Tasa de criminalidad en USA

Modelos Líneales

Trabajo final

Ignacio Acosta - Sofía Itté - Mauro Loprete

1er semestre 2021

Índice

Índice de figuras

1. Introducción

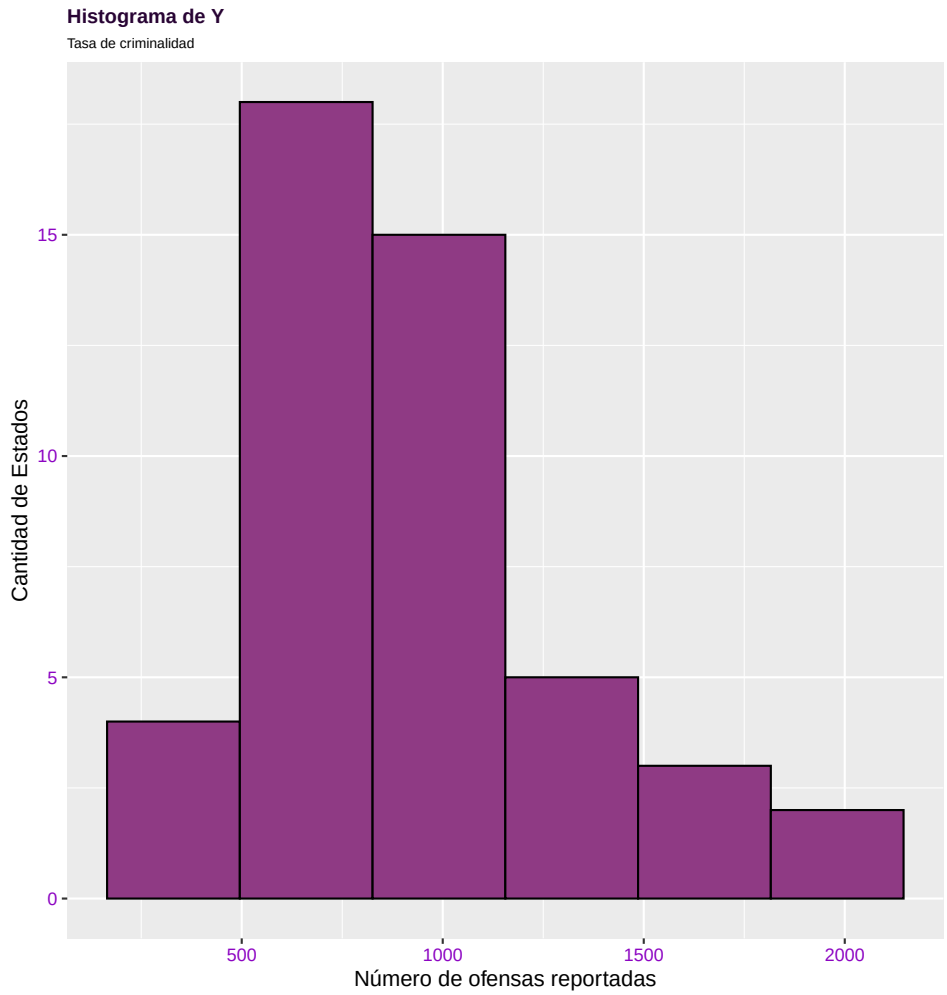


