Tasa de Criminalidad

Trabajo modelos lineales

Belén Retamosa
Valentina Torres

Introducción

Se desea explicar la tasa de criminalidad medida como cantidad de crímenes cometidos per cápita (Variable de respuesta). Para explicar dichas variables disponemos de información de cada municipio.

Cada uno de los mismos recabaron datos de 14 variables, las cuales estudiaremos y veremos la influencia que posee cada una sobre nuestra variable de interés.

- Las variables con las que trabajaremos serán:
- Crímenes cometidos por persona (crmtr)
- Probabilidad de ser arrestado (prbarr)
- Probabilidad de sentendencia con prisión (pbrpris)
- Promedio de dias sentenciados(avgsen)
- Cantidad de policías por 1000 (polpc)
- Personas por milla cuadrada (density)
- Impuesto a la renta per capita (taxpc)
- Porcentaje de "minorias" (pctmin)
- Salario semana en la construcción (wcon)
- Salario semanal de servicios financieros (wfir)
- Salario semanal en servicios (wser)
- Salario semanal empleados federales (wfed)
- Ofensa cara a cara u otras (mix)
- Porcentaje de hombres jóvenes (pctymle)

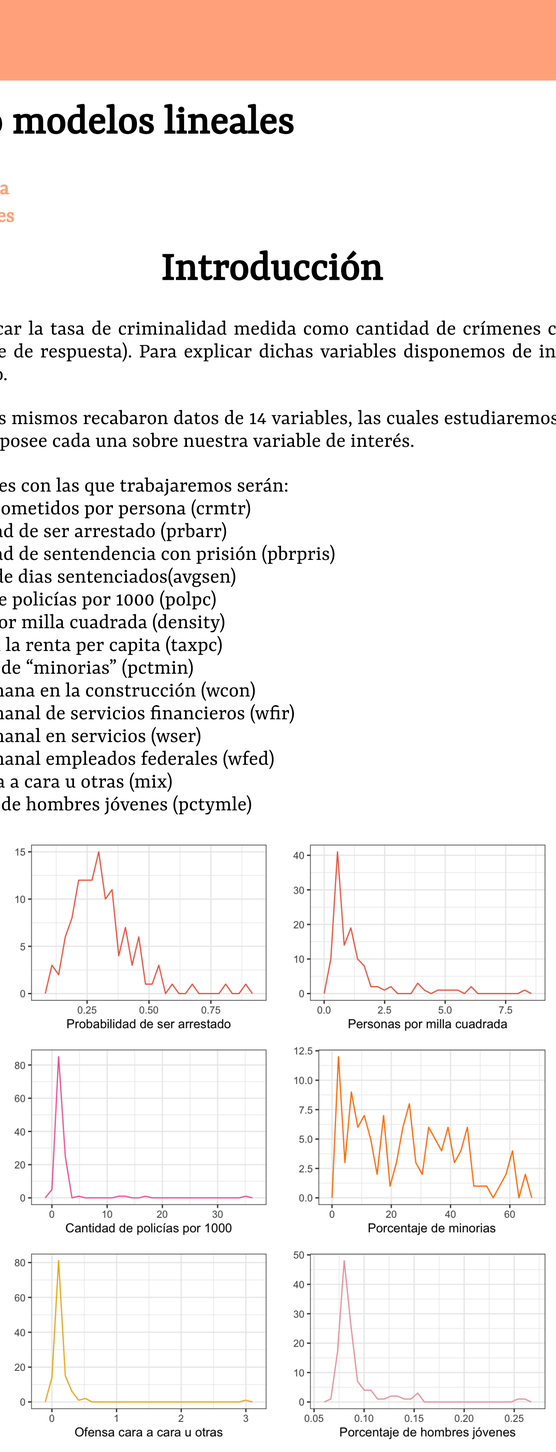


Figure 1: Visualización de variables

Para saber si las variables están correlacionadas con la variable de respuesta, se hizo el siguiente gráfico.

