Tasa de criminalidad en USA

Modelos Líneales

Trabajo final

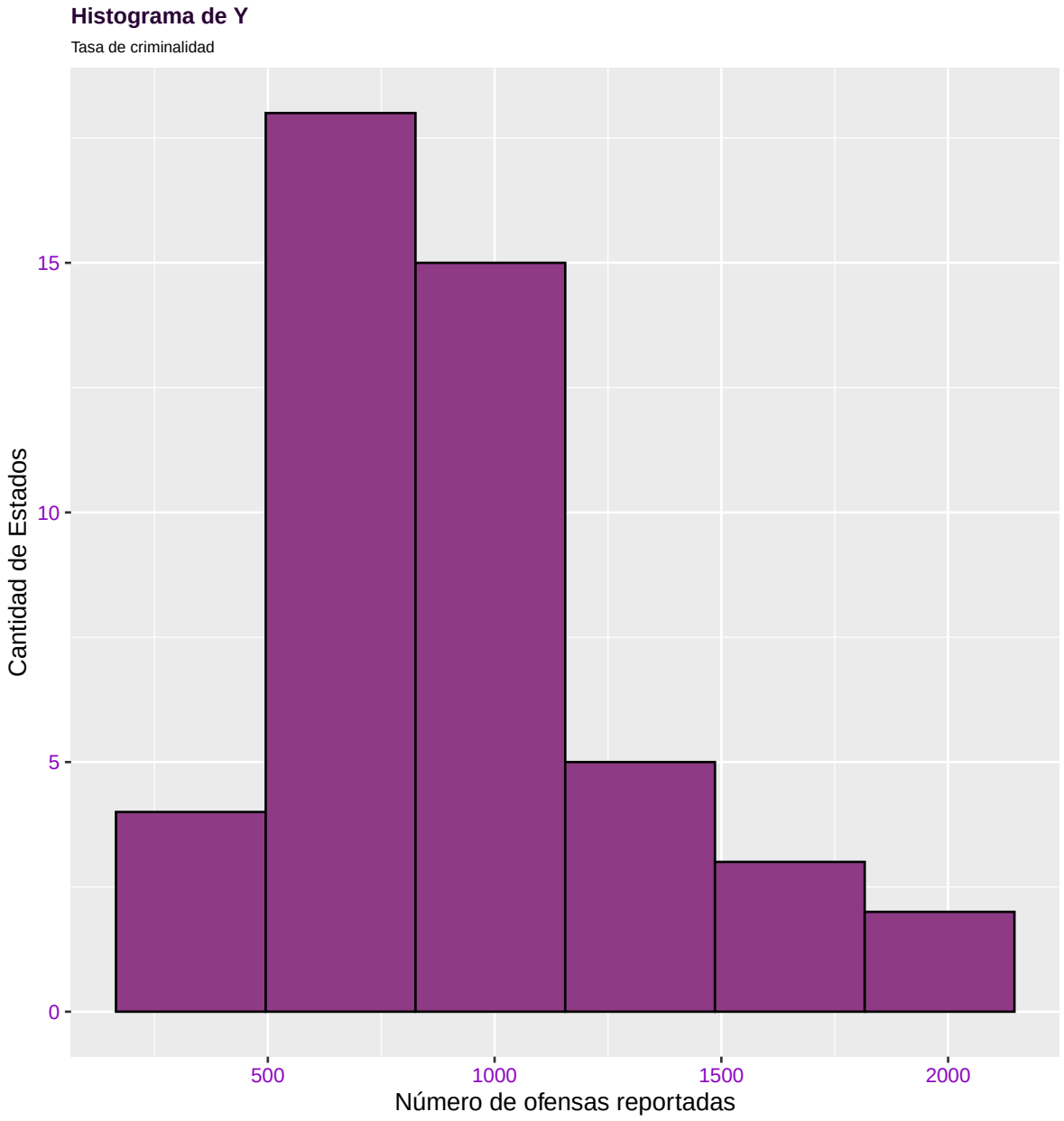
Ignacio Acosta - Sofía Itté - Mauro Loprete

1er semestre 2021

Índice

Índice de figuras

1. Introducción



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.

