Proyecto Final Regresión Avanzada

Análisis Jerárquico de asesinatos en EUA

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

Contents

1 Introducción

Desde hace mucho tiempo, los noticieros y periódicos han estado infestados de noticias violentas y este fenómeno parece ir en aumento no sólo en México sino a nivel mundial. Uno de los países particularmente violentos es Estados Unidos, donde es frecuente escuchar de masacres de decenas de personas perpetuadas a lo largo del país, y éstas pueden suceder en casi cualquier lugar, desde escuelas, restaurantes hasta iglesias, y de multiples maneras en cada condado del país.

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". En esta nota se adjudica que la disminución en las tasas fue ocasionada por tácticas políticas más agresivas, así como por un incremento del número de criminales en prisión y patrones cambiantes en el uso de drogas, sin embargo, estos no son los únicos factores que podrían influir en la tasa de asesinatos de un país. Muchos factores pueden dar origen a una racha de violencia y su nivel de influencia podría variar en cada división y cada estado del país. Es por eso que sería muy informativo tener una visión más general de cómo la demografía y circunstancia en cada división y estado puede cambiar la tasa de homicidios.

Dado que la violencia, específicamente el asesinato, es un tema muy delicado e importante, se partió de esta situación para definir el proyecto a tratar en este proyecto.

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. Sorprendentemente, en ese año la tasa de homicidios disminuyó en algunas de las ciudades más violentas de los Estados Unidos de Norteamérica.

Cabe mencionar, que la base de datos contiene información de 48 estados y 9 divisiones identificadas como: New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain y Pacific, e incluye ciertas variables que podrían considerarse como factores relacionados a la tasa de asesinatos del país.

Dado este contexto, el propósito de este proyecto es explicar la tasa de asesinatos en los diferentes condados de Estados Unidos, a través de modelos de regresión que consideren efectos por estado, división censal y variables explicativas.

En particular, en este trabajo realizamos un análisis exploratorio univariado y bivariado entre la variable de respuesta y las covariables identificadas, así como también exploramos distintos modelos de regresión para entender las relaciones subyacentes en los datos, seleccionando el modelo que mejor explique la tasa de asesinatos.

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 [1]. 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 años 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 migró en los últimos 10 años.

Adicionalmente se sabe el 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 tiene información:

Table 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

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

Estado	Número de condados	Estado	Número de condados
Indiana	48	West Virginia	14
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 jerárquico puede combinar la información de estados con más condados, como California, con estados que tienen menor cantidad de condados, como Columbia, Delaware y Kansas.

También se muestra el número de estados por división:

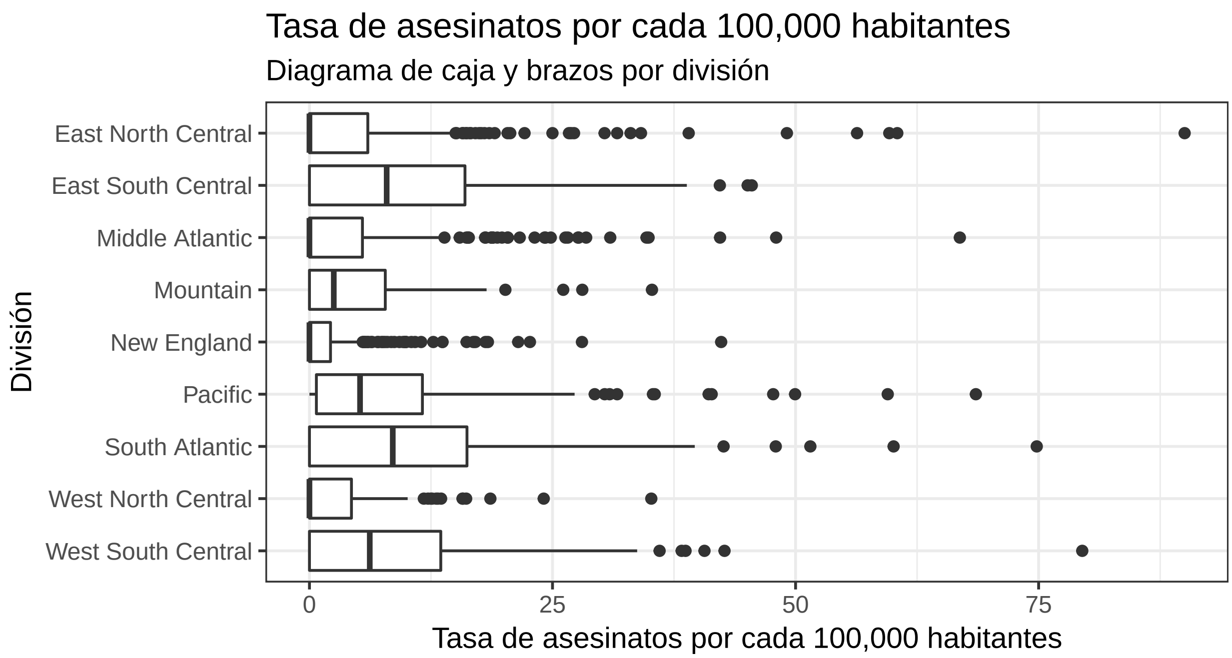
División	Número de estados
South Atlantic	9
Mountain	8
West North Central	7
New England	6
East North Central	5
Pacific	5
East South Central	4
West South Central	4
Middle Atlantic	3

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.

