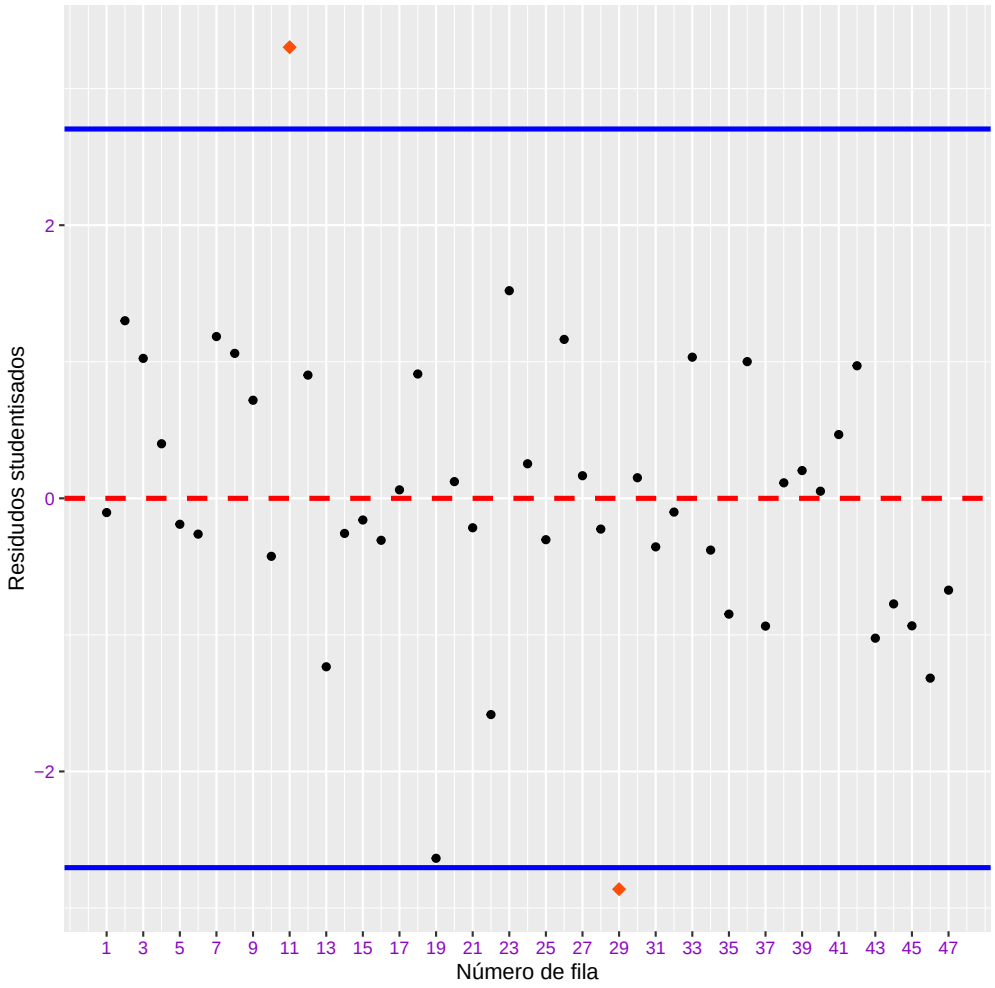
4.4. Observaciones atipicas

Analizaremos aquellas observaciones que son atipicas, es decir aquellas que tienen una dispersión mayor a las demás, pero no hacen variar las estimaciones de los regresores.

