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 (análisis de multicolinealidad).

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.

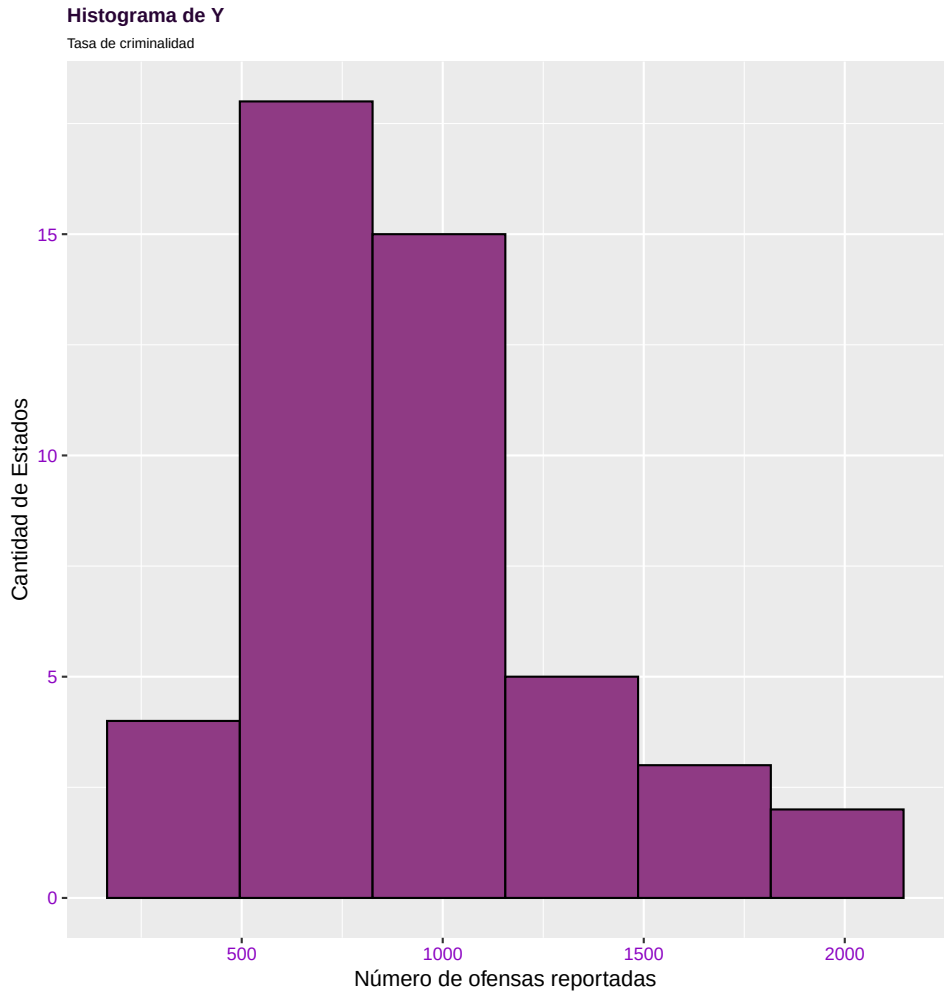


Figura 1: Histograma de la Tasa de Criminalidad

Por último, el histograma de la variable Y. Cuenta con una distribución asimétrica, tiene un intervalo modal entre los valores de 500 y 800 ofensas reportadas en casi 13 Estados. Podemos afirmar que la mayoría de las observaciones están concentradas entre 342 que es su valor mínimo y 831 que es la mediana. Dado que su valor máximo es el 1993, el histograma cuenta con una cola hacia la derecha.

