Proyecto Final Regresión Avanzada

Análisis Jerárquico de asesinatos en EUA

Ana Luisa Masetto Herrera
Arantza Ivonne Pineda Sandoval
Ixchel Meza Chávez
Saúl Caballero Ramírez

Índice

1. Introducción

La historia de Estados Unidos de Norteamérica (EUA) en temas de violencia y el crimen tiene raíces mucho más antiguas que los años recientes plagados de titulares en periódicos y noticieros anunciando tiroteos, asesinatos masivos y violencia racial.

La base de datos "*Communities and Crime Unnormalized Data Set*" combina información de datos censales de 1990 publicados por el *US Census* con los reportes de crimen en al año 1995 publicados por el *FBI*. En este año la tasa de homicidios disminuyo en algunas de las ciudades más violentas de los Estados Unidos de Norteamérica. El 13 de agosto de 1995 el periódico *The New York Times* publicó la noticia titulada *Many Cities in U.S. show sharp drop in homicide rate* [1], en esta nota se adjudica que la disminución en las tasas fue ocasionada por tácticas políticas más agresivas, incremento del número de criminales en prisión y patrones cambiantes en el uso de drogas.

Este proyecto busca explicar la tasa de asesinatos en los diferentes condados a través de modelos de regresión que consideren efectos por estado, división censal y variables explicativas. Las divisiones censales son las separaciones regionales indicadas por la oficina del censo de los Estados Unidos de Norteamérica, en la cual se identifican 9 regiones diferentes (New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain y Pacific). En particular este trabajo:

- Realiza un análisis exploratorio univariado y bivariado entre la variable de respuesta y las 9 covariables identificadas.
- Explora distintos modelos de regresión para entender las relaciones subyacentes en los datos.
- Interpreta los resultados obtenidos y selecciona el modelo que mejor explique la tasa de crímenes.

[1] https://www.nytimes.com/1995/08/13/us/many-cities-in-us-show-sharp-drop-in-homicide-rate.html

2. Descripción de la base de datos

La base de datos que se utiliza es *Communities and Crime Unnormalized Data Set* la cual se encuentra en la página de UCI [2]. Esta base contiene muchas variables sociodemográficas por condado, sin embargo, muchas de estas variables tienen valores faltantes por lo que las variables analizadas en este trabajo son:

- **Variable respuesta:**
  - **Murders**: Número de asesinatos.
- **Variables explicativas**
  - **PctBlack**: Porcentaje de la población que es Afroamericana.
  - **PctWhite**: Porcentaje de la población que es Caucásica.
  - **PctHisp**: Porcentaje de la población que es Hispana.
  - **PctPoverty**: Porcentaje de la población por debajo de nivel de pobreza.
  - **Pct12-17w2Par**: Porcentaje de niños entre 12 y 17 años que viven con ambos padres.
  - **PctNotSpeakEng**: Porcentaje de la población que no habla bien inglés.
  - **PctBornStateResid**: Porcentaje de de la población que reside en el mismo estado donde nació.
  - **GraduatespvtNotHSgrad**: Porcentaje de la población que tiene 25 o más y no se graduó de preparatoria.
  - **ForcepctWorkMom.18**: Porcentaje de madres que trabajan con hijos menores a 18 años.
  - **YearspctFgnImmg.10**: Porcentaje de la población que inmigró en los últimos 10 años.

Adicionalmente se cuenta al condado y estado al que pertenece cada observación. A continuación se muestra los estados presentes en la base de datos y el número de condados de los que se tienen información:

Cuadro 1: Estados de EUA

Estado	Número de condados	Estado	Número de condados
California	279	Kentucky	26
New Jersey	211	Rhode Island	26
Texas	162	Arkansas	25
Massachusetts	123	Colorado	25
Ohio	111	Utah	24
Michigan	108	Louisiana	22
Pennsylvania	101	New Hampshire	21
Florida	90	Arizona	20
Connecticut	71	Iowa	20
Minnesota	66	Mississippi	20
Wisconsin	60	Maine	17
Indiana	48	West Virginia	14

[2] https://archive.ics.uci.edu/ml/datasets.html

Estado	Número de condados	Estado	Número de condados
North Carolina	46	Maryland	12
New York	46	New Mexico	10
Alabama	43	South Dakota	9
Missouri	42	North Dakota	8
Illinois	40	Idaho	7
Washington	40	Wyoming	7
Georgia	37	Nevada	5
Oklahoma	36	Vermont	4
Tennessee	35	Alaska	3
Virginia	33	District of Columbia	1
Oregon	31	Delaware	1
South Carolina	28	Kansas	1

A partir de aquí se puede notar que un modelo jerarquico puede combinar la información de estados con más condados, por ejemplo, California con estados que tienen menor cantidad de condados como Columbia, Delaware y Kansas.

Tambíen se muestra el número de estados por división:

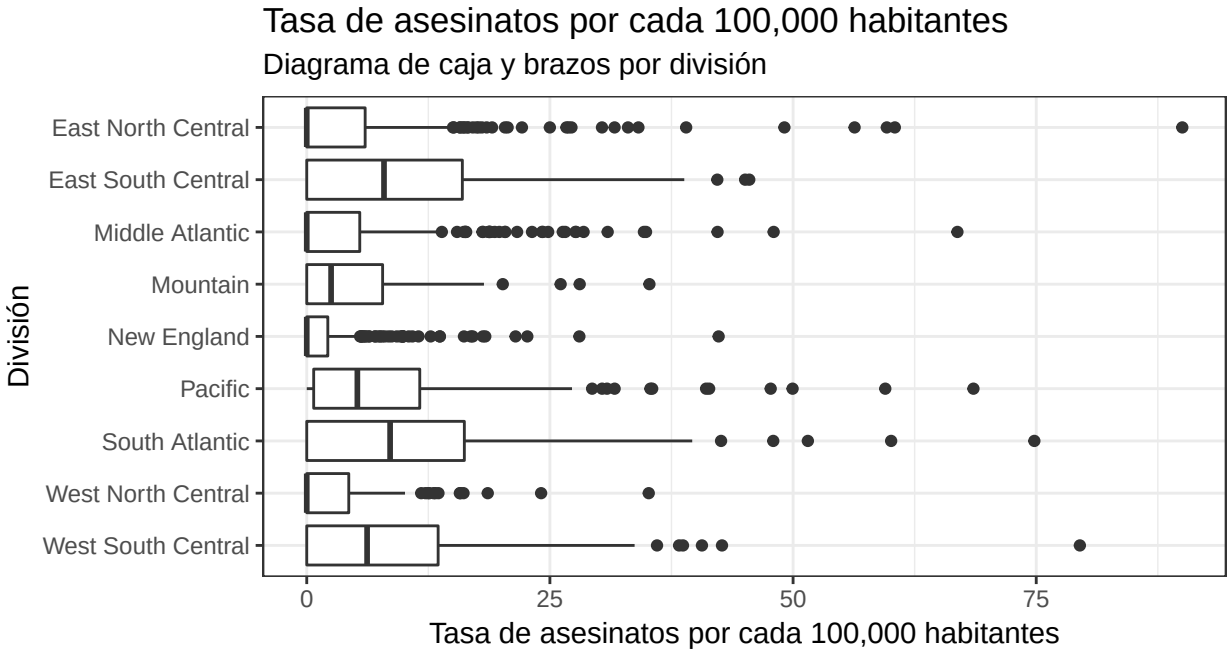
División	Número de estados
South Atlantic	9
Mountain	8
West North Central	7
New England	6
East North Central	5
Pacific	5
East South Central	4
West South Central	4
Middle Atlantic	3

Aquí también se observa que las divisiones con más estados ayudarán en la estimación de los parámetros por división a divisiones con menor número de estados.