Para detectar observatios atípicas, construimos un vector con los residuos studentisados del modelo,lo comparamos con un umbral definido por una distribución t y en el caso que la superen, se consideran atípicos.
Se definen los residudos studentisados como :

$$t_i = \frac{e}{\sqrt{\hat{\sigma_i^2}(1 - h_{ii})}}$$

Una vez calculado los residuos de esta manera, podemos definir a los outliers como aquellas observaciones que sobrepasen el umbral de una distribución $t^{df.res}_{1-0.01/2}$ (lineas continuas horizontales en el gráfico)

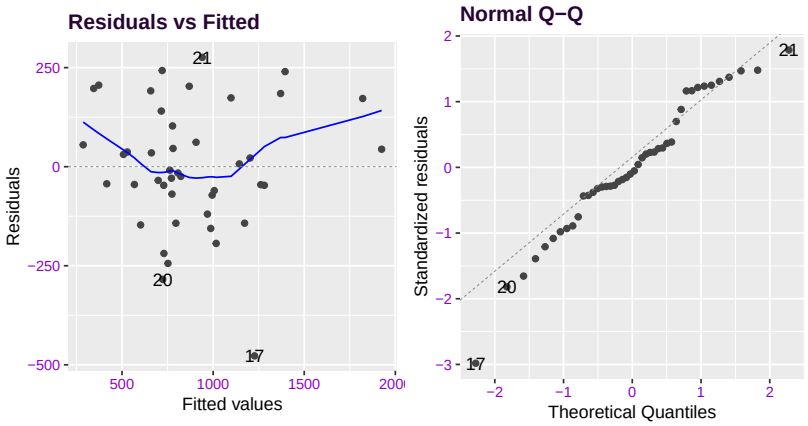


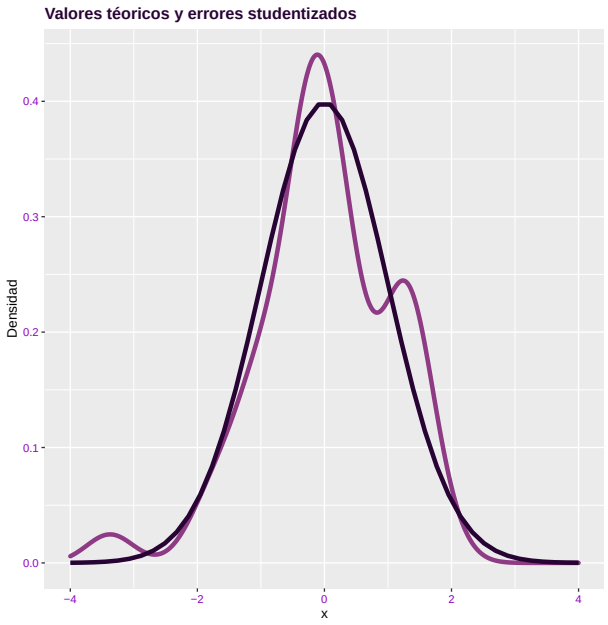
Recordando el analisis de la sección anterior, el estado número **11**, **29** y **18** presentan problemas con la distancia de Cook, con este gráfico podemos afirmar que las observaciones **29** y **11** son influyentes y atipicas, mientras que para la número **18** podemos ver que es muy cercana a cero, la esperanza de los errores por lo que es solamente atipica.
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.

4.5. Normalidad

Uno de los supuestos fundamentales al momento de construír este modelo lineal, refiere a la distribución de los errores, en este caso siguen una distribución normal. Esto permite realizar las pruebas anteriormente presentadas a lo largo de este trabajo, en el caso que este supuesto no se cumpla los estadísticos no reflejarán lo que se desea probar (su distribución no será la esperada), es por esto que es de crucial importancia chequear la distribución de los errores.

En primer se hará 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.





Con base en los gráficos se observa que la esperanza de los errores se mantiene cercana a cero, con excepción de algunas observaciones. También se debe mencionar que no es posible encontrar un patrón en la dispersión de los errores, en cambio en la gráfica QQ-plot muestra como los valores centrales de la distribución se asemeja a una distribución normal, a exepción de la observación 17, 20 y 21, lo mismo se puede apreciar en gráfica que compara la distribución de los errores (que fue construida con la función *geom density* de ggplot que calcula y dibuja la estimación de la densidad Kernel con valores por defecto) con una distribución normal.

Para poder afirmar que se cumple el supuesto de normalidad de los errores se relizan, 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

Cuadro 7: Test de Normalidad

Test	Pvalor	Estadistico	Resultado
Lillie	0.148	0.116	No rechazo normalidad
Shapiro	0.088	0.955	No rechazo normalidad
Jarque Bera	0.119	4.264	No rechazo normalidad

En base a los test se puede afirmar que la distribución de los errores sigue una distribución Normal, a continuación se hará el análisis de la homocedasticidad.