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

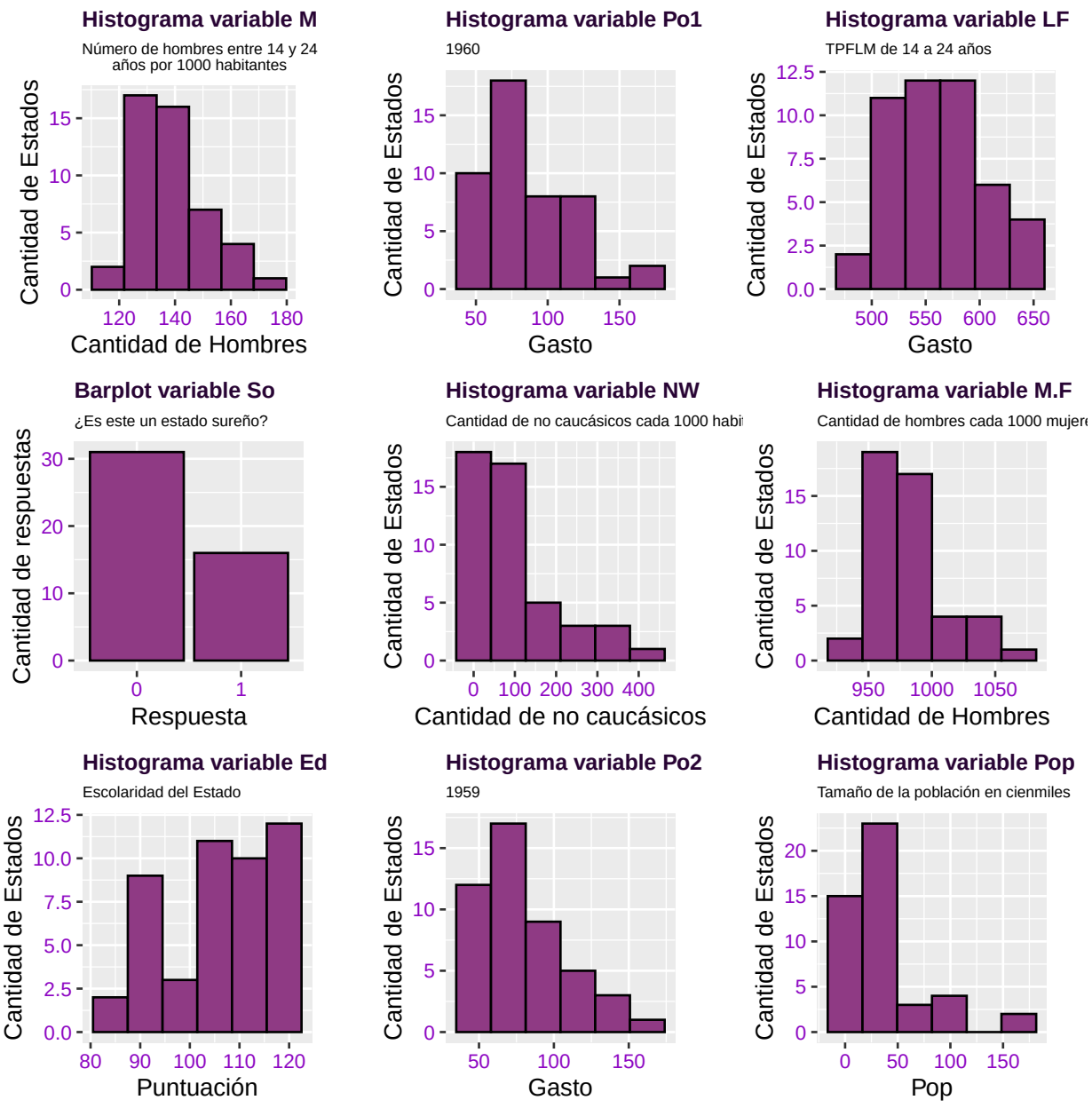


Figura 2: Histogramas (1)