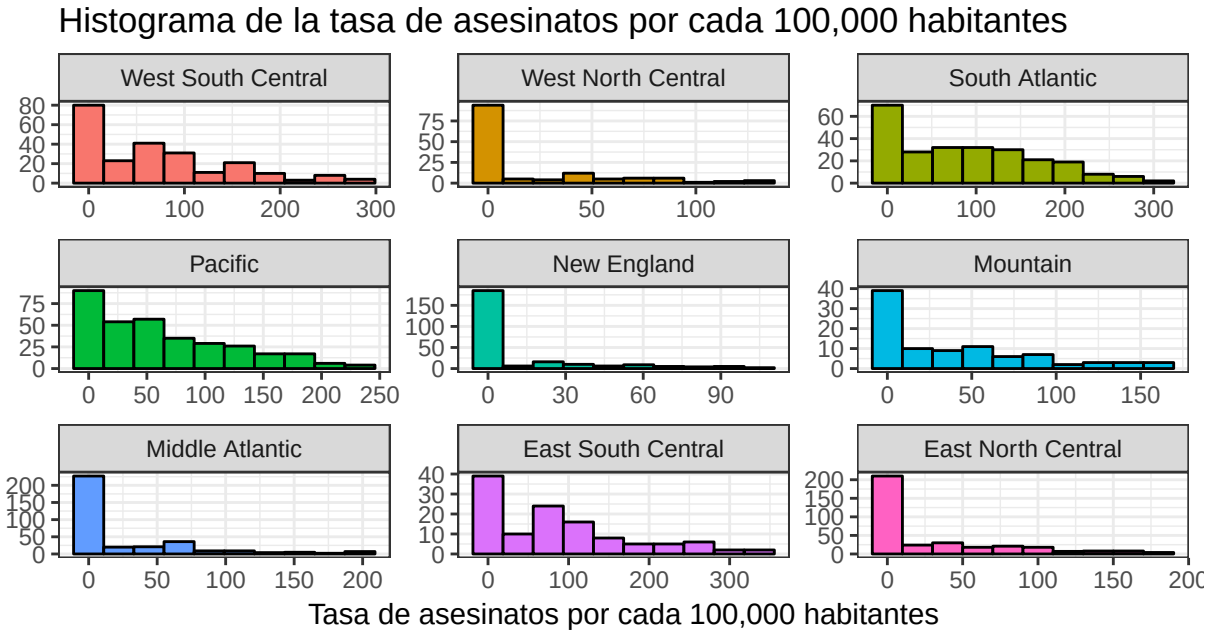
3. Análisis exploratorio de los datos

3.1. Análisis Univariado

Para entender los asesinatos en EUA se muestran las siguientes gráficas que son la tasa de asesinatos dividido entre el número de población por condado. Para facilitar la interpretación se muestran agrupados por división.



A partir de esta gráfica se puede notar que existen condados con tasas altas de asesinatos, para entender un poco más el contexto se realiza el siguiente histograma que solo conserva observaciones menores al cuantil 95:

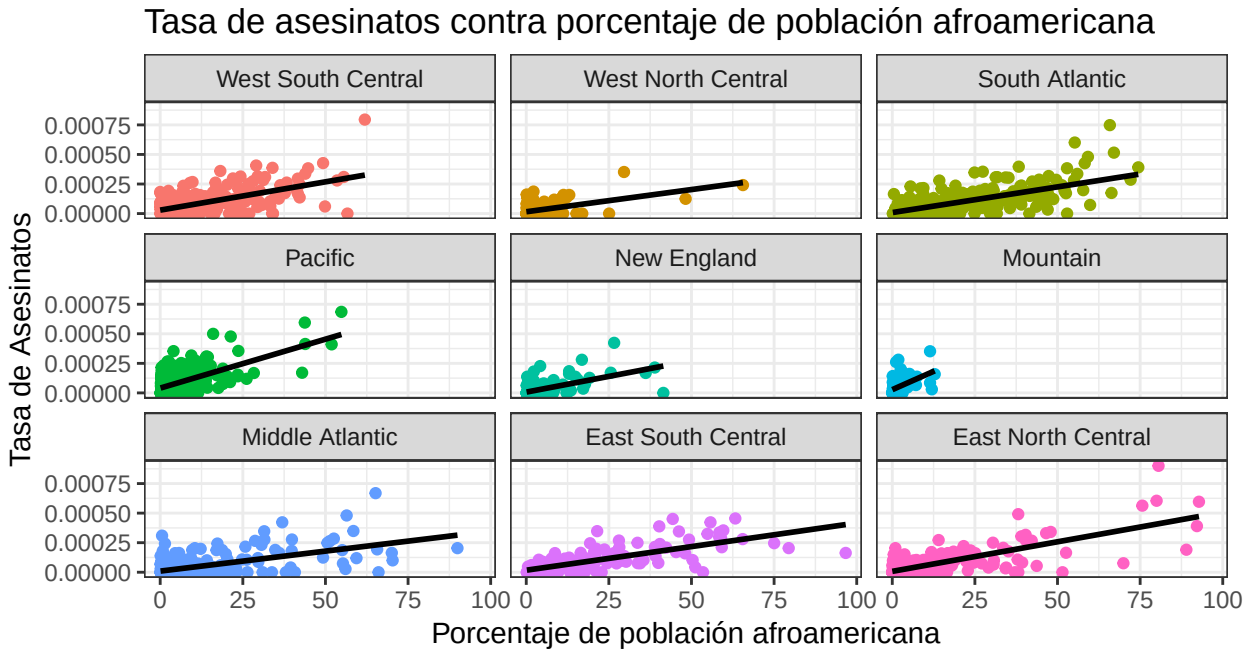


Con esta gráfica se nota que independientemente el estado existen muchos condados con tasa de asesinatos cercana al 0. Esto es un punto importante a notar a la hora del modelado. Además se puede observar que las divisiones con más condados que tienen pocos asesinatos son West North Central, New England, Middle Atlantic y East North Central, sin embargo, siguen teniendo algunos estados con cantidades altas de asesinatos.

3.2. Análisis Bivariado

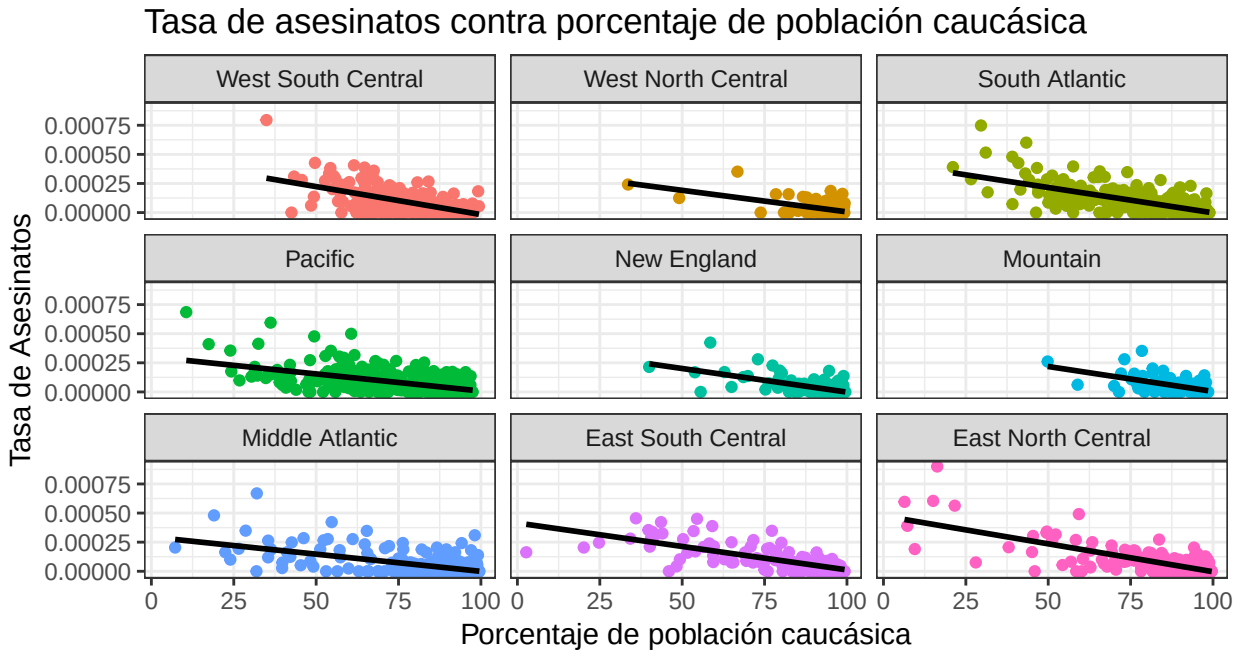
El siguiente paso es análizar las posibles relaciones entre la tasa de asesinatos y las variables sociodemográficas. Como ayuda visual se ajustaron modelos simples de la relación entre tasa de asesinatos y la variable gráficada solo para poder entender si la variable tiene o no tiene relación con el fenómeno de asesinatos. Hay que tener mucha precaución con este ajuste pues al no considerar otras variables o efectos por estado puede que la relación observada sea causada por otra variable con la que se tiene correlación.