4.6. Homocedasticidad

En esta sección se discutirá lo siguiente:

$$H_0)\sigma_i^2 = \sigma^2,\ \forall i = 1, ..., 6$$
$$H_1)\text{No } H_0$$

Para ello se hará uso de el test de Breusch-Pagan y el test de White.

El test de **White** consiste en la siguiente serie de pasos:

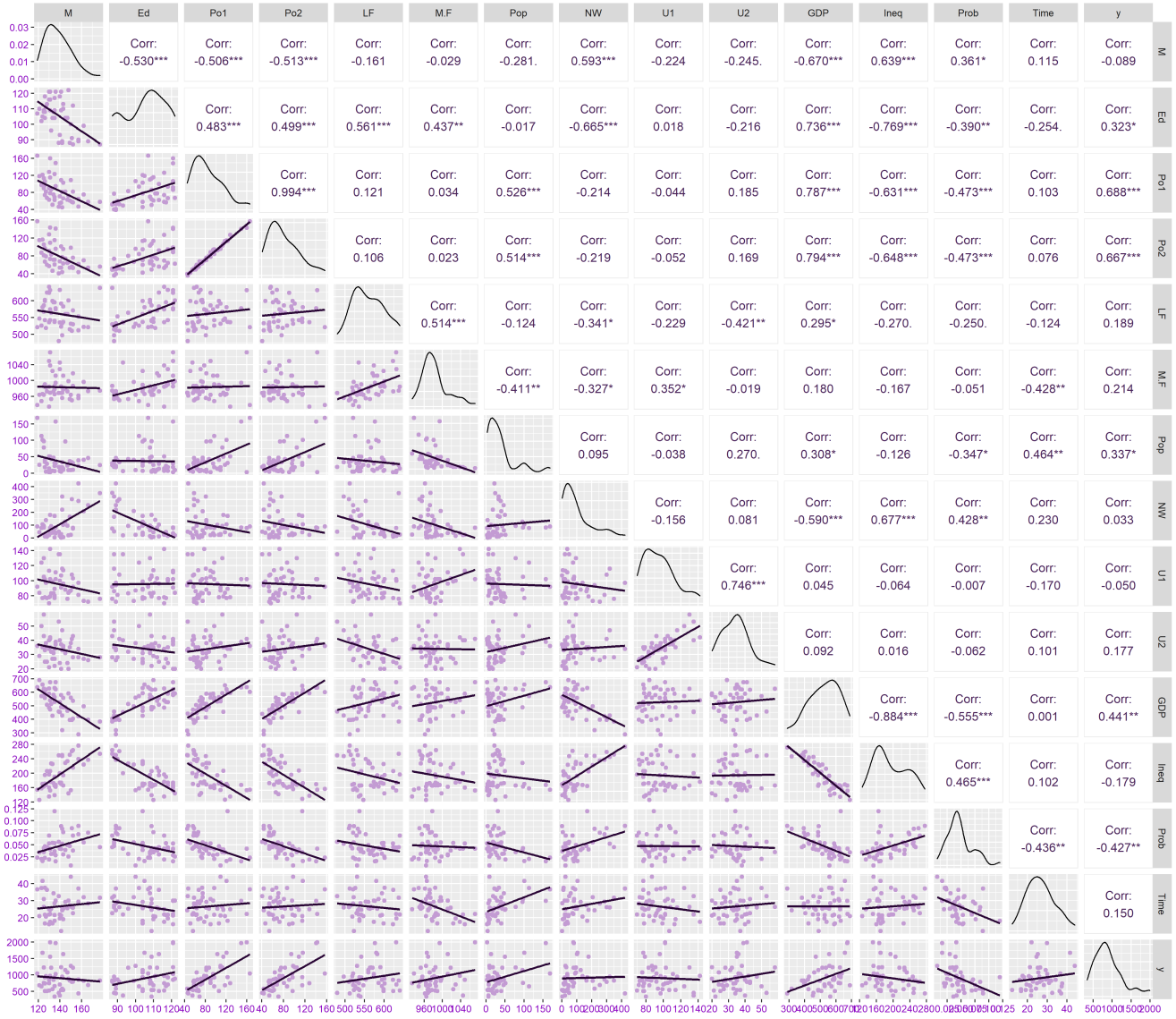
- Se estima el modelo y luego se genera el vector de errores estimados.
- Se ajusta el modelo $e^2 = X\alpha + Z\beta + W\delta + \mu$, con Z matriz de productos cruzados se las variables originales, W contiene las variables originales al cuadrado.
- Se calcula R^2 del modelo.
- Se calcula el estadístico nR^2 que se distribuye χ_k con k cantidad de variables del modelo.
- Se realiza el test.