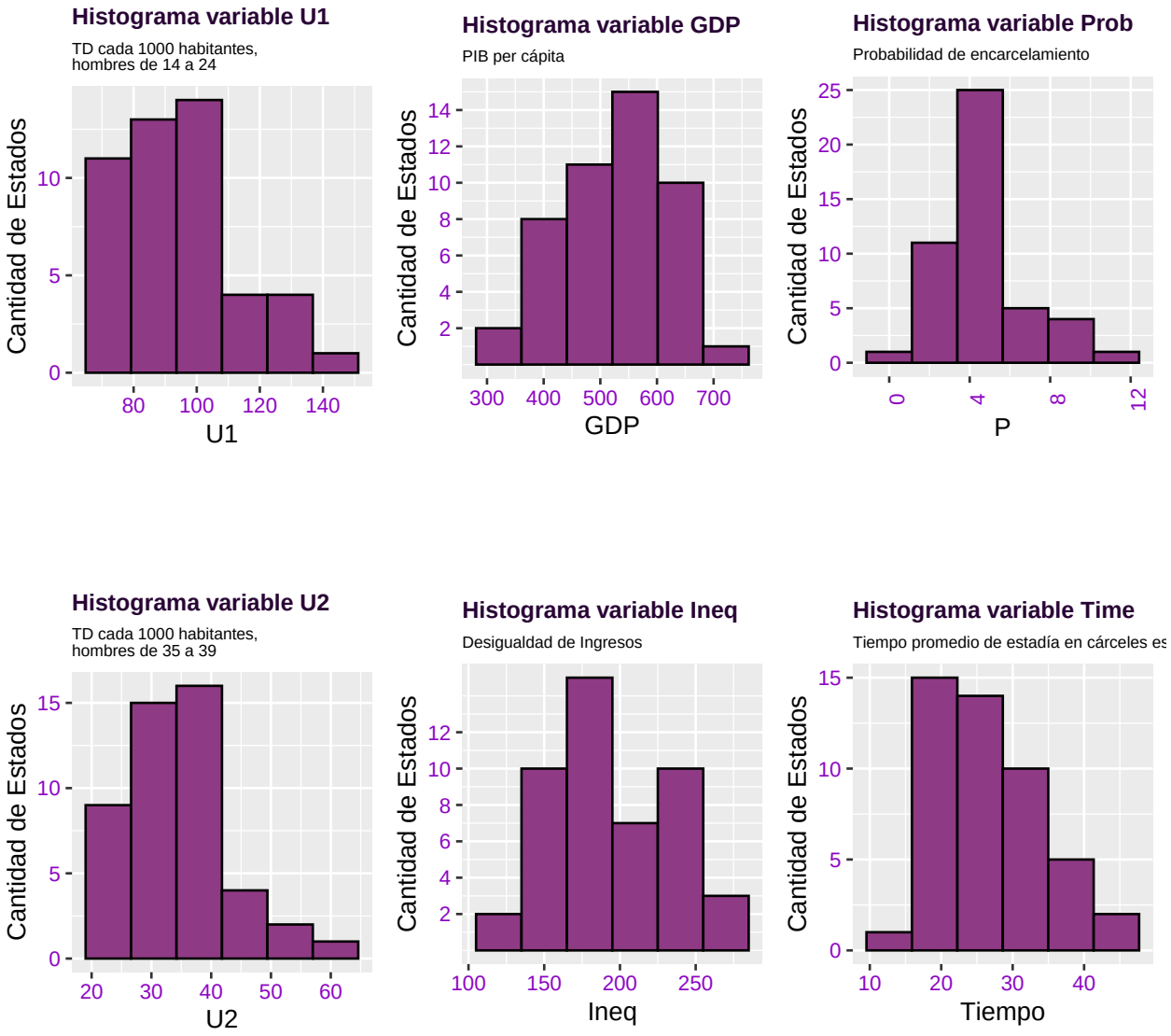


Figura 3: Histogramas (2)