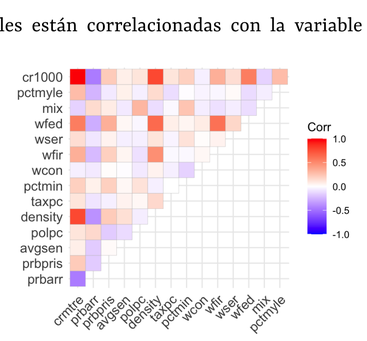
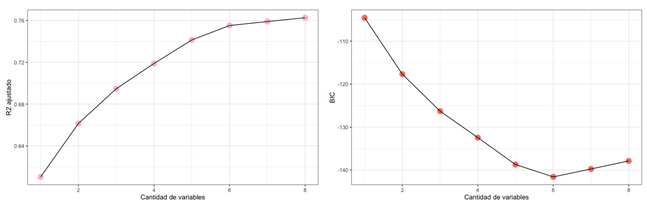


Figure 2: Gráficos de correlaciones

Se puede observar que las variables más correlacionadas con *crmtre* son *cr100*, *density*, *wfed* y *prbarr*.

cr100 tiene una correlación de 1 con *crmtre* ya que es la mismo multiplicada por 100, por lo que la quitamos de nuestros datos.

Necesitamos saber que variables son las que aportan a explicar la tasa de criminalidad. Para ello utilizaremos el R_j^2 y el BIC.



Dichos gráficos nos muestran los modelos posibles de los cuales elegiremos el que mejor se ajusta según el R_j^2 y el que mejor se ajusta según el BIC.

Una primera observación mediante el R_j^2, se podría decir que a partir de 5 variables el modelo alcanzaría a explicar con las mismas un 74% de la variabilidad de *crmtre* y va aumentando conforme aumentan las variables.

Ahora tenemos que ver si es relevante agregar una variable más para obtener un aumento en la explicación del modelo o no ya que a veces es mejor sacrificar un poco del porcentaje de la variabilidad de la Y que es explicada por las X, a cambio de tener menos variables en el modelo.

Entonces utilizamos la herramienta BIC que, a diferencia del R_j^2, queremos que este tenga el menor valor posible y así poder comparar qué modelo es mejor.

Se puede observar que el modelo que tiene menor BIC es el modelo de 6 variables, pero no es el modelo que tiene mayor R_j^2, pero como se mencionó anteriormente, estamos dispuestas a sacrificar un poco de la variabilidad de Y explicada por las X y así tener un modelo con menos variables. Por lo tanto, el modelo que mejor explica a la variable de interés se compondrá de 6 variables.
Las cuales son "prbarr", "polpc", "density", "pctmin", "mix" y "pctmyle".

Diagnóstico

Una vez obtenido nuestro modelo, podemos hacer un diagnóstico del mismo para ver si se cumplen los supuestos ya que de no cumplirse, la inferencia que se haga no será válida.

Supuestos

- Linealidad: Linealidad entre Y y las variables contenidas en X.
- Homocedasticidad: La varianza de los errores es constante.
- Normalidad: errores se distribuyen normal

También se deben de chequear:

- Multicolinealidad: Es el hecho de que una o más columnas de la matriz X puede obtenerse mediante una combinación lineal (CL) de otras columnas de X.
- Valores atípicos y/o influyentes: Observación distinta al resto, donde el o los residuos escapan a la norma. Es una observación en un conjunto de datos que, cuando se elimina, cambia drásticamente las estimaciones de coeficientes de un modelo de regresión.

Multicolinealidad

Para llevar a cabo este análisis, comenzaremos con la multicolinealidad del modelo para ello implementaremos el factor de inflación de varianza (VIF), este procedimiento se basa en la idea de que si x_j es casi combinación lineal de las demás, el R^2 de una regresión donde xj es la variable de respuesta debería ser alto.

Table 1: VIF

prbarr	polpc	density	pctmin	mix	pctmyle
1.247	1.127	1.175	1.069	1.18	1.07

Podemos notar que ninguna de nuestras variables es combinación lineal de las otras, ya que los factores de inflación están por debajo de 5 por lo tanto descartamos la presencia de multicolinealidad.

Linealidad

Una vez analizado el diagnóstico anterior nos enfocaremos en la linealidad del modelo en su conjunto y además la misma por cada variable. Para dicho análisis se deberá de ver el comportamiento que tienen los predichos y los residuos del modelo por lo que observaremos el mismo a través de un gráfico de puntos.