Primero, se observa la relación entre la tasa de asesinatos y el porcentaje de población afroamericana en el condado:



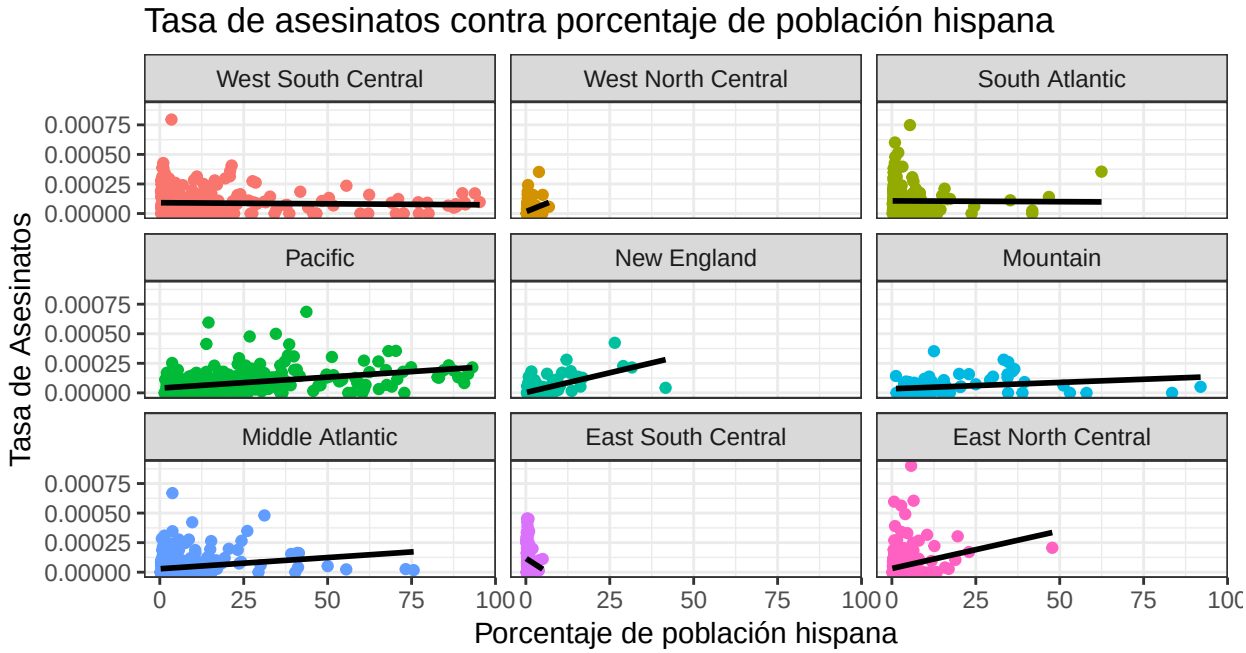
Lo primero a notar es que todos los condados dentro de las divisiones de Mountain, New England, Pacific y West North Central tienen un porcentaje de población afroamericana menor a un 60 %, lo cual llevaría a pensar que dentro de estas divisiones no existe mucha diversidad racial.

Se observa que todos las divisiones muestran que conforme el porcentaje de población afroamericana incrementa, el porcentaje de asesinatos también aumenta. Además se puede observar que dado un modelo univariado la intensidad con la que afecta esta variable en las distintas divisiones parecen ser distintos.



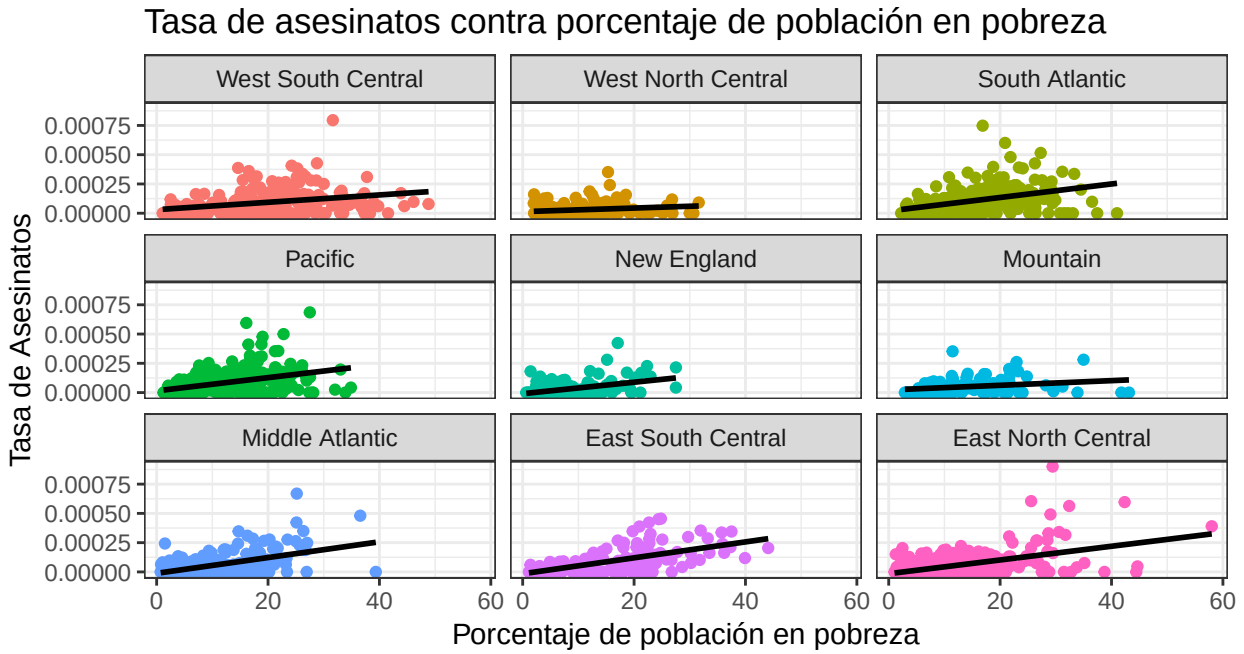
Con esta gráfica se confirma la hipótesis anterior, que en las divisiones de Mountain, New England y West North Central no hay mucha diversidad racial dentro de los condados de estas divisiones. Esto puede ser un potencial problema para la estimación de los modelos con ambas variables pues estaríamos en el caso de multicolinealidad y potencialmente los efectos de las variables se podría confundir.

Se observa que para todas las divisiones conforme aumenta el porcentaje de personas caucasicas disminuye la tasa de asesinatos.

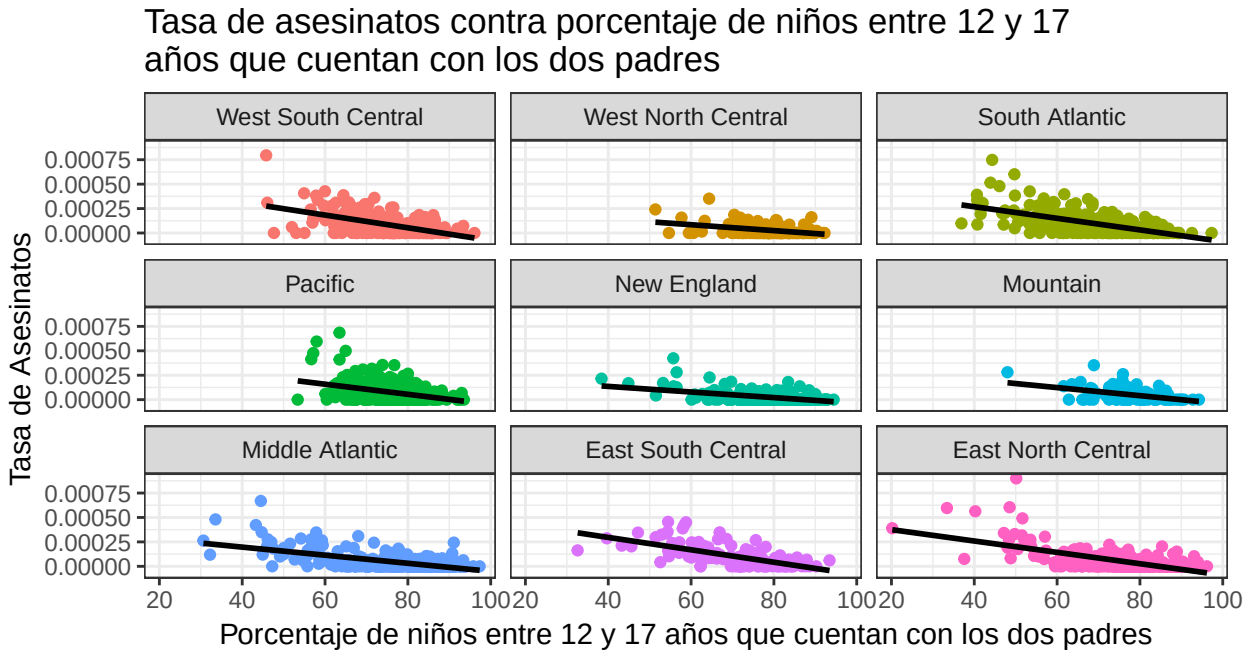


Con el porcentaje de población hispana se muestra que existen divisiones que no tienen