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

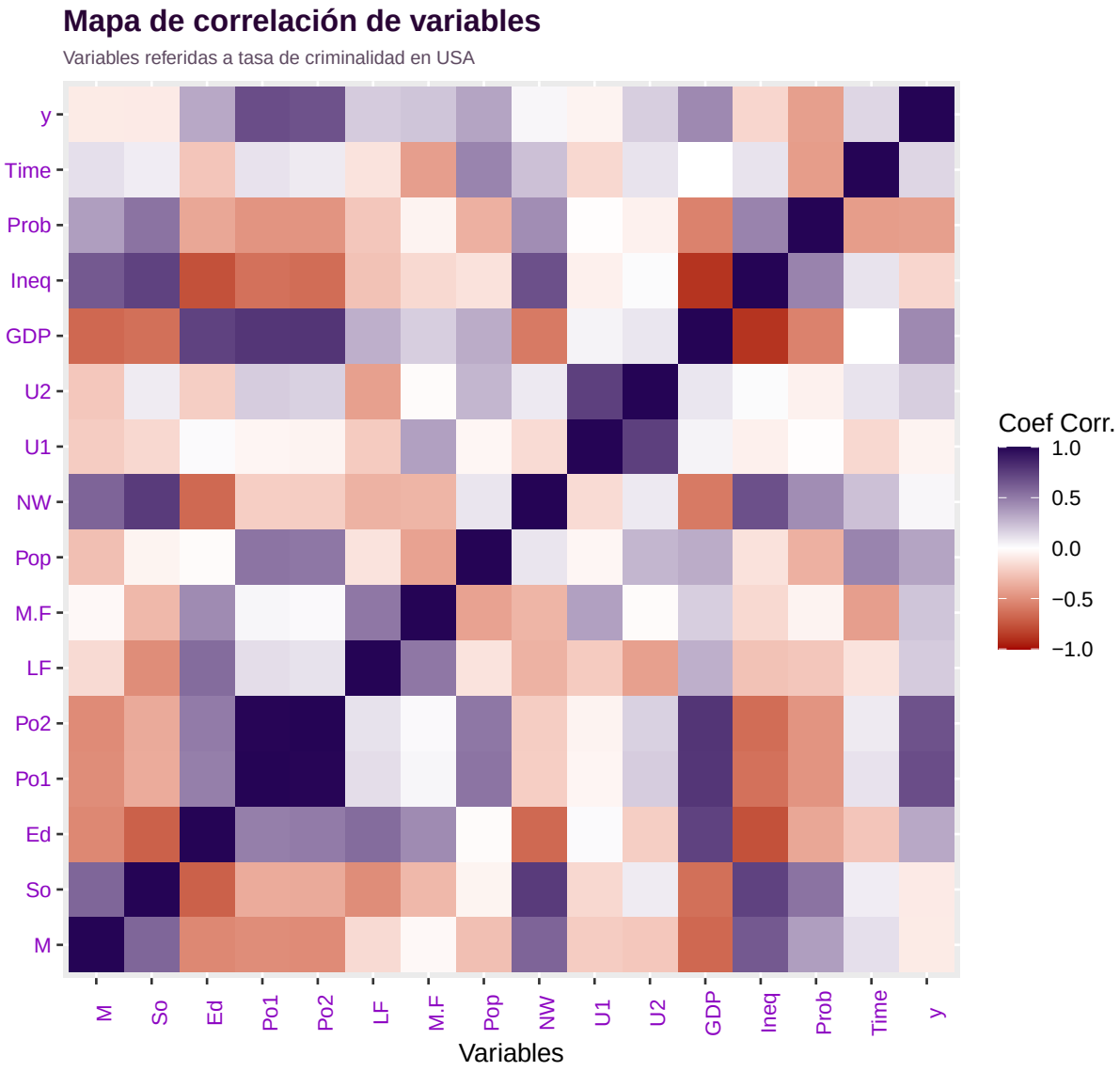
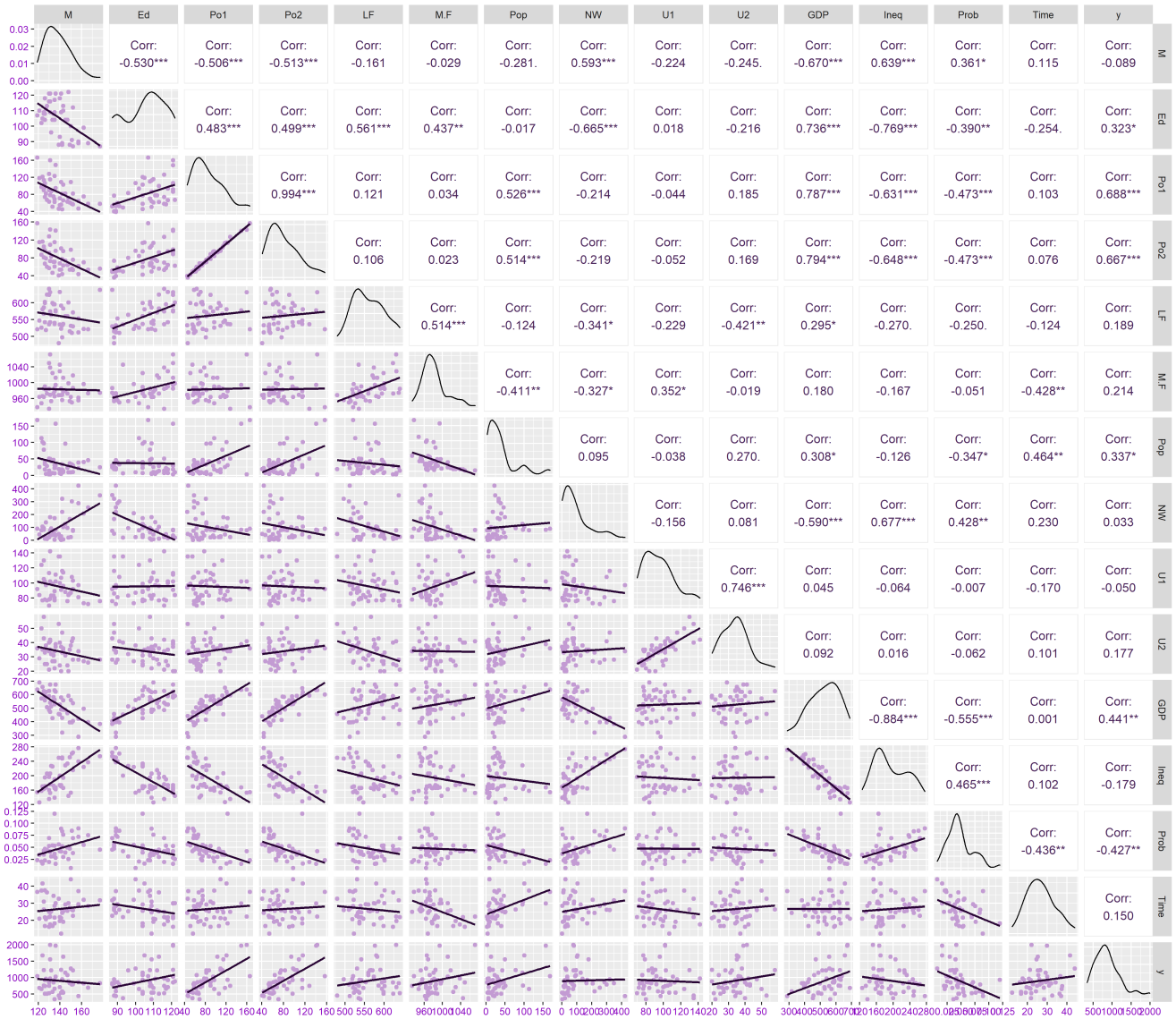