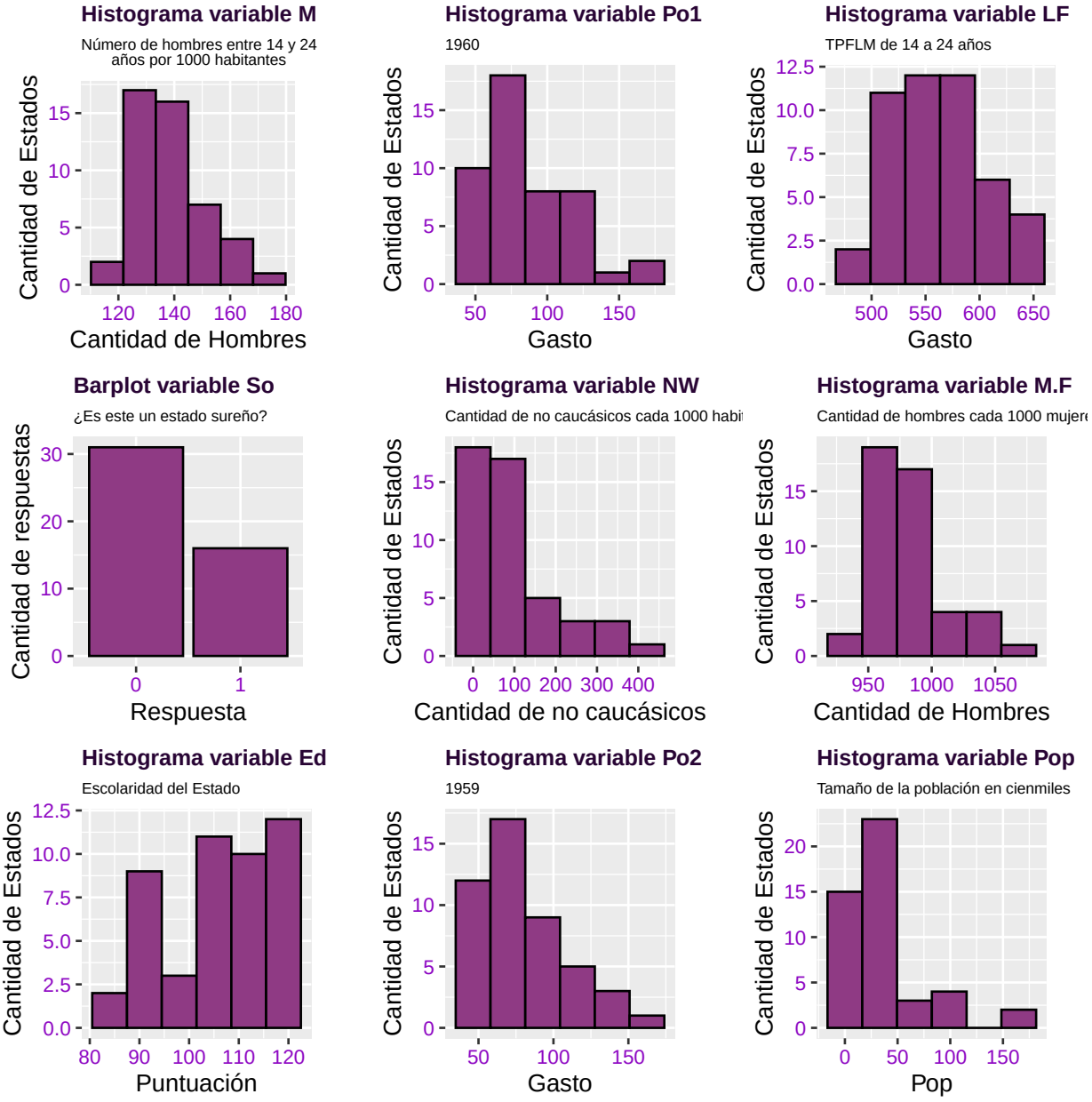
2.1. Análisis Univariado

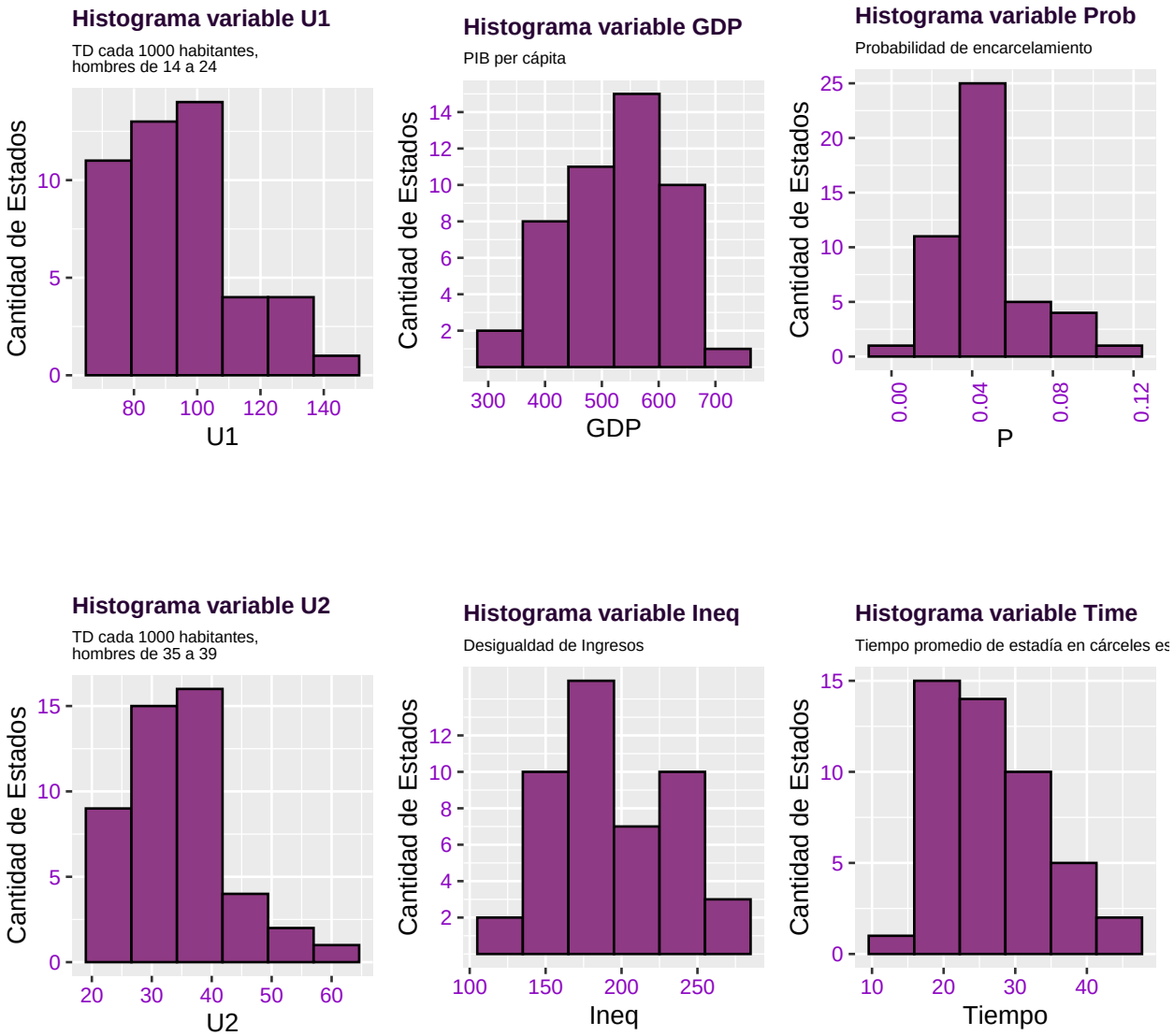
En esta primer sección se hará especial enfásis en las variables por sí mismas.