muchas personas de origen hispano dentro de sus comunidades. Además se observa que el efecto sobre la tasa de asesinatos no es claro, por ejemplo, en East North Central se observa una pendiente positiva, en Mountain una pendiente casi nula y en East South Central negativa, en esta última se podría estar observando este comportamiento debido a la poca cantidad de personas de origen hispano en los condados.

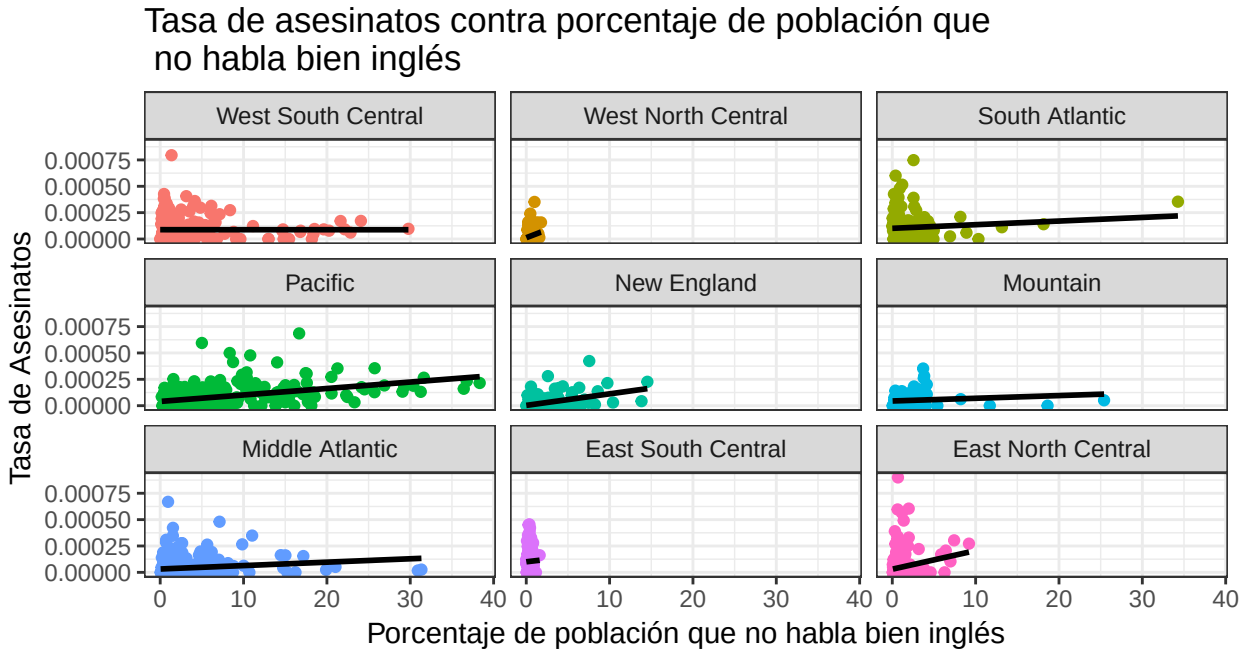


Se puede observar una posible relación positiva entre el porcentaje de población en pobreza y la tasa de asesinatos. New England parece tener los índices de pobreza más bajos de todas las divisiones (menores al 50 %).



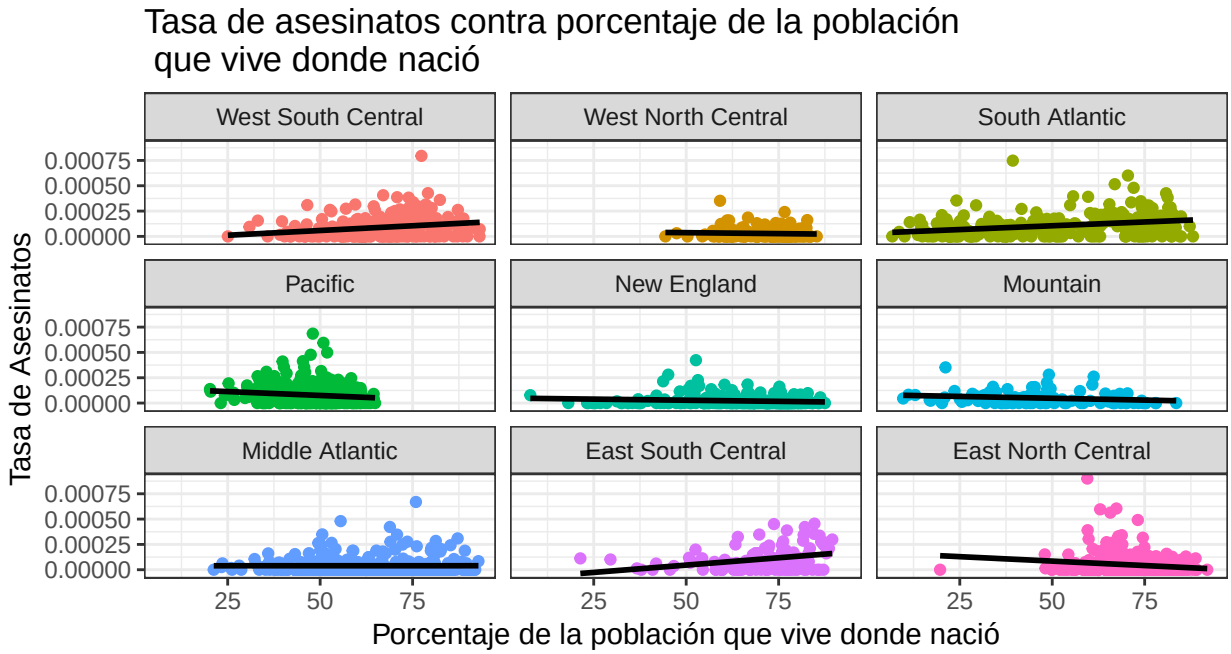
De manera general, todas las divisiones parecen mostrar un comportamiento similar que se puede dividir en dos aspectos relevantes:

- Los porcentajes de niños que cuentan con ambos padres están ubicados en los valores altos, lo que indica una población mayoritariamente está conformada por familias unidas donde los hijos viven con sus dos padres.
- La relación que muestra con la tasa de asesinatos parece ser negativa.