El test de **Breusch-Pagan** empleado se basa en:

Cuadro 8: Test de Homocedasticidad

Test	Pvalor	Estadistico	Resultado
Breusch-Pagan	0.668	4.066	No rechazo Homocedasticidad
White	0.595	10.245	No rechazo Homocedasticidad

5. Anexo



5.1. Selección de Modelos : StepWise

En este apartado se muestra la salida de la selección a pasos de Stepwise, se puede ver que el α de entrada y salida es igual a 0.15, comienza solo incluyendo la constante en el modelo, luego agrega **Po1** hasta llegar al séptimo paso y terminar el algorítmo

```
## 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: 0x000000001c272ca8>
##
## --= Step (forward) 1 =--
## Single term additions
##
```

```
## Model:
## y ~ 1
##        Df Sum of Sq     RSS    AIC F value    Pr(>F)
## <none>              6250417 524.01
## M       1      4106 6246310 525.99  0.0276  0.868824
## So      1     23206 6227211 525.85  0.1565  0.694390
## Ed      1    725865 5524551 520.58  5.5183  0.023597 *
## Po1     1   3337539 2912878 492.42 48.1231 1.788e-08 ***
## Po2     1   3096493 3153924 495.92 41.2352 9.828e-08 ***
## LF      1    227378 6023039 524.38  1.5856  0.214916
## M.F     1    471590 5778827 522.56  3.4275  0.071160 .
## Pop     1    555473 5694944 521.92  4.0966  0.049361 *
## NW      1      6196 6244221 525.97  0.0417  0.839228
## U1      1        39 6250377 526.01  0.0003  0.987098
## U2      1    198725 6051692 524.59  1.3792  0.246855
## GDP     1   1035651 5214766 518.04  8.3412  0.006102 **
## Ineq    1    144104 6106313 524.99  0.9912  0.325159
## Prob    1   1290740 4959676 515.84 10.9304  0.001943 **
## Time    1     17735 6232682 525.89  0.1195  0.731295
## ---
## 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>              2912878 492.42
## M       1    686350 2226528 482.60 12.6387 0.0009689 ***
## So      1    444671 2468207 487.13  7.3865 0.0095849 **
## Ed      1     36496 2876382 493.87  0.5202 0.4748445
## Po2     1    160925 2751953 491.92  2.3975 0.1292107
## LF      1      6046 2906832 494.33  0.0853 0.7717476
## M.F     1    135020 2777858 492.33  1.9928 0.1655889
## Pop     1      3504 2909374 494.37  0.0494 0.8252479
## NW      1    336999 2575879 489.01  5.3640 0.0256344 *
## U1      1       616 2912262 494.41  0.0087 0.9262832
## U2      1      9482 2903396 494.28  0.1339 0.7163006
## GDP     1    421443 2491435 487.54  6.9354 0.0118648 *
## Ineq    1    931365 1981513 477.47 19.2711 7.779e-05 ***
## Prob    1       916 2911962 494.41  0.0129 0.9101458
## Time    1     24418 2888460 494.05  0.3466 0.5592728
## ---
## 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>              1981513 477.47
## M       1    162014 1819500 475.71  3.5617 0.066398 .
## So      1      2446 1979068 479.41  0.0494 0.825193
## Ed      1    399620 1581893 469.56 10.1049 0.002852 **
## Po2     1     22716 1958797 478.96  0.4639 0.499737
## LF      1     91565 1889948 477.39  1.9379 0.171585
## M.F     1    266756 1714757 473.11  6.2226 0.016843 *
## Pop     1     48237 1933277 478.38  0.9980 0.323794
## NW      1     25754 1955759 478.89  0.5267 0.472206
## U1      1      6645 1974869 479.32  0.1346 0.715660
## U2      1      6907 1974606 479.31  0.1399 0.710334
## GDP     1     56175 1925339 478.20  1.1671 0.286477
## Prob    1     74359 1907155 477.79  1.5596 0.218988