Figura 4: Mapa de correlación de variables incluidas


M
Ed
Po1
Po2
LF
M.F
Pop
NW
U1
U2
GDP
Ineq
Prob
Time
y
Corr: -0.530***
Corr: -0.506***
Corr: -0.513***
Corr: -0.161
Corr: -0.029
Corr: -0.281.
Corr: 0.593***
Corr: -0.224
Corr: -0.245.
Corr: -0.670***
Corr: 0.639***
Corr: 0.361*
Corr: 0.115
Corr: -0.089
Corr: 0.483***
Corr: 0.499***
Corr: 0.561***
Corr: 0.437**
Corr: -0.017
Corr: -0.665***
Corr: 0.018
Corr: -0.216
Corr: 0.736***
Corr: -0.769***
Corr: -0.390**
Corr: -0.254.
Corr: 0.323*
Corr: 0.994***
Corr: 0.121
Corr: 0.034
Corr: 0.526***
Corr: -0.214
Corr: -0.044
Corr: 0.185
Corr: 0.787***
Corr: -0.631***
Corr: -0.473***
Corr: 0.103
Corr: 0.688***
Corr: 0.106
Corr: 0.023
Corr: 0.514***
Corr: -0.219
Corr: -0.052
Corr: 0.169
Corr: 0.794***
Corr: -0.648***
Corr: -0.473***
Corr: 0.076
Corr: 0.667***
Corr: 0.514***
Corr: -0.124
Corr: -0.341*
Corr: -0.229
Corr: -0.421**
Corr: 0.295*
Corr: -0.270.
Corr: -0.250.
Corr: -0.124
Corr: 0.189
Corr: -0.411**
Corr: -0.327*
Corr: 0.352*
Corr: -0.019
Corr: 0.180
Corr: -0.167
Corr: -0.051
Corr: -0.428**
Corr: 0.214
Corr: 0.095
Corr: -0.038
Corr: 0.270.
Corr: 0.308*
Corr: -0.126
Corr: -0.347*
Corr: 0.464**
Corr: 0.337*
Corr: -0.156
Corr: 0.081
Corr: -0.590***
Corr: 0.677***
Corr: 0.428**
Corr: 0.230
Corr: 0.033
Corr: 0.746***
Corr: 0.045
Corr: -0.064
Corr: -0.007
Corr: -0.170
Corr: -0.050
Corr: 0.092
Corr: 0.016
Corr: -0.062
Corr: 0.101
Corr: 0.177
Corr: -0.884***
Corr: -0.555***
Corr: 0.001
Corr: 0.441**
Corr: 0.465***
Corr: 0.102
Corr: -0.179
Corr: -0.436**
Corr: -0.427**
Corr: 0.150