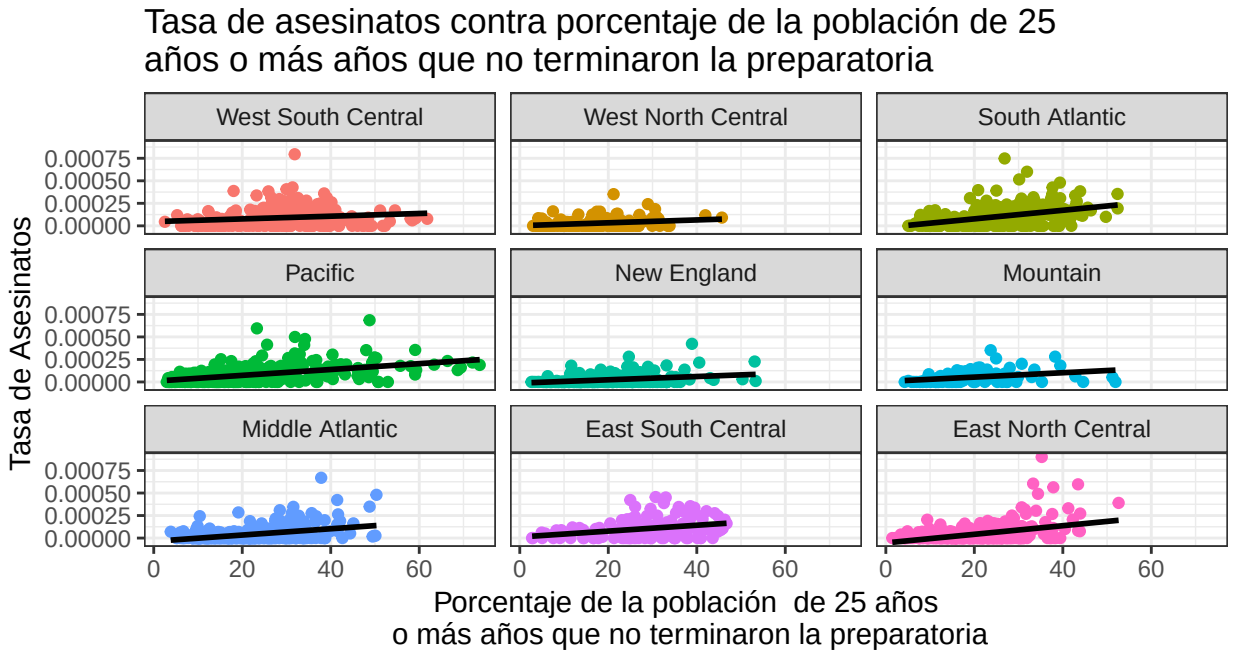
Se observan tres tipos de comportamientos distintos:

- En las divisiones de East South Central y West North Central tienen porcentajes cercanos a cero, es decir, casi todos hablan bien inglés; sin embargo, las tasas de asesinatos son muy variantes e incluso altas recorriendo desde el 0 % hasta alcanzar el .05 %.
- Las divisiones de East North Central y New England: sus observaciones puntuales muestran porcentajes menores al 15 %, es decir, que sí existe una proporción que no se puede ignorar de personas que hablan mal el inglés. Las observaciones se encuentran variando mucho respecto a la tasa de asesinatos. Se trata por lo tanto de un comportamiento o patrones no concluyentes respecto a la relación entre ambas variables.
- Las divisiones de Middle Atlantic, Mountain, Pacific, South Atlantic y West South Central muestran porcentajes más altos de personas que hablan mal el inglés (alcanzando hasta 40 %). Mientras que en casi todas las divisiones no parece haber una relación directa con la tasa de asesinatos, las división de Pacific parecería sí tener una mayor tasa de asesinatos a mayor porcentaje de personas que no hablan bien el inglés.

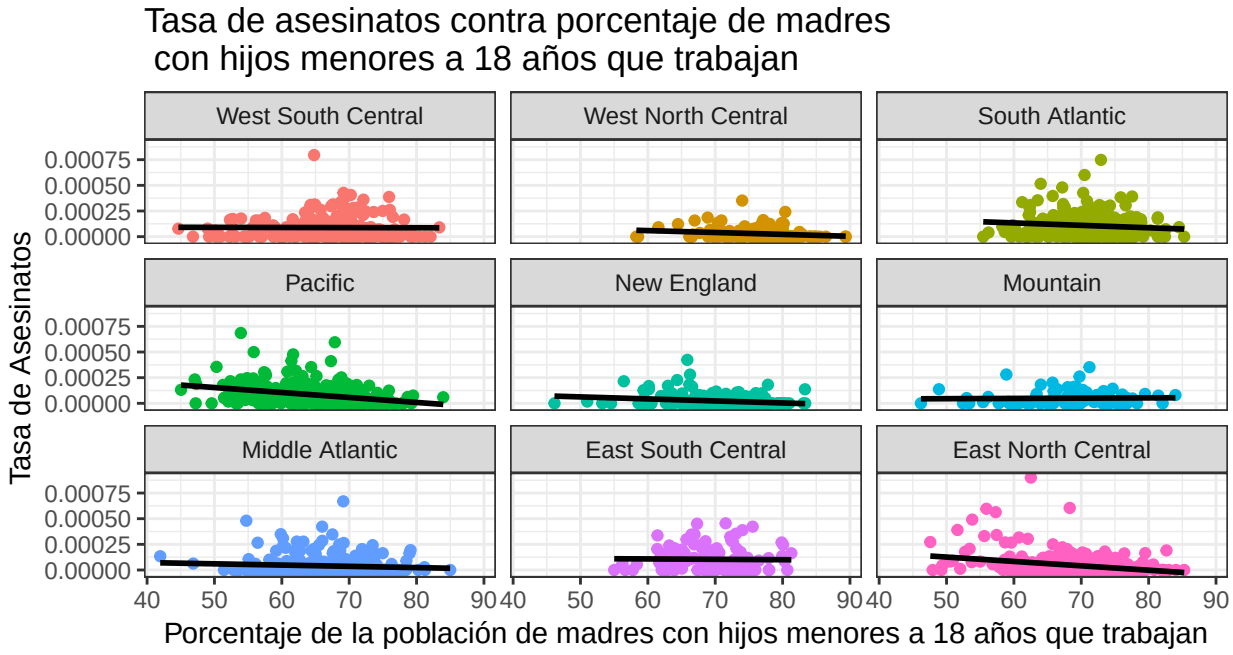


- En las divisiones de East North Central, East South Central y West North Central existen muchas personas oriundas de su división. No existe una relación evidente con la tasa de asesinatos ya que las tasas altas de asesinatos se encuentran dispersas sin importar si existen altos o bajos porcentajes de personas oriundas. Para la división de East South Central parece haber una ligera tendencia creciente.
- Las divisiones de Middle Atlantic, Mountain y New England tienen comportamientos muy planos ya que los porcentajes de oriundos cubren un espectro muy amplio (desde el 0 % hasta poco más del 75 %) y las tasas de asesinatos para éstos varían mucho sin hacer distinción entre porcentajes altos o bajos de personas oriundas.
- La división de Pacific muestra un comportamiento particular: para población oriunda de alrededor del 50 %, las tasas de asesinatos son muy altas, pero una vez que las tasas de oriundos se acercan a los extremos, las tasas de asesinatos parecen disminuir. Los datos parecen estar acotados a tasas de oriundos menores al 75 % y centrados en tasas del 50 %.
- Las divisiones de South Atlantic y West South Central: aunque estas divisiones presentan un alto espectro de tasas de oriundos (desde el 0 % hasta más del 75 %), las aproximaciones lineales muestran una tendencia creciente con la tasa de asesinatos que se comprobará posteriormente con los resultados del modelo.

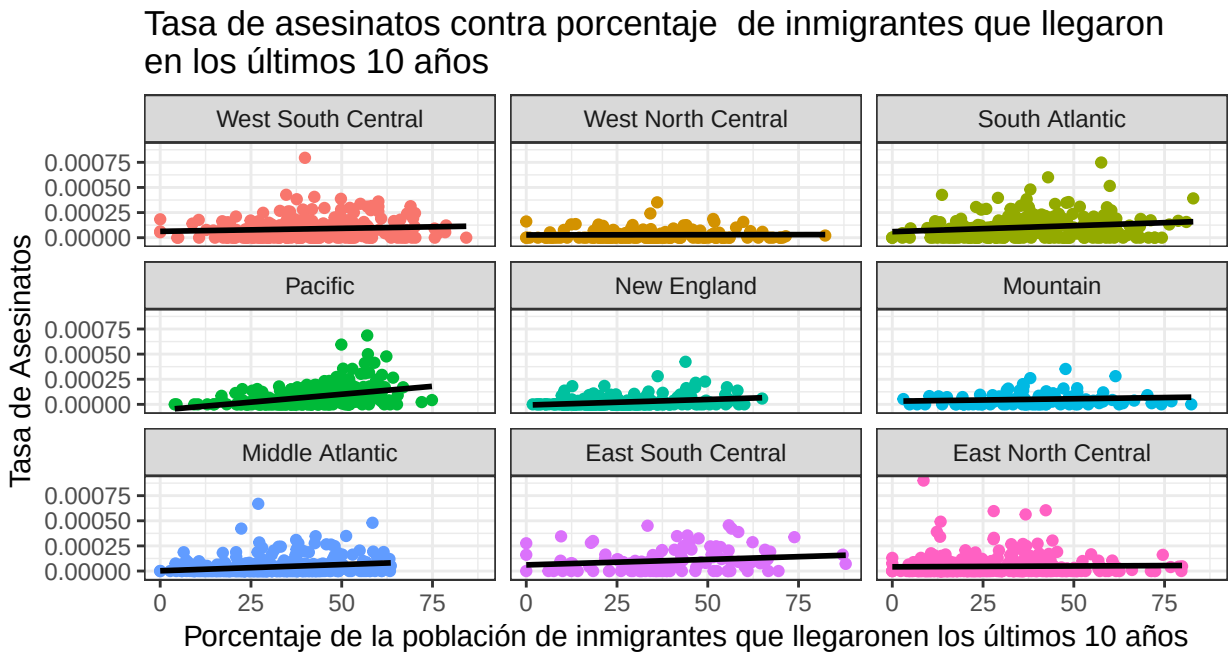
De este análisis parece no haber relación directa con un factor de inmigración afecte la tasa de asesinatos, ya que sin importar si las personas son o no originarias de su residencia actual, las tasas de asesinatos son muy dispersas.