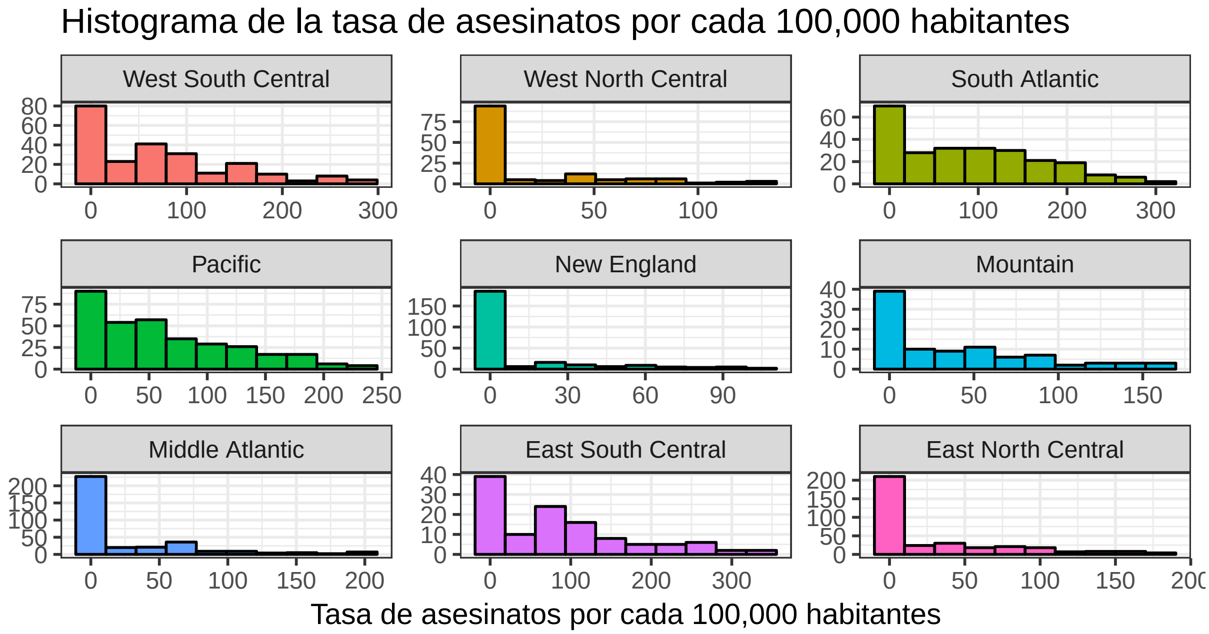
3 Análisis exploratorio de los datos

3.1 Análisis Univariado

Para entender los asesinatos en EUA se muestran las siguientes gráficas que son la tasa de asesinatos. 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 sólo conserva observaciones menores al cuantil 95:

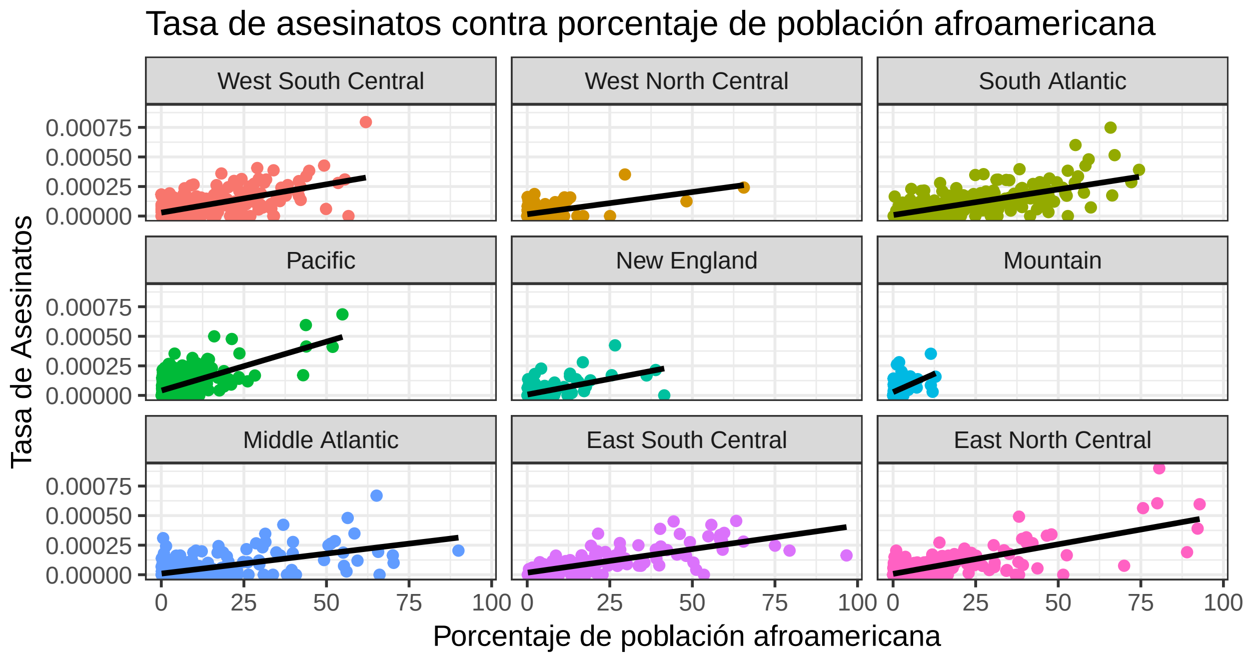


Con esta gráfica se nota que independientemente del estado, existen muchos condados con tasa de asesinatos cercana al 0. Esto es un punto importante a notar a la hora del modelado. Además se puede observar que las divisiones con más condados que tienen pocos asesinatos son West North Central, New England,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

3.2 Análisis Bivariado

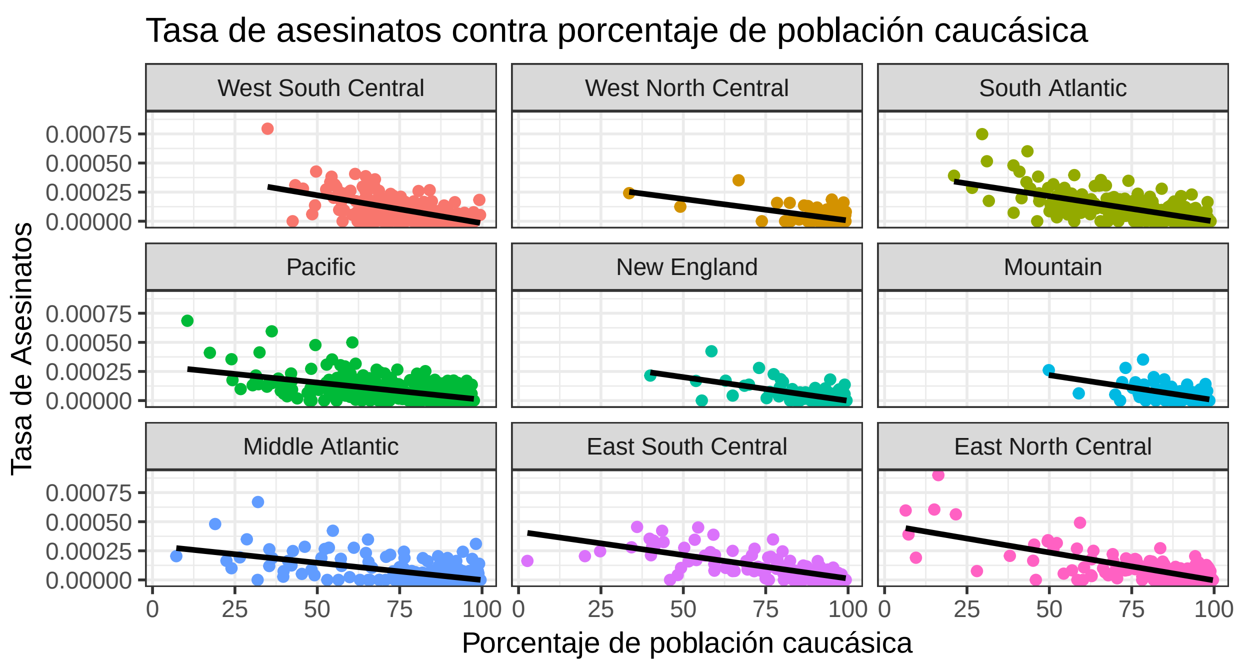
El siguiente paso es análizar las posibles relaciones entre la tasa de asesinatos y las variables sociodemográficas. Como ayuda visual se ajustaron modelos simples de la relación entre tasa de asesinatos y la variable gráficada, sólo para poder entender si la variable tiene o no tiene relación con el 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:



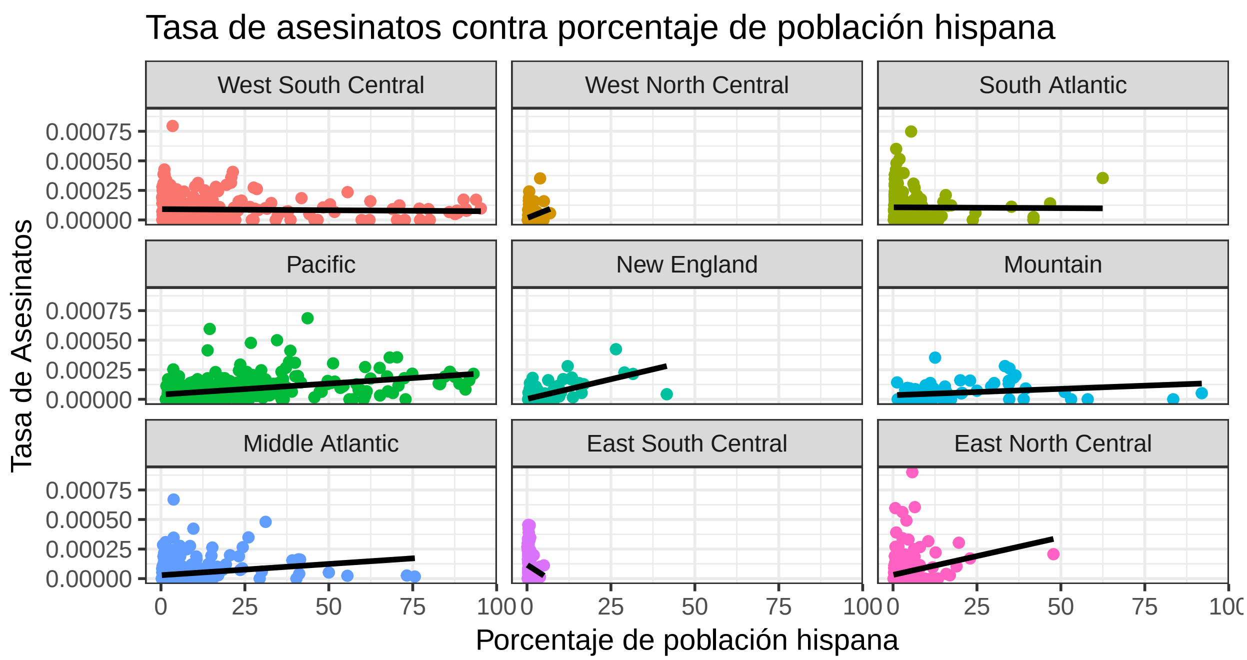
Lo primero a notar es que todos los condados dentro de las divisiones de Mountain, New England, Pacific y West North Central tienen un porcentaje de población afroamericana menor a un 60%, lo cual llevaría a pensar que dentro de estas divisiones no existe mucha diversidad racial.

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.