Figura 1: Histograma de la Tasa de Criminalidad

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

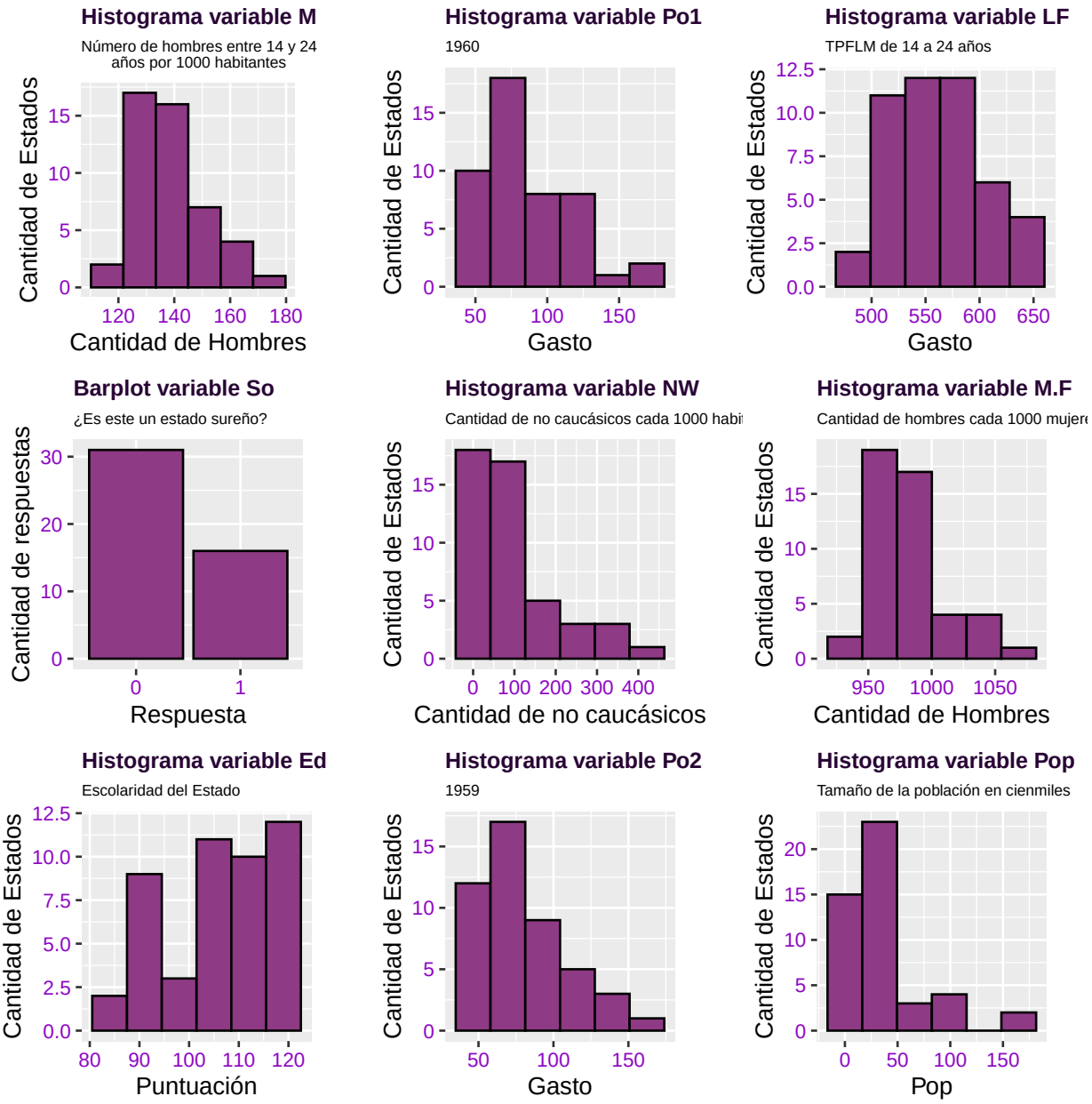


Figura 2: Histogramas (1)