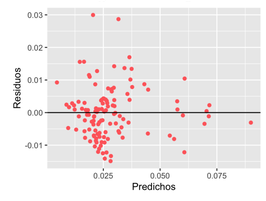


Figure 3: Gráficos de residuos parciales

Si bien los puntos están más concentrados en la zona izquierda del gráfico no presentan un patrón propiamente dicho por lo que podríamos decir que nuestro modelo es lineal. Por lo que analizaremos la linealidad en las respectivas variables explicativas.

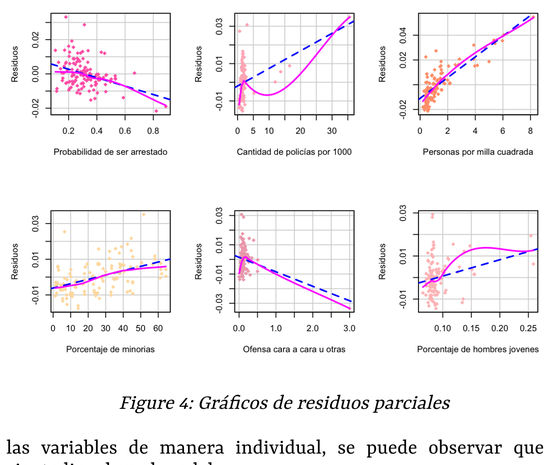


Figure 4: Gráficos de residuos parciales

Al tratar las variables de manera individual, se puede observar que no hay un comportamiento lineal en el modelo.

Homocedasticidad

Para chequear si el modelo cumple o no con el supuesto de homocedasticidad, se utiliza nuevamente el gráfico del comportamiento que tienen los residuos estudentizados externamente con respecto a los predichos, y donde si no se observa ningún patrón en él, se puede decir que hay homocedasticidad.

Si bien no es claro el comportamiento de los puntos en el gráfico, podríamos decir que no hay un patrón como tal. Dado esto lo analizaremos mediante el test de hipótesis de Breusch-Pagan:

$H_0) Varianza\ constante$

$H_1) Varianza\ no\ constante$

Nos interesa NO rechazar la hipótesis nula.

Table 2: Breusch Pagan

statistic	p.value	parameter	method	alternative
6.673489	0.3521005	6	Koenker (studentised)	greater

Mediante la prueba de hipótesis podemos ver que el p-valor es mayor que el nivel de significación a un 1% por lo que no rechazamos la hipótesis nula, esto implica que, si se cumple y por lo tanto los errores son homocedásticos, lo cual significa que las varianzas de los residuos son todas iguales.

Supuesto de normalidad

Otro supuesto que nos interesaba ver era la normalidad.

$H_0) W \sim Normal$

$H_1) no\ H_0$

Con el cual nos interesa no rechazar la H_0

Vamos a utilizar los test Shapiro Wilk el cual se basa en la comparación de los cuantiles empíricos y teóricos bajo el supuesto de normalidad.

Table 3: Shapiro test

0.002

Rechazamos la hipótesis nula y por ende los residuos del modelo no presentan normalidad.

Valores atípicos o influyentes

Por último nos queda ver las observaciones atípicas y su influencia.

H_0) La observación i NO es atípica

H_1) La observación i SI es atípica

Table 4: Valores atípicos e influyentes

	pvalor	Bonferroni
5	0.000	0.015
41	0.000	0.033
109	0.006	0.698

Obtuvimos 2 observaciones atípicas dado que sus respectivos p-valores hacen que rechacemos la hipótesis nula con un nivel de significación del 1% y una observación influyente.

Incumplimiento de los supuestos

Ya que en nuestro modelo no se cumple la linealidad comenzaremos con quitar las observaciones atípicas, pero como se observa en la tabla anterior, sobre todo la observación 109 tiene una influencia grande sobre el modelo y esto puede afectar el resto de los supuestos, por lo que luego de quitarla se deberán de volver a analizar los mismos.