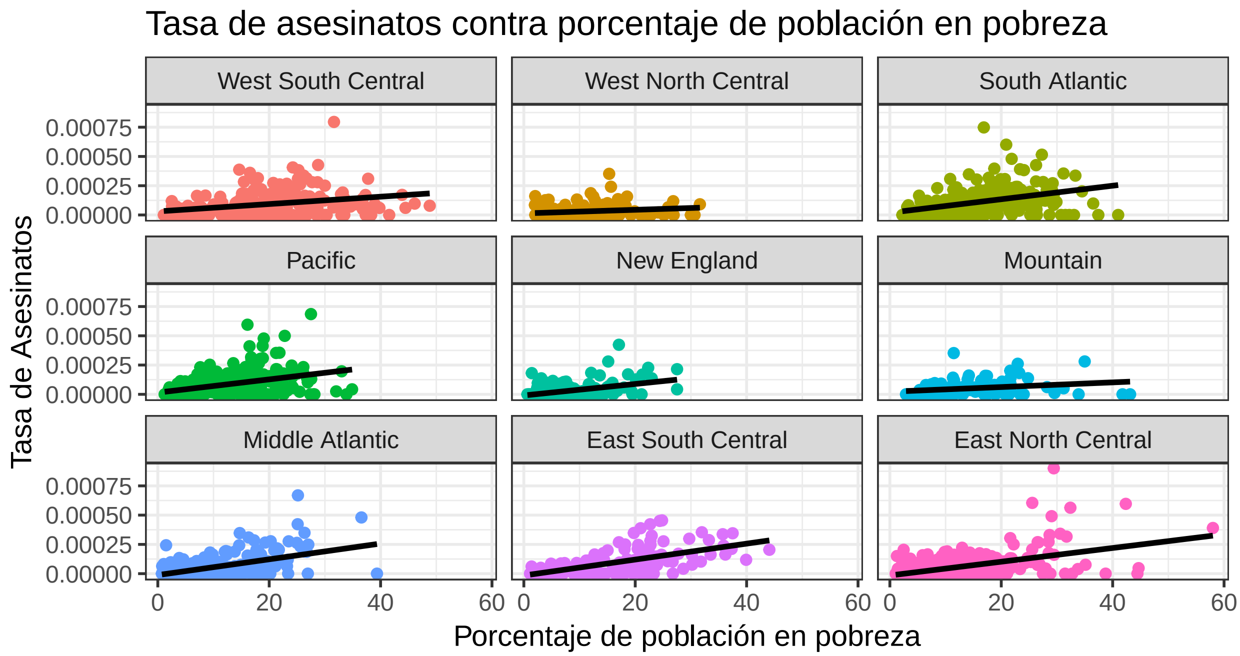


Con esta gráfica se confirma la hipótesis anterior, que en las divisiones de Mountain, New England y West North Central no hay mucha diversidad racial dentro de los condados de estas divisiones. Esto puede ser un potencial problema para la estimación de los modelos con ambas variables pues estaríamos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

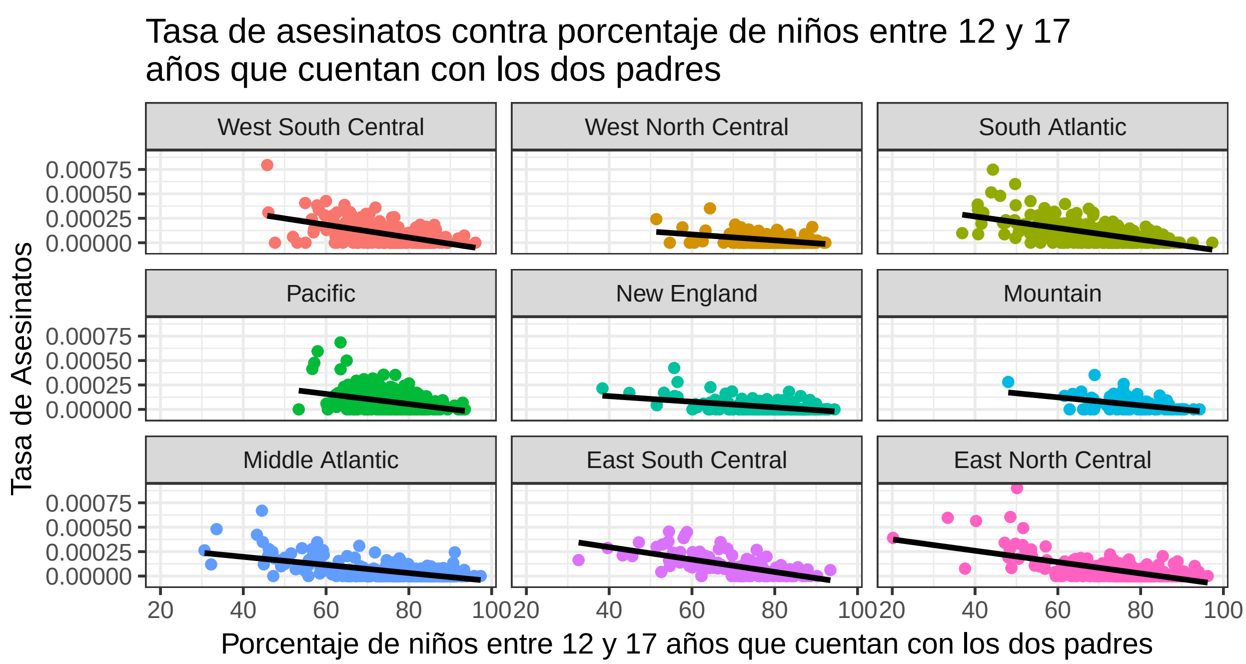
Se observa que para todas las divisiones, conforme aumenta el porcentaje de personas caucásicas, disminuye la tasa de asesinatos.



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 tiene 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.



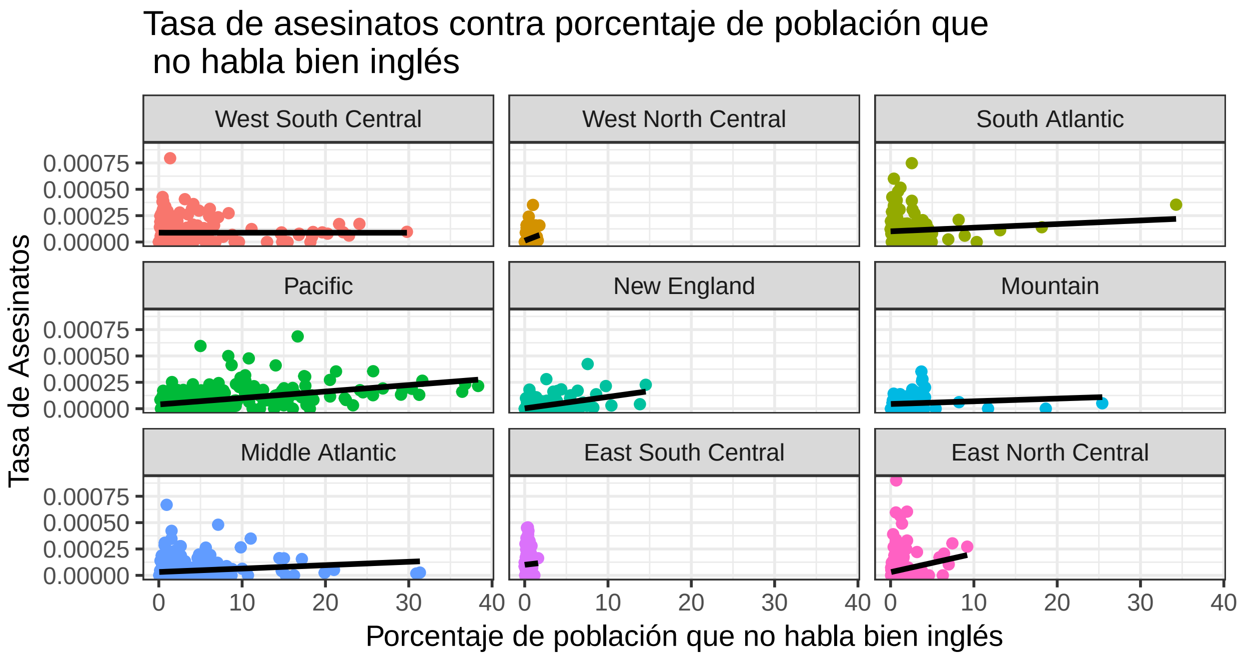
Se puede observar una posible relación positiva entre el porcentaje de población en pobreza y la tasa de asesinatos. New England parece tener los índices de pobreza más bajos de todas las divisiones (menores al 50%).