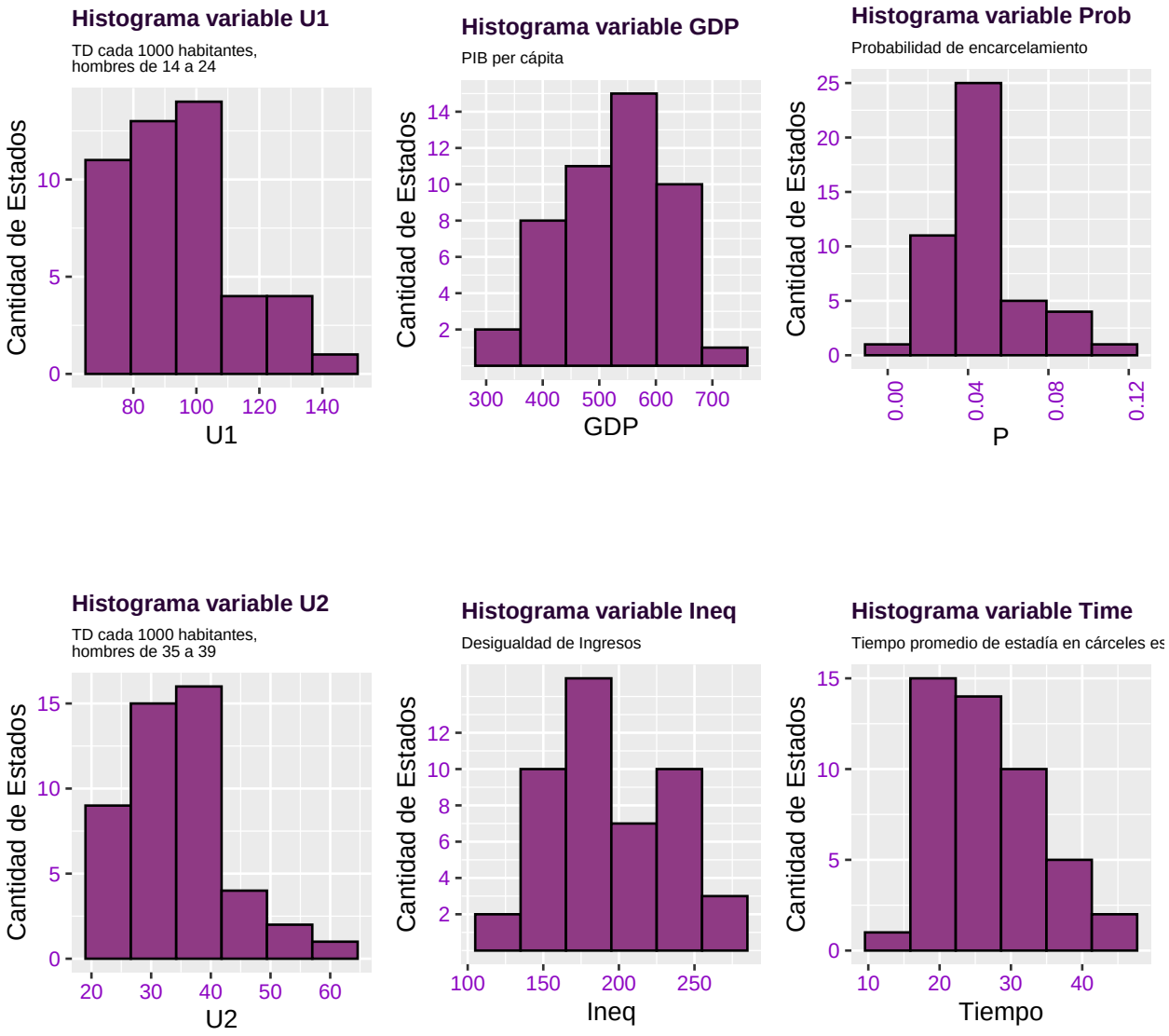


Figura 3: Histogramas (2)

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	0.0069	0.0327	0.0421	0.0544	0.1198	0.0471	48.2827
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