3. Especificación y selección de modelos

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
63.452	233.818	7.655	0	12	34

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 $RSE^2 = 43707,93$, 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	-5251.313	1715.203	-3.062	0.004	Se rechaza H0
Número de hombres entre 14 y 24 / 1000	8.874	4.460	1.990	0.055	No se rechaza H0
Variable indicadora de loes estados del sur	225.677	134.660	1.676	0.103	No se rechaza H0
Indice que refleja la escolaridad del estado	12.767	6.578	1.941	0.061	No se rechaza H0
Gasto per cápita en policía en 1960	9.689	2.392	4.050	0.000	Se rechaza H0
Tasa de participación masculina 14 y 24 / 1000	0.971	1.467	0.662	0.513	No se rechaza H0
Número de hombres por cada 1000 mujeres	2.950	2.177	1.355	0.184	No se rechaza H0
Tamaño de la población por 1.000.000 habitantes	0.232	1.396	0.166	0.869	No se rechaza H0
Tasa de desmempleo urbana hombres 14 y 24 por 1000	-5.809	4.609	-1.260	0.216	No se rechaza H0
Tasa de desmempleo urbana hombres 35 y 39 por 1000	19.382	9.160	2.116	0.042	Se rechaza H0
Producto bruto interno per cápita	-1.236	0.884	-1.398	0.171	No se rechaza H0
Probabilidad de encarcelamiento	-48.711	23.519	-2.071	0.046	Se rechaza H0
Tiempo promedio de estadía en cárceles estatales	-0.030	7.322	-0.004	0.997	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).

3.2. 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 entre 14 y 24 / 1000	2.643	No hay problema de colinealidad
Variable indicadora de loes estados del sur	3.500	No hay problema de colinealidad
Indice que refleja la escolaridad del estado	4.557	No hay problema de colinealidad
Gasto per cápita en policía en 1960	4.252	No hay problema de colinealidad
Tasa de participación masculina 14 y 24 / 1000	2.958	No hay problema de colinealidad
Número de hombres por cada 1000 mujeres	3.462	No hay problema de colinealidad
Tamaño de la población por 1.000.000 habitantes	2.377	No hay problema de colinealidad
Tasa de desmempleo urbana hombres 14 y 24 por 1000	5.811	No hay problema de colinealidad
Tasa de desmempleo urbana hombres 35 y 39 por 1000	5.035	No hay problema de colinealidad
Producto bruto interno per cápita	6.126	No hay problema de colinealidad
Probabilidad de encarcelamiento	2.406	No hay problema de colinealidad
Tiempo promedio de estadía en cárceles estatales	2.265	No hay problema de colinealidad

En base a esto, podemos afirmar que nuestro modelo no presenta problemas con la colinealidad y es por esto que continuaremos con la selección a pasos por el método Stepwise.