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

- Los porcentajes altos de niños que cuentan con ambos padres podrían indicar una población conformada en su mayoría por familias unidas aunque la dinámica de estas familias queda fuera del alcance de este trabajo.

- La relación que muestra con la tasa de asesinatos parece ser negativa.

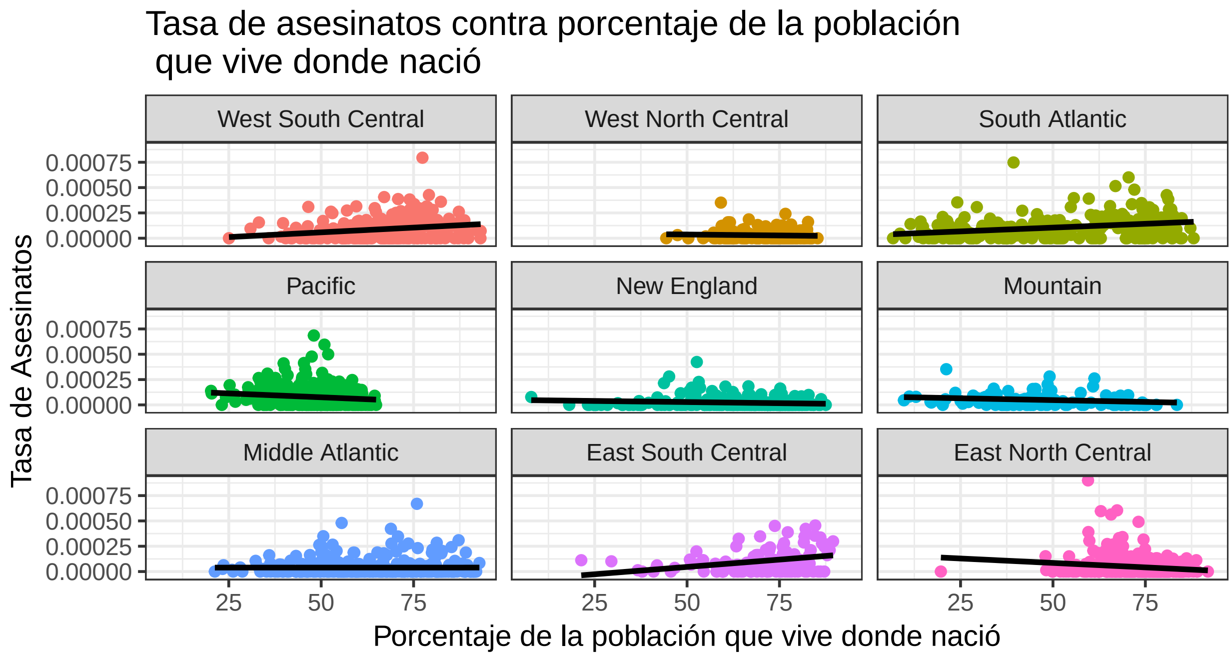


Se observan tres tipos de comportamientos distintos:

- 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%.

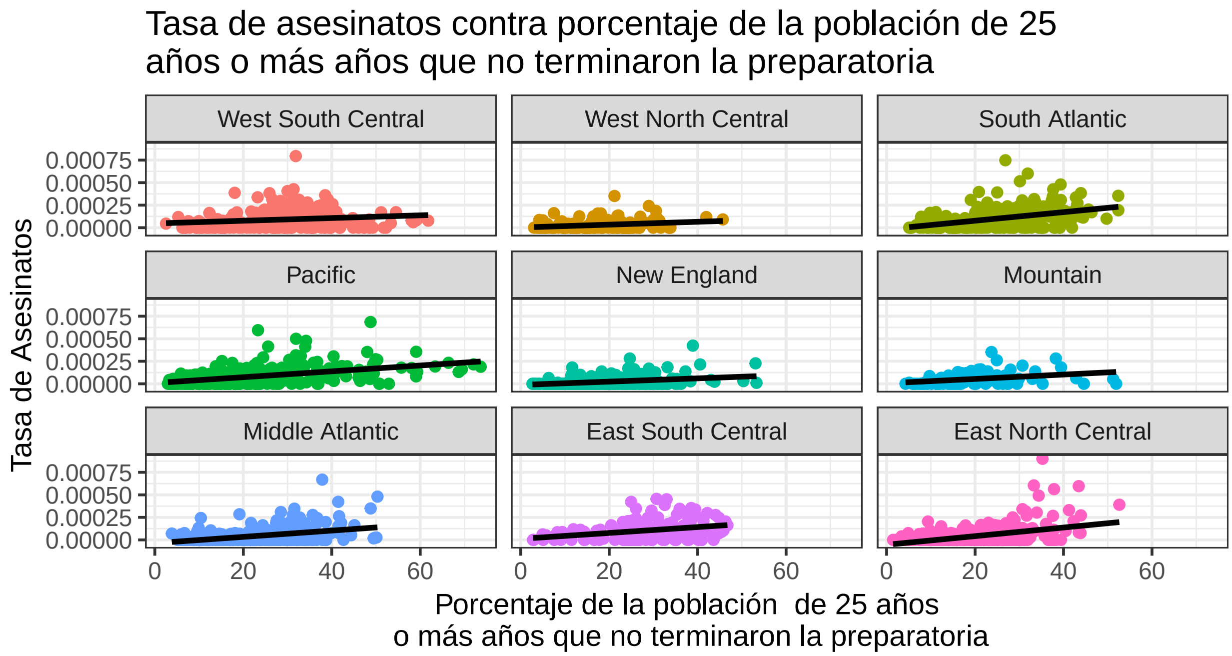
- Las divisiones de East North Central y New England: sus observaciones puntuales muestran porcentajes menores al 15%, es decir, que sí existe una proporción que no se puede ignorar de personas que hablan mal el inglés. Las observaciones se encuentran variando mucho respecto a la tasa de asesinatos. 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. La división de Pacific parecería sí tener una mayor tasa de asesinatos a mayor porcentaje de personas que no hablan bien el inglés.