## Time    1      1246 1980267 479.44  0.0252 0.874749
## ---
## 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>              1581893 469.56
## M       1    232602 1349291 464.56  6.7232 0.01333 *
## So      1     23076 1558817 470.91  0.5773 0.45192
## Po2     1     39581 1542313 470.44  1.0009 0.32327
## LF      1      4129 1577764 471.44  0.1021 0.75108
## M.F     1     65074 1516819 469.71  1.6732 0.20345
## Pop     1     12861 1569032 471.20  0.3197 0.57504
## NW      1      1591 1580302 471.51  0.0393 0.84395
## U1      1     18306 1563588 471.05  0.4566 0.50321
## U2      1     26898 1554995 470.80  0.6746 0.41644
## GDP     1      4980 1576913 471.42  0.1232 0.72752
## Prob    1    103514 1478380 468.58  2.7307 0.10647
## Time    1     21233 1560660 470.96  0.5306 0.47071
## ---
## 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>              1349291 464.56
```

```
## So      1      2303 1346988 466.48  0.0650 0.80017
## Po2     1     50071 1299220 464.90  1.4645 0.23369
## LF      1      9938 1339353 466.23  0.2820 0.59851
## M.F     1     40031 1309260 465.23  1.1618 0.28788
## Pop     1      1276 1348015 466.52  0.0360 0.85057
## NW      1     22201 1327090 465.83  0.6357 0.43022
## U1      1     87235 1262056 463.62  2.6266 0.11336
## U2      1    149794 1199497 461.38  4.7455 0.03565 *
## GDP     1     36869 1312422 465.34  1.0675 0.30804
## Prob    1    108374 1240917 462.88  3.3187 0.07637 .
## Time    1      4027 1345264 466.43  0.1138 0.73776
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## --= Step (forward) 6 =--
## Single term additions
##
## Model:
## y ~ Po1 + Ineq + Ed + M + U2
##        Df Sum of Sq     RSS    AIC F value  Pr(>F)
## <none>              1199497 461.38
## So      1      1211 1198286 463.34  0.0374 0.84772
## Po2     1     42642 1156855 461.79  1.3638 0.25034
## LF      1       943 1198554 463.35  0.0291 0.86548
## M.F     1     11053 1188444 462.97  0.3441 0.56103
## Pop     1      6246 1193251 463.15  0.1937 0.66243
## NW      1     25736 1173761 462.43  0.8113 0.37357
## U1      1      2041 1197457 463.31  0.0631 0.80313
## GDP     1     20961 1178536 462.61  0.6581 0.42243
## Prob    1    126860 1072637 458.46  4.3760 0.04337 *
## Time    1      2658 1196839 463.28  0.0822 0.77597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## --= Step (forward) 7 =--
## Single term additions
##
## Model:
## y ~ Po1 + Ineq + Ed + M + U2 + Prob
##        Df Sum of Sq     RSS    AIC F value Pr(>F)
## <none>              1072637 458.46
## So      1     10351 1062286 460.04  0.3508 0.5574
## Po2     1     31664 1040973 459.14  1.0950 0.3023
## LF      1       779 1071858 460.43  0.0262 0.8724
## M.F     1     11081 1061557 460.01  0.3758 0.5437
## Pop     1     22553 1050084 459.53  0.7732 0.3851
## NW      1      2971 1069666 460.34  0.1000 0.7537
## U1      1      3890 1068747 460.30  0.1310 0.7195
## GDP     1         0 1072637 460.46  0.0000 0.9979
```

```
## Time    1     31921 1040717 459.13  1.1042 0.3004
##
## Call:
## lm(formula = y ~ Po1 + Ineq + Ed + M + U2 + Prob, data = Datos)
##
## Coefficients:
## (Intercept)          Po1         Ineq           Ed            M           U2
##   -5171.280       12.108        7.016       19.835       10.575        8.743
##        Prob
##     -37.296
```