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

2.1.1. Histogramas y Barplots





Los histogramas de las variables M, Po1, Nw, Po2, M.F, Pop, U1, U2, Prob y Time cuentan con una distribución asimétrica. Cuenta con menor variabilidad entre el valor mínimo y la mediana, dejado asi mayor variabilidad de observaciones entre la mediana y el valor máximo generando una cola hacia la derecha. Por otro lado, los histogramas de las variables Ed y GDP también cuentan con una distribución asimétrica, pero en este caso cuenta con menor variabilidad entre la mediana y el valor máximo, dejando mayor variabilidad entre el valor mínimo y la mediana generando asi una cola hacia la izquierda. La variable LF cuenta con una distribución casi simétrica ya que su media y su mediana solo difieren en 1. Tiene un intervalo modal que va desde el valor 530 a 590. La variable ineq cuenta con una distribución asimétrica, tiene un intervalo modal entre los valores de ... a.... Cuenta con menor variabilidad entre el primer y el tercer cuartil. 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.

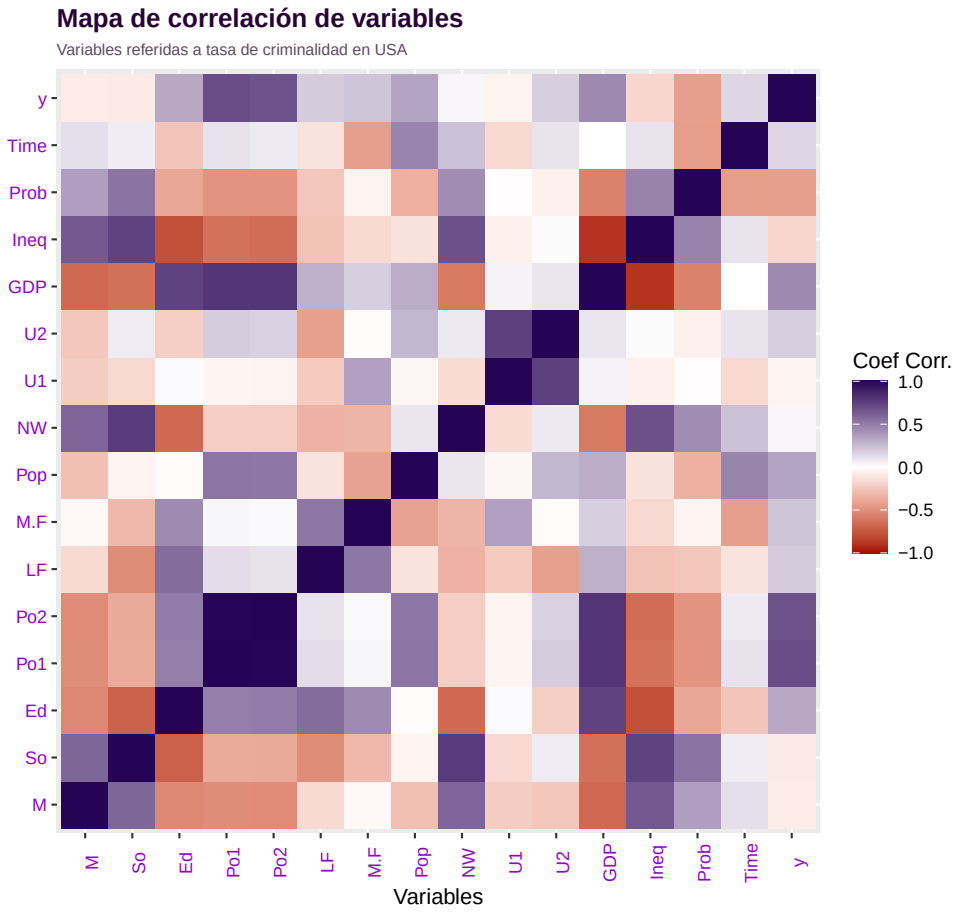
2.1.2. Medidas de resumen

Continuando el ánalisis de los gráficos presentados anteriormente, se presenta en forma de tabla el resumen de las variables númericas. A