Quitando los valores atípicos e influyentes se nos cumplen los supuestos de multicolinealidad, normalidad y homocedasticidad, pero no la linealidad, por ende, para obtener dicho cumplimiento optamos por aplicar logaritmo a *polpc*, *mix* y *pctmyle*, ya que puede resolver problemas de linealidad y son más faciles de interpretar.

Quedando el modelo:

$$Crmtre = \beta_0 + \beta_1 Prbarr + \beta_2 log(Polpc) + \beta_3 Density + \beta_4 Pctmin + \beta_5 log(Mix) + \beta_6 log(Pctmyle)$$

Luego de transformar las variables vemos que ahora la linealidad de las mismas es más clara, por lo que podríamos decir que dicho supuesto se cumple.

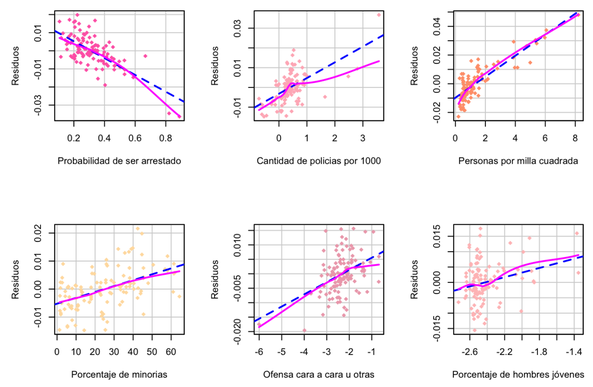


Figure 5: Gráficos de residuos parciales

Pero dado que para lograr la linealidad del modelo tuvimos que transformar ciertas variables deberemos de chequear si los supuestos se cumplen. Nuevamente comenzamos con la multicolinealidad, seguido de la homocedasticidad y por último la normalidad.

Como podemos ver al realizar el VIF los valores obtenidos son menores a 5 por lo tanto nuestro modelo no presenta multicolinealidad.

Al chequear la homocedasticidad obtenemos un p-valor mayor a nuestro nivel de significación del 1% utilizado durante todo el trabajo, de esta manera no rechazamos nuestra hipótesis nula y por lo tanto el modelo es homocedastico.

Por último, al realizar el *Shapiro Wilk* obtenemos un p-valor mayor a nuestro nivel de significación mencionado anteriormente por lo que tampoco rechazamos nuestra hipótesis nula y por ende nuestro modelo es normal.

Conclusión

Luego de llegar a que todos los supuestos se cumplen y por ende la inferencia que se realice sobre el modelo tendrá validez podemos proseguir a interpretar los estimadores de nuestro modelo de forma correcta. Para ello visualizaremos el summary del mismo.

	crmtre		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.052	0.037 – 0.068	**<0.001**
prbarr	-0.046	-0.060 – -0.033	**<0.001**
polpc [log]	0.008	0.006 – 0.010	**<0.001**
density	0.007	0.006 – 0.008	**<0.001**
pctmin	0.000	0.000 – 0.000	**<0.001**
mix [log]	0.004	0.002 – 0.006	**<0.001**
pctmyle [log]	0.008	0.002 – 0.013	**0.009**
Observations	117		
R^2 / R^2 adjusted	0.852 / 0.844		

Como se puede observar el modelo es globalmente significativo dado que tiene un p-valor menor al 1% por lo tanto alguna de nuestras variables seleccionadas sirve para explicar el mismo. Luego de observar esto e ir a detalle en cada una de estas vemos que todas ellas tienen un p-valor menor al 1% por lo tanto podemos afirmar que todas ayudan a explicar a la variable de interés.

Podemos ver que si bien tenemos menos variables que al inicio del análisis el R^2 es de 85%, significando que nuestro modelo es capaz de captar el 85% de variabilidad de "crmtre" que puede explicarse mediante las variables explicativas.

Algunas interpretaciones del modelo:

Entre dos observaciones que difieren solamente en la probabilidad de arresto, esperamos que la tasa de criminalidad sea menor en 4,64%.

Por otro lado, el aumento porcentual de la cantidad de policías cada 1000 personas hacen que aumente el porcentaje de crímenes cometidos por persona en un 0,79%, dejando las demás variables constantes.