3.3. Método Stepwise

```
## Stepwise regression (forward-backward), alpha-to-enter: 0.15, alpha-to-remove: 0.15
##
## Full model: y ~ M + So + Ed + Po1 + LF + M.F + Pop + U1 + U2 + GDP + Prob +
##     Time
## <environment: 0x00000000237bdd90>
##
##      Step InOut     RSS    AIC  R2pred      Cp F value    Pr(>F)
## Po1     1     1 3627626 532.94 0.39260 23.3539 40.3566 9.338e-08 ***
## M       2     1 3010885 526.18 0.48030 14.0729  9.0128  0.004407 **
## M.F     3     1 2749621 523.91 0.51089 11.2941  4.0858  0.049498 *
## Prob    4     1 2576451 522.85 0.52091 10.1266  2.8229  0.100351
## So      5     1 2295000 519.42 0.55641  6.9785  5.0281  0.030410 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

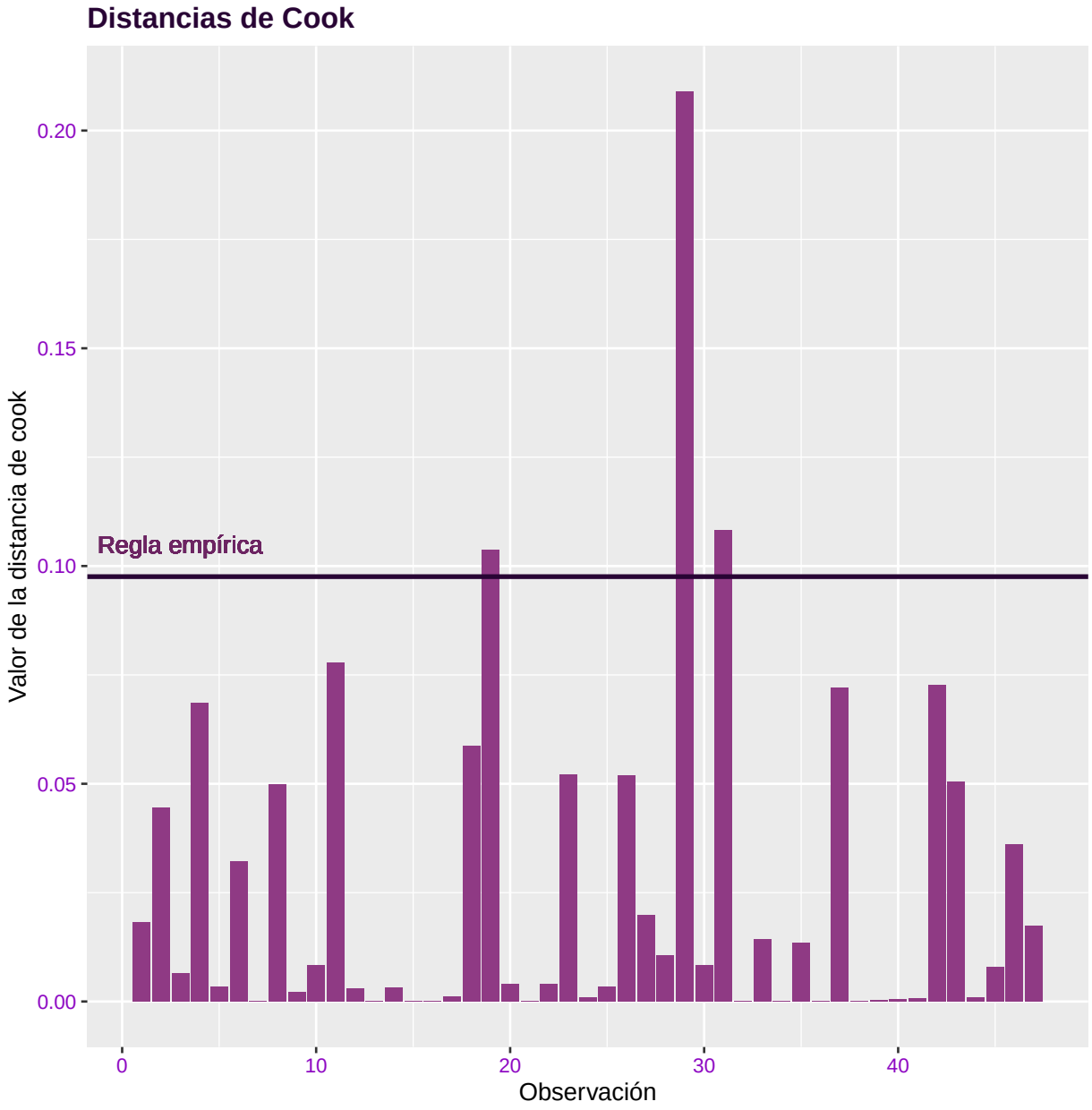
Variable	Estimación	Error estandar	Estadístico F	P valor	$\left(H_0^{\alpha=0.05}\right) \beta_i = 0$
Intercepto	-4354.40060	1297.991695	-3.354722	0.0017203	Se rechaza H0
Gasto per cápita en policía en 1960	10.05256	1.461604	6.877755	0.0000000	Se rechaza H0
Número de hombres entre 14 y 24 / 1000	7.40109	3.779373	1.958285	0.0570240	No se rechaza H0
Número de hombres por cada 1000 mujeres	3.59328	1.285195	2.795901	0.0078413	Se rechaza H0
Probabilidad de encarcelamiento	-49.88420	19.551501	-2.551425	0.0145548	Se rechaza H0
Variable indicadora de loes estados del sur	241.04366	107.496487	2.242340	0.0304103	Se rechaza H0

3.4. 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 $\frac{4}{n-k-1}$ puede verse que las observaciones **19**, **29** (de manera excesiva) y **31** 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

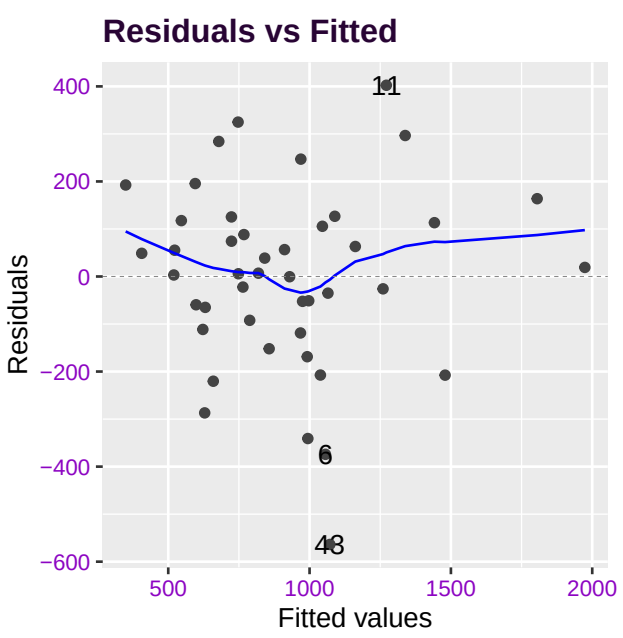
alejado de la realidad.

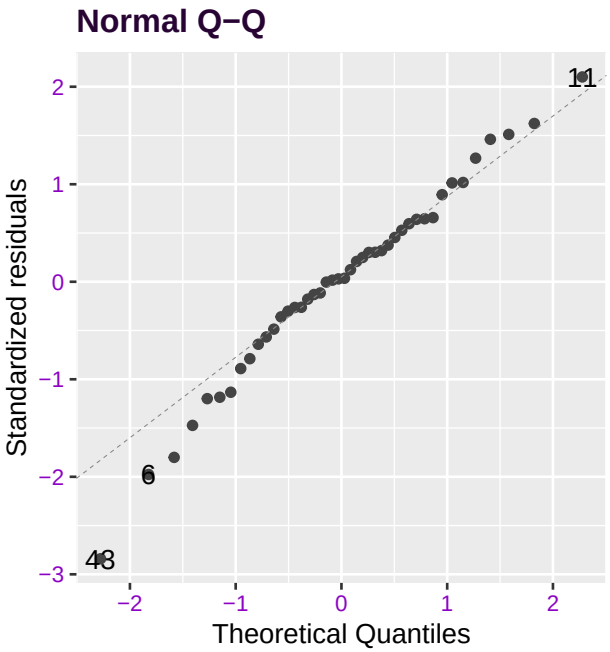
4. Diagnostico

5. Conclusiones

6. Normalidad