En todas las divisiones la tasa de asesinatos es baja en condados con bajo porcentaje de población sin preparatoria y en general en todas las divisiones la tasa de asesinatos tiende a aumentar conforme aumenta el porcentaje de gente sin preparatoria terminada. Este comportamiento es más claro en East North Central y Middle Atlantic, mientras que en las demás divisiones, pese a que también hay una tendencia a aumentar la tasa de asesinatos, los condados con más mortalidad no son los que tienen una proporción más alta de la población sin certificado de preparatoria.



El porcentaje de madres con hijos menores a 18 años parece no tener un efecto claro a simple vista.



Se identifican los siguientes patrones:

- Las divisiones de Mountain, New England y West North Central: tienen comportamientos similares. Los porcentajes de inmigrantes están distribuidos ampliamente entre 0 % y 100 % (solo New England parece tener tasas de inmigrantes menores al 75 %) para los cuales existen mayoritariamente tasas bajas menores al 0.05 %. No se identifica una relación directa entre ambas variables.
- Las divisiones East North Central, South Atlantic y West South Cental: muestran relaciones planas no concluyentes respecto a la relación de la covariable con las tasas de asesinatos. Sin embargo, para estas divisiones, existen algunas o varias observaciones de condados que superan el 0.05 % en tasas de asesinatos.
- Las divisiones de East South Central, Middle Atlantic y Pacific: presentan aproximaciones lineales con tendencia creciente; conforme aumentan los porcentajes de inmigrantes en los últimos 10 años, aumentan las tasas de asesinatos. Es interesante el caso de la división de Pacific, ya que de las tres anteriores es aquella con la relación positiva más marcada.

4. Modelos

Para plantear los modelos sea y_i el número de asesinatos por condado y n_i la población total por condado donde $i \in \{1, \ldots, 2215\}$. Para representar los distintos estados se usará el subíndice $s \in \{1, \ldots, 48\}$ y para representar las divisiones se utiliza el subíndice $d \in \{1, \ldots, 9\}$. X_i representa un conjunto de covariables por condado y se asume que existen k covariables.

4.1. Modelo de intercepto variante por divisiones sin covariables

El objetivo del primer modelo fue tener un modelo sencillo con interceptos variantes por división y sin covaribles para evaluar el desempeño de distintas verosimilitudes y ligas. En particular, se prueban dos verosimilitudes distintas: *Poisson* y *Binomial*. Para los modelos con verosimilitud *Binomial* se prueban las ligas: *logística, probit, log-log* y *log-log complementaria*(clog-log), y para la distribución *Poisson* solo se utiliza la liga logarítmica. Los modelos probados son los siguientes:

Modelo Poisson

$$\begin{aligned}
y_i &\sim Po(n_i\lambda_i) && i \in \{1, \ldots, 2215\} \\
\lambda_i &= e^{\theta_d} && i \in \{1, \ldots, 2215\},\ d \in \{1, \ldots, 9\} \\
\theta_d &\sim N(\phi, \sigma_\phi^2) && d \in \{1, \ldots, 9\} \\
\phi &\sim N(0, 10) \\
\sigma_\phi &\sim \Gamma(0.001, 0.001)
\end{aligned}$$

Modelo Binomial

$$\begin{aligned}
y_i &\sim Bin(n_i, \pi_i) && i \in \{1, \ldots, 2215\} \\
\pi_i &= f(\theta_d) && i \in \{1, \ldots, 2215\},\ d \in \{1, \ldots, 9\} \\
\theta_d &\sim N(\phi, \sigma_\phi^2) && d \in \{1, \ldots, 9\} \\
\phi &\sim N(0, 10) \\
\sigma_\phi &\sim \Gamma(0.001, 0.001)
\end{aligned}$$

$$f(x) = \begin{cases} logit^{-1}(x) & \text{Liga Logistica} \\ \Phi(x) & \text{Liga Probit} \\ e^{-e^x} & \text{Liga log-log} \\ 1 - e^{-e^x} & \text{Liga log-log complementaria} \end{cases}$$

Las distribuciones iniciales para los hiperparámetros ϕ y σ_ϕ son distribuciones vagas, es decir, que no contienen mucha información acerca de los hiperparámetros reales.

Para poder comparar los distintos modelos se utilizó el criterio de información de Watanabe-Akaike (WAIC), el cual se calcula de la siguiente forma:

$$WAIC = -2(log(f(y|\theta)) - P)$$

donde el primer componente es el logaritmo de la predictiva posterior y P es una estimación del número de parámetros efectivos en el modelo (Gelman, Hwang, and Vehtari 2014). Como cualquier critério de información se busca el valor más bajo del WAIC.

El WAIC calculado para cada modelo es el siguiente:

Verosimilitud	Liga	WAIC
Binomial	Probit	19920
Binomial	Loglog	19910
Binomial	Cloglog	19902
Binomial	Logit	19892
Poisson	Exponencial	19891

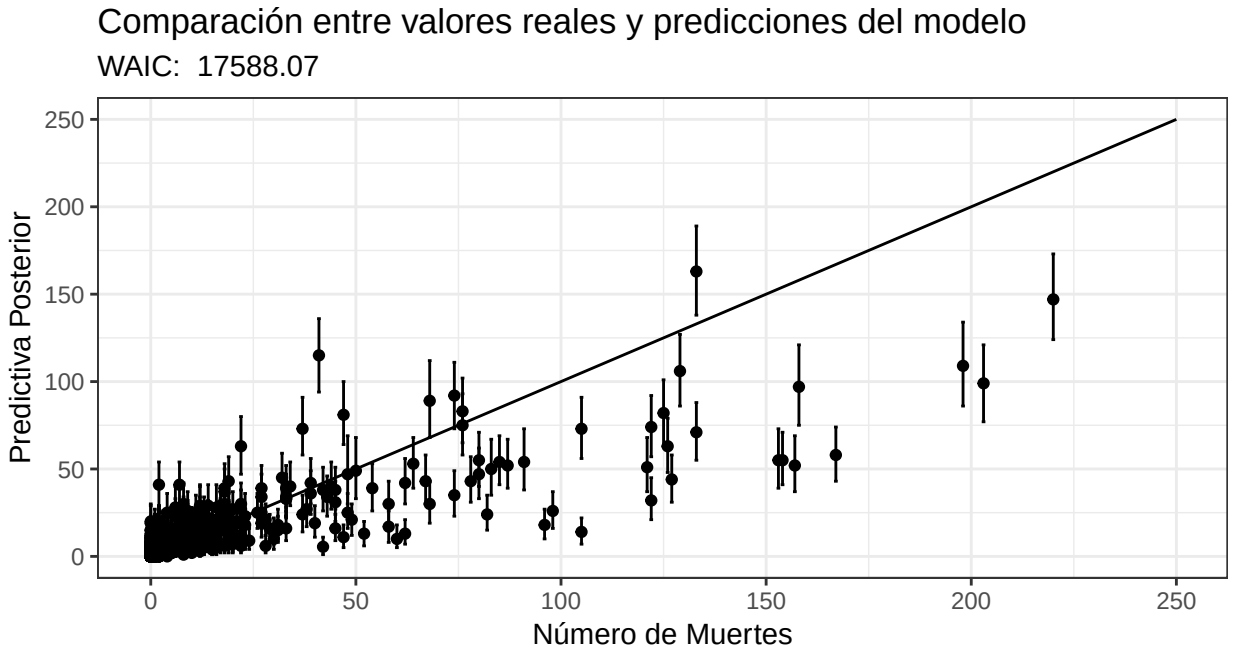
dado el criterio del WAIC, el mejor modelo es el *Poisson* con liga exponencial. Es por esta razón que para los siguientes modelos se utilizará la verosimilitud *Poisson* con liga *exponencial*.