5.2. Script de R

Aquí se mostrará el código R utilizado para el trabajo, para reproducirlo solo hace falta incluir en el mismo directorio de trabajo el archivo **USCrime.txt**, en el proceso se desplegaran las diferentes gráficas y tablas, así como el modelo completo *ModeloCompleto* el modelo reducido *ModeloRed* y el modelo reducido intervenido como *ModeloRedInter* en ambiente de trabajo.

6. Referencias

Libros consultados

[3] Julian James Faraway. Linear models with R. 2005.

[6] John Fox y Sanford Weisberg. An R Companion to Applied Regression. Third. Thousand Oaks CA: Sage, 2019. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

[22] Alvin C. Rencher y G. Bruce Schaalje. Linear models in statistics. Wiley-Interscience, 2008.

Paquetes de R

[1] Stefan Milton Bache y Hadley Wickham. magrittr: A Forward-Pipe Operator for R. R package version 2.0.1. 2020. URL: https://CRAN.R-project.org/package=magrittr.

[2] Andrew Bray y col. infer: Tidy Statistical Inference. R package version 0.5.4. 2021. URL: https://CRAN.R-project.org/package=infer.

[4] Thomas Farrar. skedastic: Heteroskedasticity Diagnostics for Linear Regression Models. R package version 1.0.3. 2021. URL: https://github.com/tjfarrar/skedastic.

[5] Thomas J. Farrar. skedastic: Heteroskedasticity Diagnostics for Linear Regression Models. R Package Version 1.0.0. Bellville, South Africa, 2020. URL: https://github.com/tjfarrar/skedastic.

[7] John Fox, Sanford Weisberg y Brad Price. car: Companion to Applied Regression. R package version 3.0-10. 2020. URL: https://CRAN.R-project.org/package=car.

[8] John Fox, Sanford Weisberg y Brad Price. carData: Companion to Applied Regression Data Sets. R package version 3.0-4. 2020. URL: https://CRAN.R-project.org/package=carData.

[9] Juergen Gross y Uwe Ligges. nortest: Tests for Normality. R package version 1.0-4. 2015. URL: https://CRAN.R-project.org/package=nortest.

[10] Lionel Henry y Hadley Wickham. purrr: Functional Programming Tools. R package version 0.3.4. 2020. URL: https://CRAN.R-project.org/package=purrr.

[11] Masaaki Horikoshi y Yuan Tang. ggfortify: Data Visualization Tools for Statistical Analysis Results. 2018. URL: https://CRAN.R-project.org/package=ggfortify.

[12] Masaaki Horikoshi y Yuan Tang. ggfortify: Data Visualization Tools for Statistical Analysis Results. R package version 0.4.11. 2020. URL: https://github.com/sinhrks/ggfortify.

[13] Torsten Hothorn y col. lmtest: Testing Linear Regression Models. R package version 0.9-38. 2020. URL: https://CRAN.R-project.org/package=lmtest.

[14] Max Kuhn. dials: Tools for Creating Tuning Parameter Values. R package version 0.0.9. 2020. URL: https://CRAN.R-project.org/package=dials.

[15] Max Kuhn. modeldata: Data Sets Used Useful for Modeling Packages. R package version 0.1.0. 2020. URL: https://CRAN.R-project.org/package=modeldata.

[16] Max Kuhn. tune: Tidy Tuning Tools. R package version 0.1.5. 2021. URL: https://CRAN.R-project.org/package=tune.

[17] Max Kuhn y Davis Vaughan. parsnip: A Common API to Modeling and Analysis Functions. R package version 0.1.6. 2021. URL: https://CRAN.R-project.org/package=parsnip.