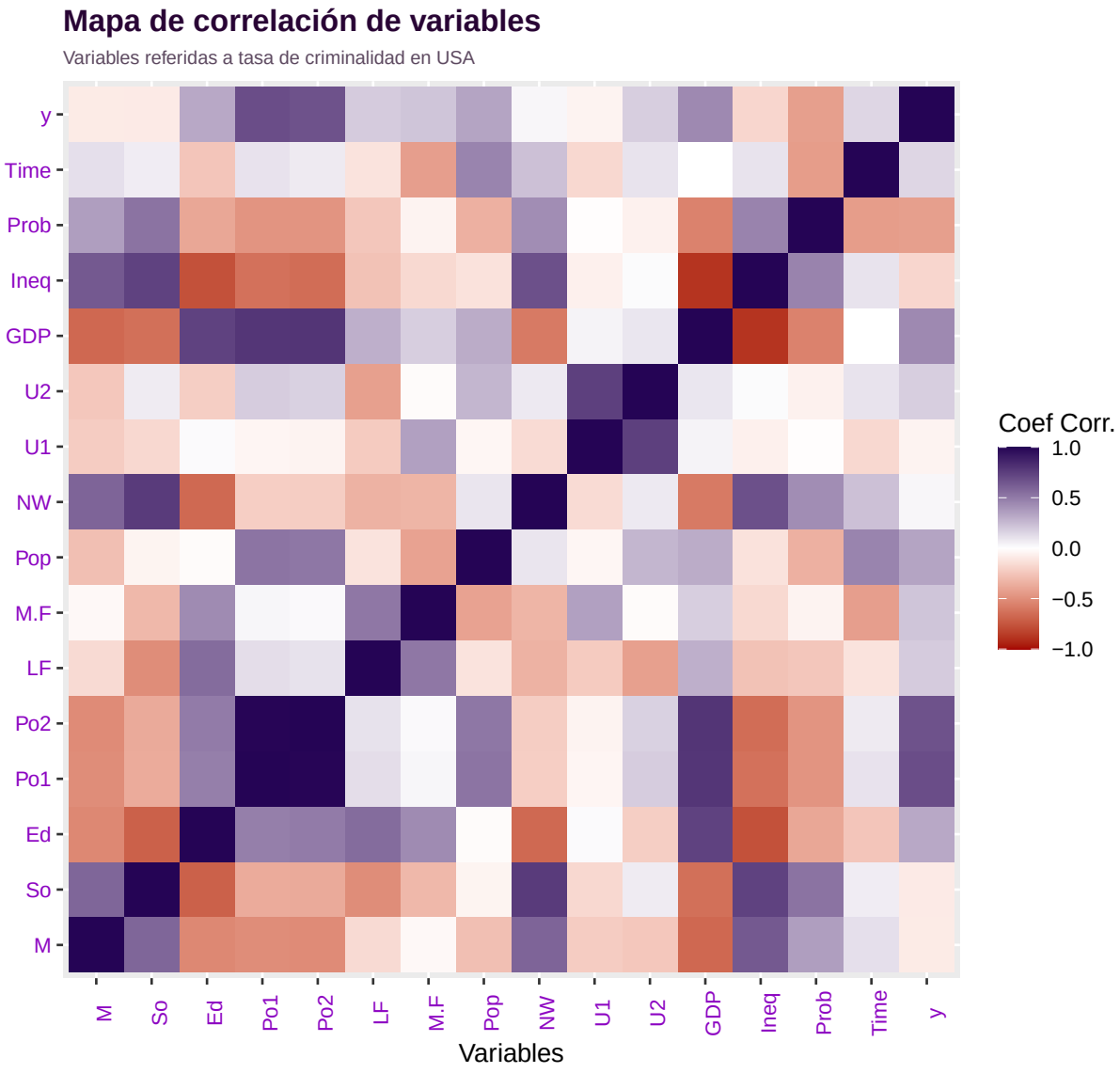


Figura 4: Mapa de correlación de variables incluidas

3. Especificación y selección de modelos

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)

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**.

$$\rho_{y,Ineq} = -0,179 \qquad \rho_{y,GDP} = 0,441 \qquad \rho_{y,P01} = 0,688 \qquad \rho_{y,P02} = 0,667 \qquad \rho_{y,Ed} = 0,323$$

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

Trás esto, se decide que **Ineq** y **P02** 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].

Otro aspecto a considerar es , ¿qué tan correlacionadas se encuentran las mismas variables con y?. Un caso partícular es el de las variables **S01** y **NW** que claramente se encuentran relacionadas (con un coeficiente de correlación de 0,767). A su vez resalta como ambas variables tienen una correlación con **y** prácticamente despreciable

$$\rho_{y,S0} = -0,09 \qquad \rho_{y,NW} = 0,03$$

[1] Frugalidad y moderación en los gastos.