Cuadro 1: Resumen descriptivo (Variables númericas)

Variable	Min	1er Qu.	Mediana	3er Qu.	Max	Media	Desvío.Std
Número de Hombres 14-24 / 1.000	119	130	136	146	177	139	13
Indice Escolaridad	87	98	108	114	122	106	11
Gasto per cápita 1.960	45	62	78	104	166	85	30
Gasto per cápita 1.959	41	58	73	97	157	80	28
Tasa participación masculina 14-24 por 1.000	480	530	560	593	641	561	40
Hombres cada 1.000 mujeres	934	964	977	992	1071	983	29
Población cada 100.000	3	10	25	42	168	37	38
Número de no caucásicos cada 1.000 habitantes	2	24	76	132	423	101	103
Tasa desempleo urbana Hombres 14-24 por 1.000	70	80	92	104	142	95	18
Tasa desempleo urbana Hombres 35-39 por 1.000	20.0	27.5	34.0	38.5	58.0	34.0	8.4
Producto bruto interno per cápita	288	460	537	592	689	525	96
Desigualdad ingreso	126	166	176	228	276	194	40
Probabilidad Encarcelamiento	0.0069	0.0327	0.0421	0.0544	0.1198	0.0471	0.0227
Tiempo de estadía en carceles	12.2	21.6	25.8	30.5	44.0	26.6	7.1
Tasa de criminalidad	342	658	831	1058	1993	905	387

2.1.3. Correlación entre variables



3. Especificación y selección de modelos

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

3.1. Modelo completo

Como una primera aproximación, se construye un modelo donde se incluyen todas las variables de nuestra tabla de datos.

Cuadro 2: 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

En base al cuadro anterior se puede apreciar que si hacemos hincapie al R_a^2 que hace referencia a la variabilidad de que podemos explicar con nuestro modelo respecto a la tasa de criminalidad, se puede considerar un valor *aceptable*.
Si consideramos la siguiente prueba $H_0)\beta_1 = \beta 2 = \cdots = \beta_k = 0$ y el siguiente estadístco de prueba:

$$F_{obs} = \frac{SCE/Regresion.gl}{RSE^2} \tag{1}$$

Siendo SCE la suma de cuadrados explicados por la regresión $\sum (\hat{y}_i - \bar{y})^2$ y $RSE^2 = SCR/Residos.gl = \frac{\sum (y_i - \hat{y}_i)^2}{Residuos.gl}$ el cuadrado del error estandar de los Residuos, resultando para este caso particular $SCE =$ 5,525,982 y $RSE^2 =$ 43707,93 podemos obtener el F_{obs} que nos permite rechazar H_0. De esta manera, podemos decir que el modelo es globalmente significativo,a continuación veremos el test de significación de cada variable.

Cuadro 3: 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

En base al cuadro anterior se puede ver que no existe evidencia estadístca suficiente para rechazar que no existe una relación lineal de las variables incluidas para explicar la *tasa de criminalidad* a exepción de la *Cantidad de Hombres de 14-24 cada mil habitantes,la desigualdad en el ingreso* y la *probabilidad de encarcelamiento.*
En resumen, el modelo es globalmente significativo, pero en solo 3 variables podemos afirmar que sean significativas en el modelo, es por esto que respetando el concepto de parsimonia y por una alta correlación entre variables, debemos de trabajar en la selección de las variables a utilizar.

3.2. Aca iría la selección de modelos

4. Diagnostico

5. Conclusiones