[18] Max Kuhn y Hadley Wickham. Tidymodels: a collection of packages for modeling and machine learning usi 2020. URL: https://www.tidymodels.org.

[19] Max Kuhn y Hadley Wickham. tidymodels: Easily Install and Load the Tidymodels Packages. R package version 0.1.3. 2021. URL: https://CRAN.R-project.org/package=tidymodels.

[20] Kirill Müller. here: A Simpler Way to Find Your Files. R package version 1.0.1. 2020. URL: https://CRAN.R-project.org/package=here.

[21] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[23] Tyler Rinker y Dason Kurkiewicz. pacman: Package Management Tool. R package version 0.5.1. 2019. URL: https://github.com/trinker/pacman.

[24] David Robinson, Alex Hayes y Simon Couch. broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.6. 2021. URL: https://CRAN.R-project.org/package=broom.

[25] Barret Schloerke y col. GGally: Extension to ggplot2. R package version 2.1.1. 2021. URL: https://CRAN.R-project.org/package=GGally.

[26] Yuan Tang, Masaaki Horikoshi y Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages". En: The R Journal 8 (2 2016). URL: https://journal.r-project.org/.

[27] Adrian Trapletti y Kurt Hornik. tseries: Time Series Analysis and Computational Finance. R package version 0.10-48. 2020. URL: https://CRAN.R-project.org/package=tseries.

[28] Hadley Wickham. reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package. R package version 1.4.4. 2020. URL: https://github.com/hadley/reshape.

[29] Hadley Wickham. "Reshaping Data with the reshape Package". En: Journal of Statistical Software 21.12 (2007), págs. 1-20. URL: http://www.jstatsoft.org/v21/i12/.

[30] Hadley Wickham. tidyr: Tidy Messy Data. R package version 1.1.3. 2021. URL: https://CRAN.R-project.org/package=tidyr.

[31] Hadley Wickham y Jim Hester. readr: Read Rectangular Text Data. R package version 1.4.0. 2020. URL: https://CRAN.R-project.org/package=readr.

[32] Hadley Wickham y Dana Seidel. scales: Scale Functions for Visualization. R package version 1.1.1. 2020. URL: https://CRAN.R-project.org/package=scales.

[33] Hadley Wickham y col. dplyr: A Grammar of Data Manipulation. R package version 1.0.6. 2021. URL: https://CRAN.R-project.org/package=dplyr.

[34] Hadley Wickham y col. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.3.3. 2020. URL: https://CRAN.R-project.org/package=ggplot2.

[35] Hadley Wickham y col. "Welcome to the tidyverse". En: Journal of Open Source Software 4.43 (2019), pág. 1686. DOI: 10.21105/joss.01686.

[36] Yihui Xie. "knitr: A Comprehensive Tool for Reproducible Research in R". En: Implementing Reproducible C Ed. por Victoria Stodden, Friedrich Leisch y Roger D. Peng. ISBN 978-1466561595. Chapman y Hall/CRC, 2014. URL: http://www.crcpress.com/product/isbn/9781466561595.

[37] Yihui Xie. knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.33. 2021. URL: https://yihui.org/knitr/.

[38] Achim Zeileis y Gabor Grothendieck. "zoo: S3 Infrastructure for Regular and Irregular Time Series". En: Journal of Statistical Software 14.6 (2005), págs. 1-27. DOI: 10.18637/jss.v014.i06.

[39] Achim Zeileis, Gabor Grothendieck y Jeffrey A. Ryan. zoo: S3 Infrastructure for Regular and Irregular Time R package version 1.8-9. 2021. URL: https://zoo.R-Forge.R-project.org/.

[40] Achim Zeileis y Torsten Hothorn. "Diagnostic Checking in Regression Relationships". En: R News 2.3 (2002), págs. 7-10. URL: https://CRAN.R-project.org/doc/Rnews/.

[41] Hao Zhu. kableExtra: Construct Complex Table with kable and Pipe Syntax. R package version 1.3.4. 2021. URL: https://CRAN.R-project.org/package=kableExtra.