4.2. Modelo de intercepto variante por estado y por división sin covariables

En este modelo se le agrega una jerarquía al modelo, es decir, ahora los interceptos varían por estado y estos a su vez dependen de los hiperparámetros de las distintas divisiones a las que pertenecen y finalmente al hiperparámetro que representa a EUA. El modelo se define de la siguiente forma:

$$
\begin{aligned}
y_i &\sim Po(n_i\lambda_i) && i \in \{1,\dots,2215\} \\
\lambda_i &= exp(\beta_s) && i \in \{1,\dots,2215\},\ s \in \{1,\dots,48\} \\
\beta_s &\sim N(\theta_d, \sigma^2_{\beta d}) && s \in \{1,\dots,48\},\ d \in \{1,\dots,9\} \\
\theta_d &\sim N(\phi, \sigma^2_{\phi}) && d \in \{1,\dots,9\} \\
\phi &\sim N(0, 10) \\
\sigma_\phi &\sim \Gamma(0.001, 0.001)
\end{aligned}
$$

A continuación se muestra el desempeño del modelo:



La gráfica muestra la recta de 45ř la cual representa un modelo predictivo perfecto, y se puede apreciar que los puntos están muy alejados de esta curva, por lo que no es un buen modelo. Por otra parte el WAIC disminuyó a comparación de los primeros modelos, lo cual nos indica que la dirección tomada es la correcta, pero aun falta refinar el modelo.

4.3. Modelo de intercepto variante por estado y división y pendientes fijas

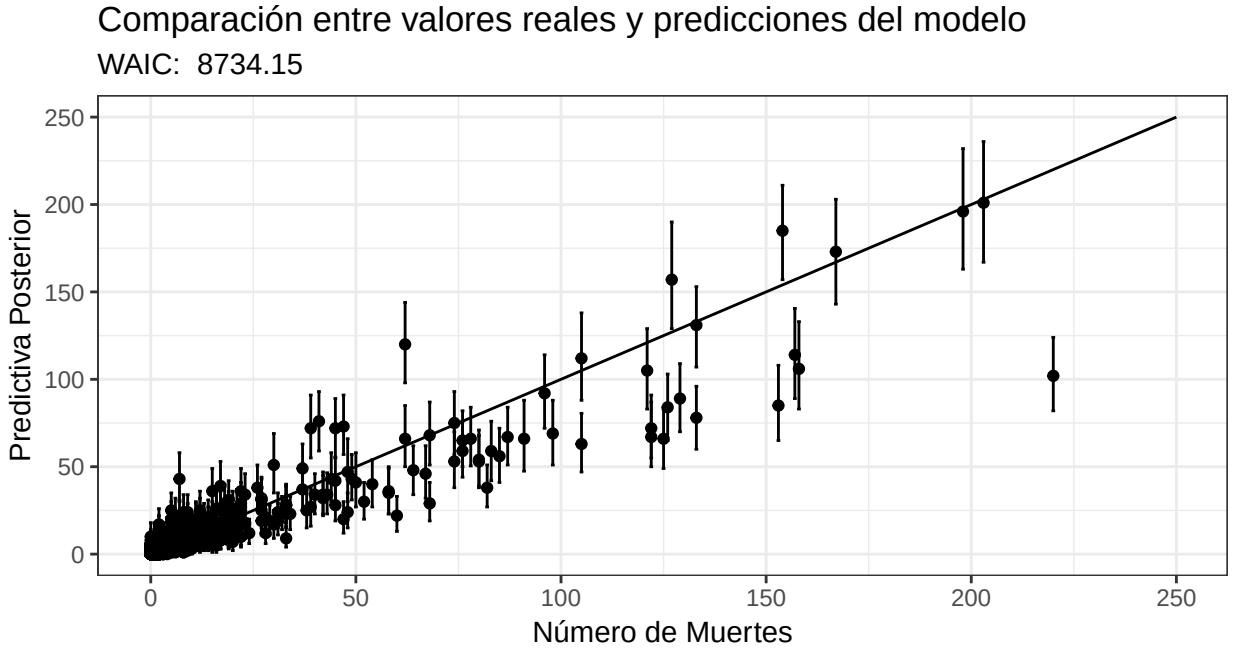
Para empezar a mejorar el modelo se agregaron las siguientes covariables:

- Porcentaje de la población que es afroamericana.
- Porcentaje de la población por debajo de nivel de pobreza.
- Porcentaje de la población que no habla bien inglés.
- Porcentaje de de la población que reside en el mismo estado donde nació.
- Porcentaje de la población que tiene 25 o más y no se graduó de preparatoria.
- Porcentaje de madres que trabajan con hijos menores a 18 años.
- Porcentaje de la población que inmigró en los últimos 10 años.

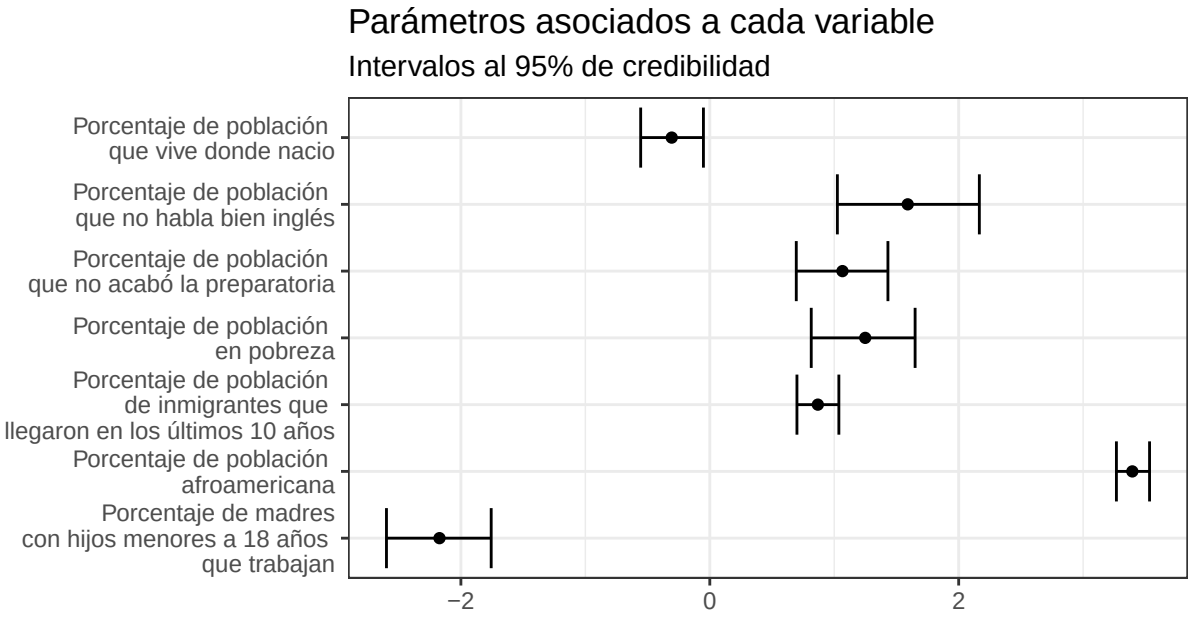
y se planteo el mismo modelo anterior solo que ahora el modelo considera covariables de la siguiente forma:

$$
\begin{aligned}
y_i &\sim Po(n_i\lambda_i) && i \in \{1,\dots,2215\} \\
\lambda_i &= exp(\beta_{0s} + X'\beta) && i \in \{1,\dots,2215\},\ s \in \{1,\dots,48\} \\
\beta_{0s} &\sim N(\theta_{0d}, \sigma^2_{\beta_0 d}) && s \in \{1,\dots,48\},\ d \in \{1,\dots,9\} \\
\theta_{0d} &\sim N(\phi, \sigma^2_{\phi}) && d \in \{1,\dots,9\} \\
\phi &\sim N(0, 10) \\
\sigma_{\phi} &\sim \Gamma(0.001, 0.001) \\
\beta_j &\sim N(0, 1) && j \in \{1,\dots,k\}
\end{aligned}
$$