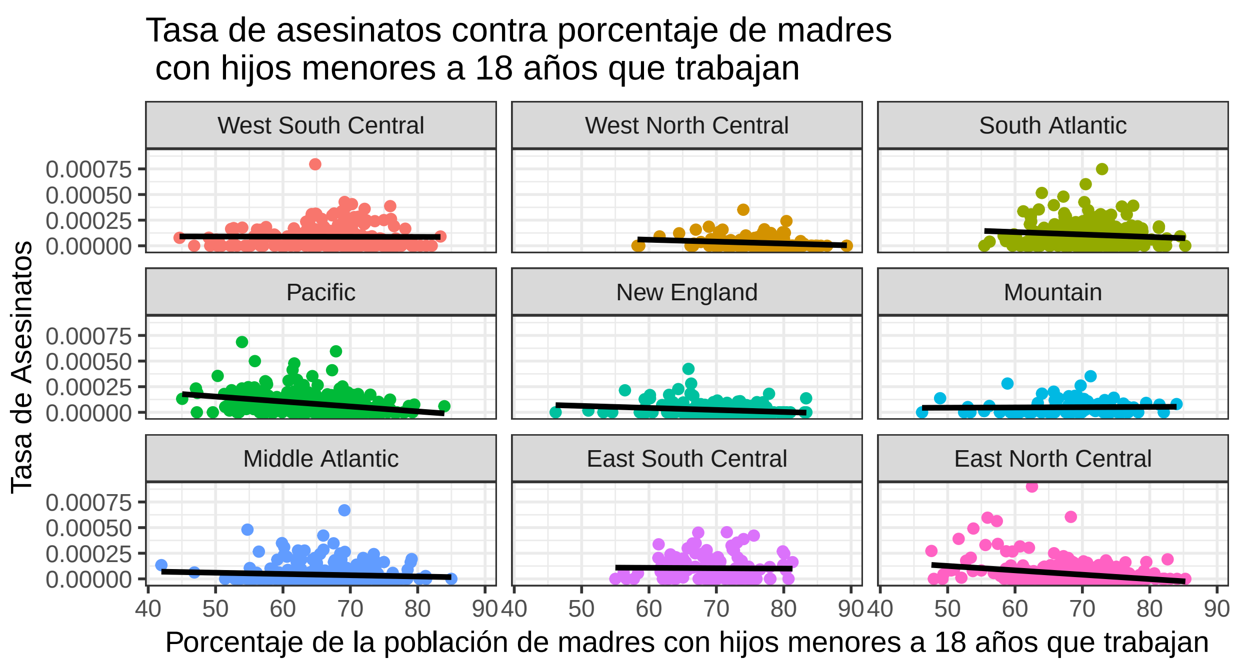


- En las divisiones de East North Central, East South Central y West North Central existen muchas personas oriundas de su división. No existe una relación evidente con la tasa de asesinatos ya que las tasas altas de asesinatos se encuentran dispersas sin importar si existen altos o bajos porcentajes 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 esta población.
- 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%.
- Aunque las divisiones de South Atlantic y West South Central 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.

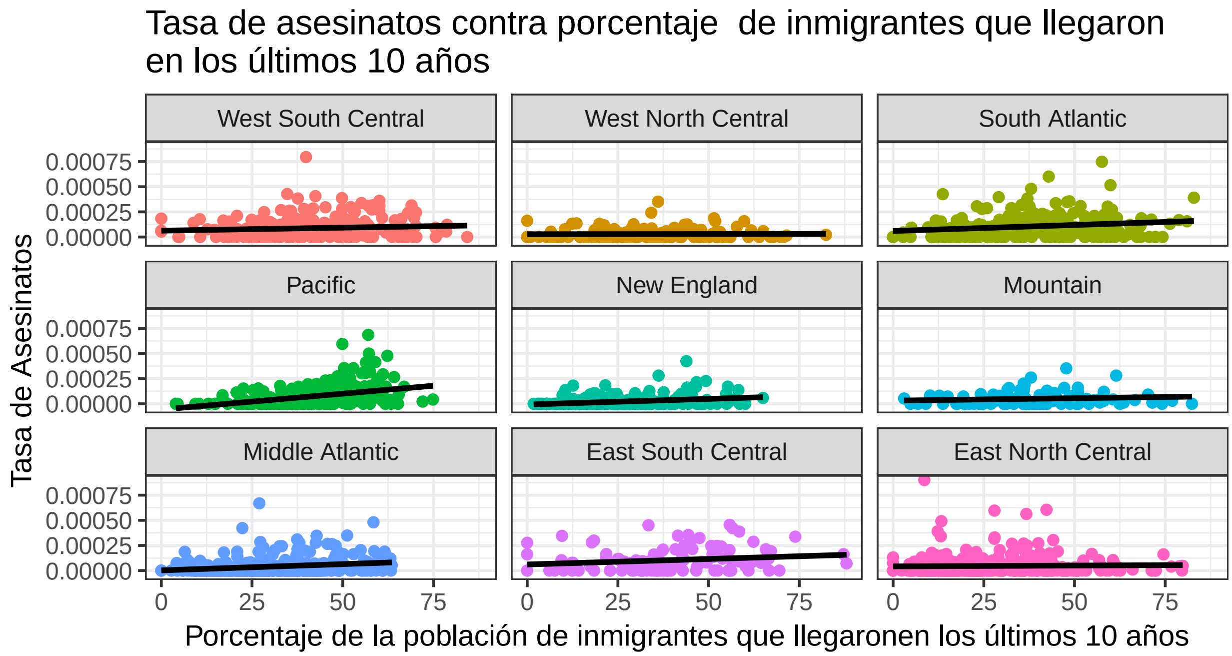
Dado este análisis, aparentemente la inmigración no afecta la tasa de asesinatos, ya que sin importar si las personas son o no originarias de su residencia actual, las tasas de asesinatos son muy dispersas, sin embargo esto se analizará más adelante.