```
## Stepwise regression (forward-backward), alpha-to-enter: 0.15, alpha-to-remove: 0.15
##
## Full model: y ~ M + So + Ed + Po1 + LF + M.F + Pop + U1 + U2 + GDP + Prob +
##     Time
## <environment: 0x0000000026800600>
##
##      Step InOut     RSS    AIC  R2pred      Cp F value    Pr(>F)
## Po1     1     1 2739055 489.71 0.53357 27.7426 58.4037 1.778e-09 ***
## M       2     1 2172348 481.51 0.61707 15.7267 10.6958   0.00218 **
## Prob    3     1 1934376 478.41 0.63456 11.8412  4.9209   0.03227 *
## GDP     4     1 1720910 475.26 0.67248  8.5617  4.8377   0.03384 *
## M.F     5     1 1589939 473.78 0.68936  7.3225  3.1303   0.08488 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





Considerando los gráficos anteriormente mencionados podemos ver que la esperanza de los errores se mantiene cercana a cero, a diferencia de algunas observaciones, a su vez , no encontramos un patrón en la dispersión de los errores, en cambio en la gráfica QQ-plot podemos ver que en valores centrales de la distribución se asemeja a una distribución normal a exepción de la observación 46,11 y 6.

```
## Error: Problem with ‘summarise()‘ column ‘Shapiro‘.
## i ‘Shapiro = shapiro.test(as.numeric(.))‘.
## x ‘Shapiro‘ must be a vector, not a ‘htest‘ object.
## Error in stopifnot(is.numeric(x)): el argumento "x" está ausente, sin valor por omisión
```

En base a los test puedo afirtmar que la distribución de los errores es normal.
Me tranque en como presentar una tabla , lo miro luego