En base a lo último es claro que estas 2 variables no serán significativas para explicar a **y** y que en cualquier proceso de selección computarizado serán descartadas.

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

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 $RSE^2 = 43707,93$, 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	-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	-4855.266	2272.375	-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 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 *U1,Ineq y Prob.*

¿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. Aca iría la selección de modelos

4. Diagnostico

5. Conclusiones

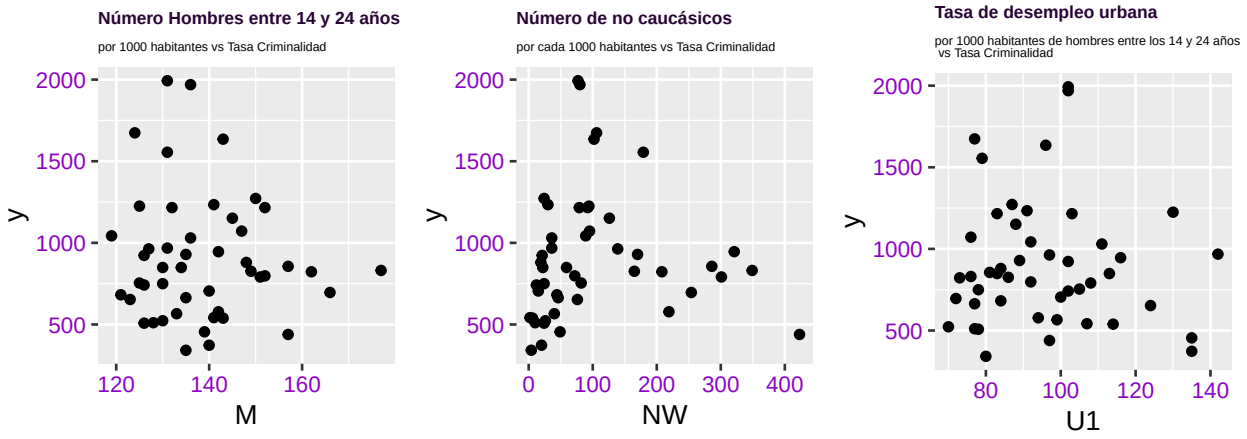
6. discutir

En este apartado se explicitará el procedimiento de especificación y selección de variables para que dada la información disponible explicar la *Tasa de Criminalidad en el 1960 para los Estados de USA*.

Considerando los ánalisis anteriores, antes de asignarle una forma a cada variable, utilizando diagramas de dispersión la para ver la relación que tienen con la variable Respuesta, haciendo enfasis en aquellas que en el mapa de correlación se presentan colores gris claro o blancos, en concreto:

- **M** : *Número de hombres entre 14 y 24 años por 1.000 habitantes*
- **NW** : *Número de no caucásicos por cada 1000 habitantes*
- **U1** : *Tasa de desempleo urbana por 1000 habitantes de hombres entre los 14 y 24 años*

La justificación de esto se debe a que el coeficiente correlación toma en cuenta solo relaciones lineales por lo que deberíamos de indagar un poco mas la interacción, siendo crucial en la inclusión en el modelo.



A la luz de los gráficos no podemos establecer una transformación para ninguna de las tres variables, aunque hay que considerar que para la variable **NW** podemos asegurar que no existe un patrón que indique relación entre las dos variables y a exepción de **Po1** *El gasto per cápita en policía por el gobierno en 1960* y **Po2** *El gasto per cápita en policía por el gobierno en 1960*, la correlación es considerable, estando implicita en las demás variables por lo que no se incluirá en el modelo completo.

Por otra parte, para **M** y **U1** no podemos establecer una forma çonocida"pero podemos afirmar una relación creciente por lo no que serán descartadas, de esta manera, el modelo completo tendrá todas las variables a exepción de *Número de no caucásicos por cada 1.000 habitantes* **NW**.

En la siguiente sección mostraremos el procedimiento que en primera instancia, un modelo completo sin transformaciones previas de las variables independientes.