El desempeño del modelo es el siguiente:



Se puede observar una reducción dástrica al WAIC y el ajuste del modelo ha mejorado bastante. Además se muestran los parámetros de las covariables dentro del modelo:



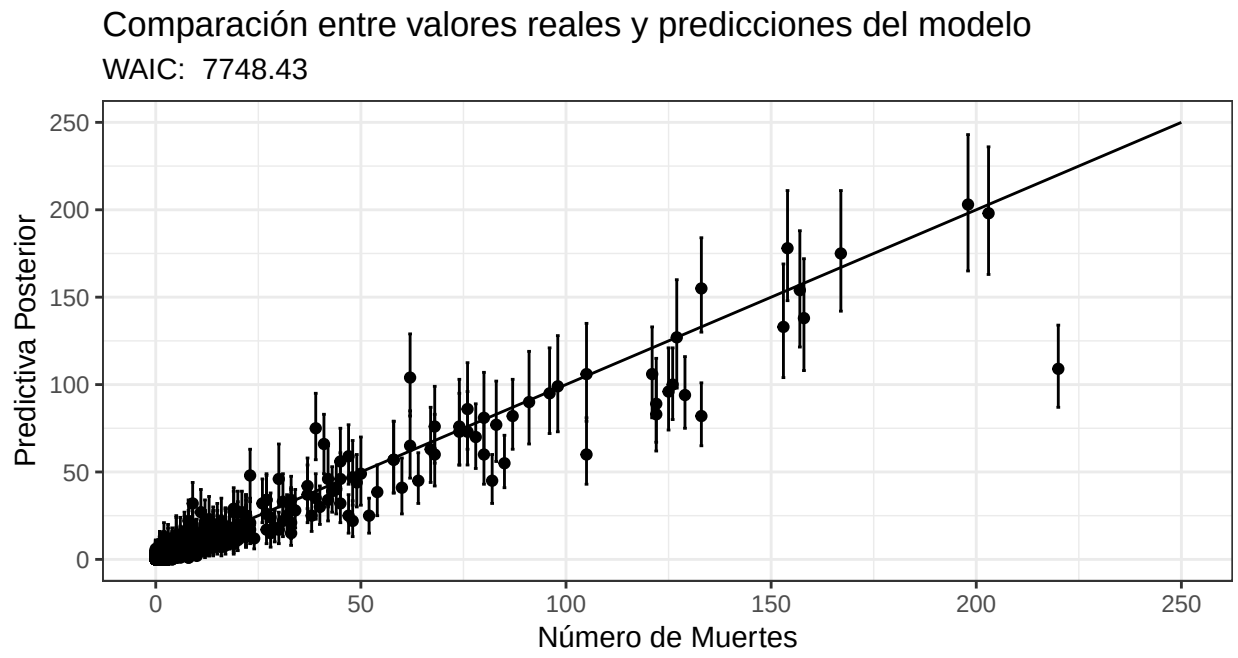
Estas variables se seleccionaron de tal forma que ninguna contuviera en su intervalo de credibilidad al 0 para confirmar que las variables tuvieran un efecto sobre la tasa de asesinatos. Se puede observar que la variable que más contribuye de manera positiva es el porcentaje de población afroamericana y después la que más contribuye de forma negativa es el porcentaje de madres con hijos menores a 18 que trabajan, esto fue una sorpresa pues no se observaba un efecto claro en el análisis exploratorio de los datos, pero controlando por efectos divisionales, estatales y las demás covariables muestra un efecto negativo en la tasa de asesinatos en EUA.

4.4. Modelo de intercepto variante por estado y división y pendientes variables por estado y división

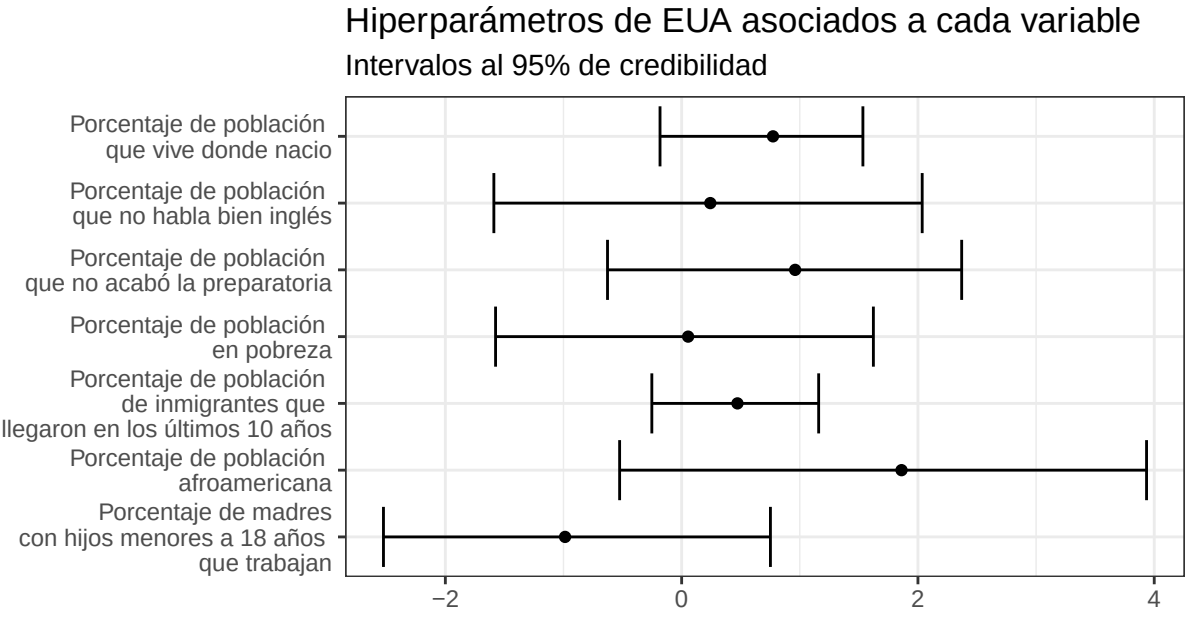
Sea X_i un conjunto de k covariables para la observación i, el modelo se define de la siguiente forma:

$$
\begin{aligned}
y_i &\sim Po(n_i\lambda_i) && i \in \{1,\dots,2215\}\\
\lambda_i &= exp(\beta_{0s} + X'\beta_s) && i \in \{1,\dots,2215\},\ s \in \{1,\dots,48\}\\
\beta_{0s} &\sim N(\theta_{0d}, \sigma^2_{\beta_0 d}) && s \in \{1,\dots,48\},\ d \in \{1,\dots,9\}\\
\beta_{sj} &\sim N(\theta_{dj}, \sigma^2_{\beta d}) && s \in \{1,\dots,48\},\ d \in \{1,\dots,9\}\ j \in \{1,\dots,k\}\\
\theta_{0d} &\sim N(\phi, \sigma^2_{\phi}) && d \in \{1,\dots,9\}\\
\theta_{dj} &\sim N(\phi_j, \sigma^2_{\phi_j}) && d \in \{1,\dots,9\}\ j \in \{1,\dots,k\}\\
\phi &\sim N(0, 10)\\
\sigma_\phi &\sim \Gamma(0.001, 0.001)\\
\phi_j &\sim N(0, 1) && j \in \{1,\dots,k\}
\end{aligned}
$$

A continuación se muestra el resultado predictivo del modelo:



Se puede observar que el WAIC disminuyó, sin embargo, no tuvo una caída tan grande como entre el modelo 3 y el modelo 4. A continuación se muestran los efectos globales para EUA de cada variable:



A partir de esta gráfica podemos ver que cuando se dividen los efectos por división y estado de todas las variables no hay contribuciones que no contengan al 0 dentro de su intervalo de credibilidad, por lo que este modelo es poco interpretable y por lo tanto será descartado a pesar que su WAIC sea menor.

4.5. Modelo seleccionado

Se seleccionó el modelo 3 para modelar la tasa de asesinatos en EUA y por lo tanto aquí es donde se hará la interpretación del modelo.

Referencias

Gelman, Andrew, Jessica Hwang, and Aki Vehtari. 2014. "Understanding Predictive Information Criteria for Bayesian Models." *Statistics and Computing* 24 (6): 997–1016. https://doi.org/10.1007/s11222-013-9416-2.