En todas las divisiones la tasa de asesinatos es baja en condados con bajo porcentaje de población sin preparatoria. 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 esta tiene 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 covariables 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* sólo se utiliza la liga logarítmica. Los modelos probados son los siguientes:

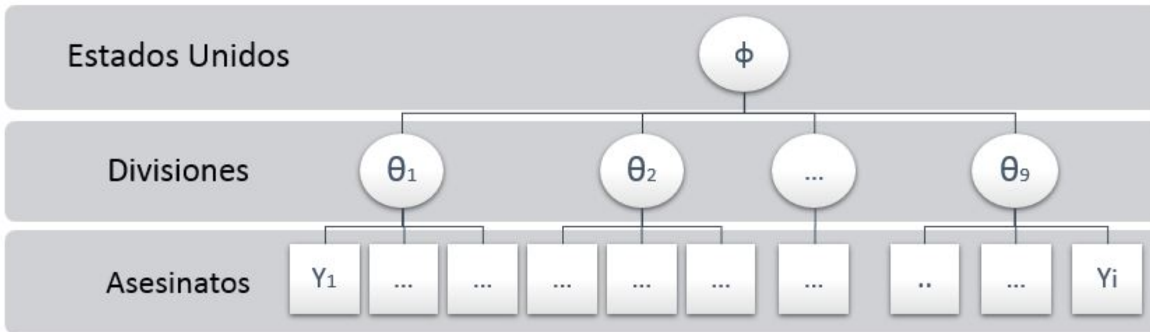
Modelo Poisson

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

Modelo Binomial

$$
\begin{aligned}
y_i &\sim Bin(n_i, \pi_i) & i \in \{1, \ldots, 2215\} \\
\pi_i &= f(\theta_d) & i \in \{1, \ldots, 2215\},\ d \in \{1, \ldots, 9\} \\
\theta_d &\sim N(\phi, \sigma_\phi^2) & d \in \{1, \ldots, 9\} \\
\phi &\sim N(0, 10) & \\
\sigma_\phi &\sim \Gamma(0.001, 0.001) & \\
f(x) &= \begin{cases} logit^{-1}(x) & \text{Liga Logistica} \\ \Phi(x) & \text{Liga Probit} \\ e^{-e^x} & \text{Liga log-log} \\ 1 - e^{-e^x} & \text{Liga log-log complementaria} \end{cases} &
\end{aligned}
$$

La representación de la jerarquía:



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 criterio 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	Logarítmica	19891

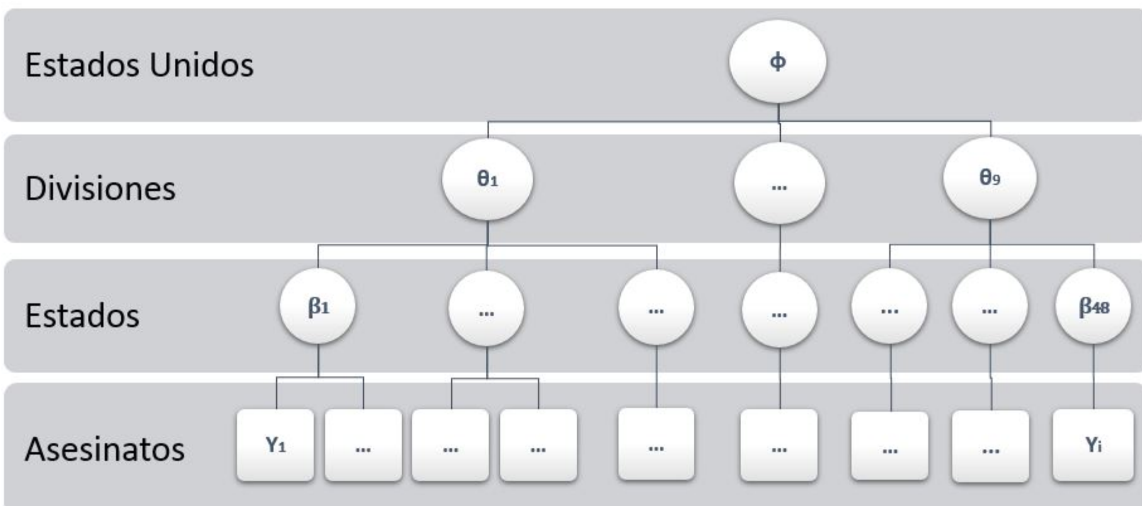
dado el criterio del WAIC, el mejor modelo es el *Poisson* con liga logarítmica. Es por esta razón que para los siguientes modelos se utilizará la verosimilitud *Poisson* con liga *logarítmica*.

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

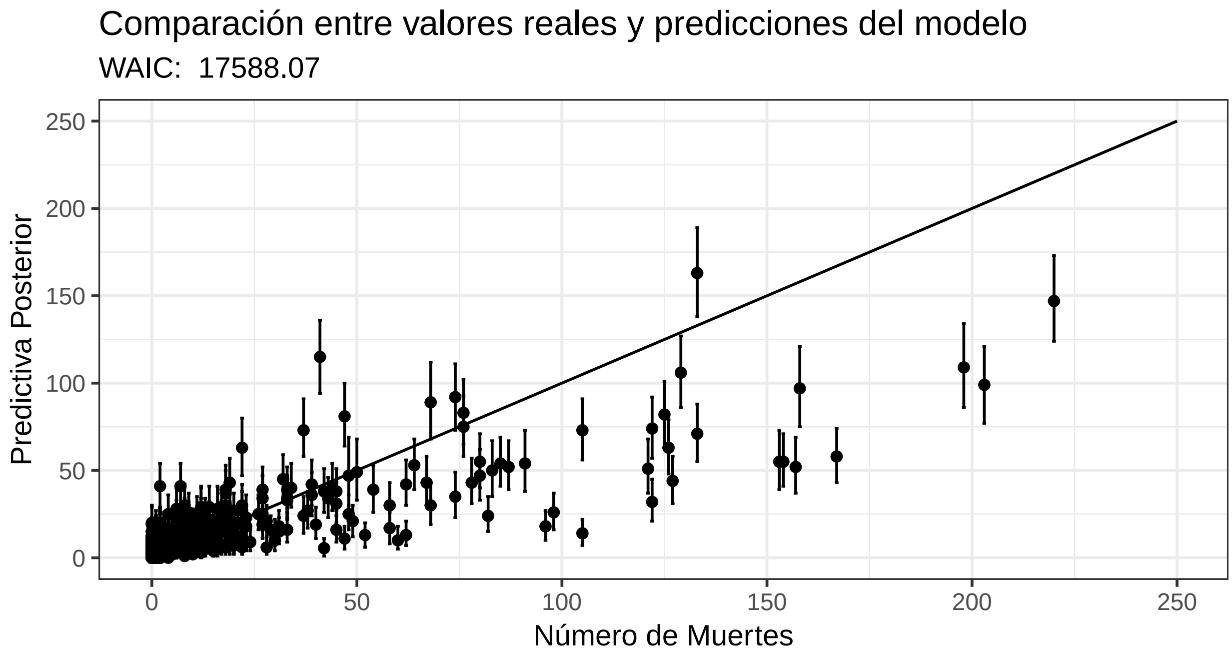
En este modelo se le agrega una jerarquía al modelo, es decir, ahora los interceptos varían por estado y estos a su vez dependen de los hiperparámetros de las distintas divisiones a las que pertenecen y finalmente al hiperparámetro que representa a EUA. El modelo se define de la siguiente forma:

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

La representación de la jerarquía:



A continuación se muestra el desempeño del modelo:



La gráfica muestra la recta de 45° 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

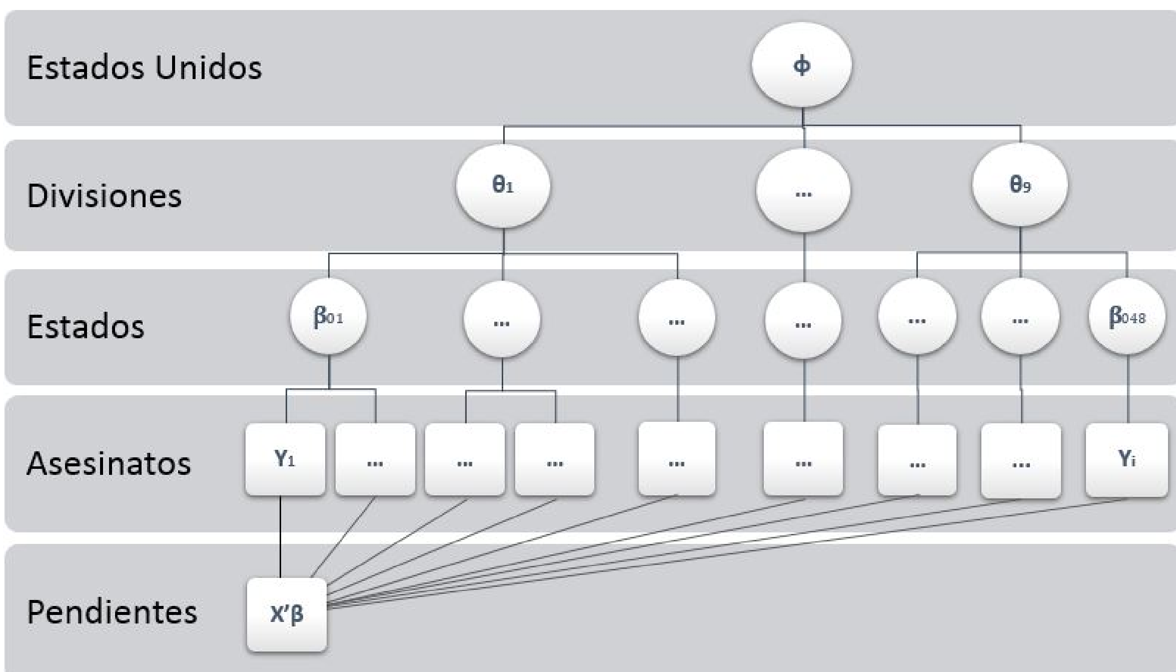
Para empezar a mejorar el modelo se agregaron las siguientes covariables:

- 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:

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

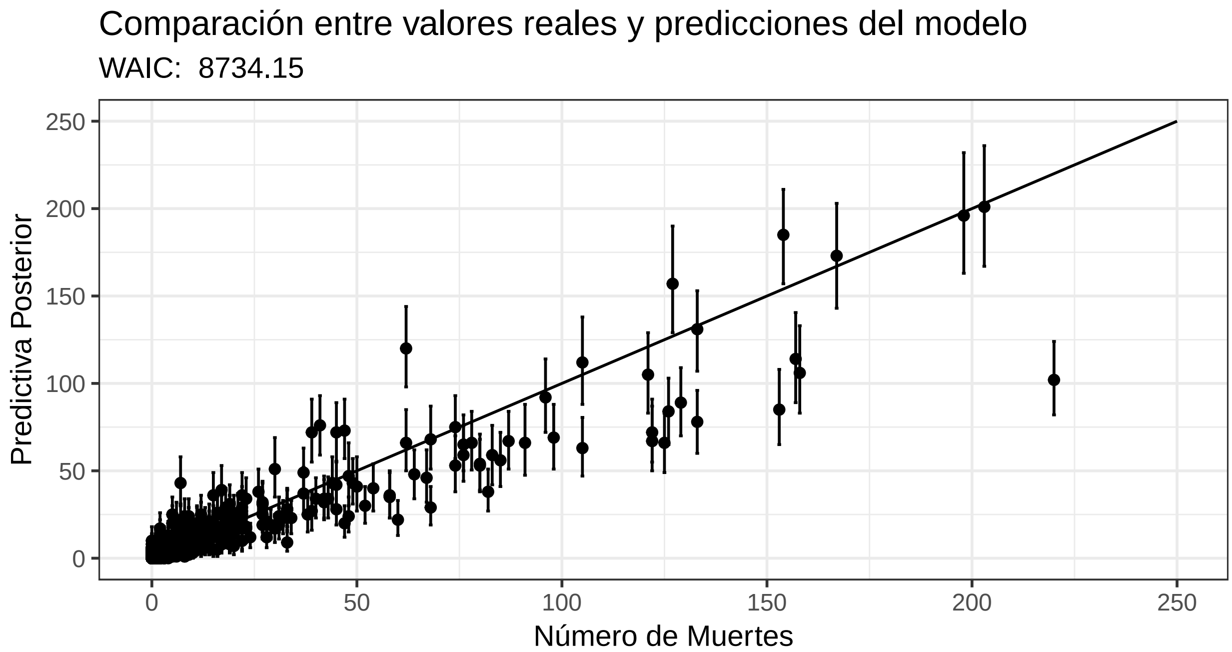
A continuación se muestra la representación de la jerarquía:



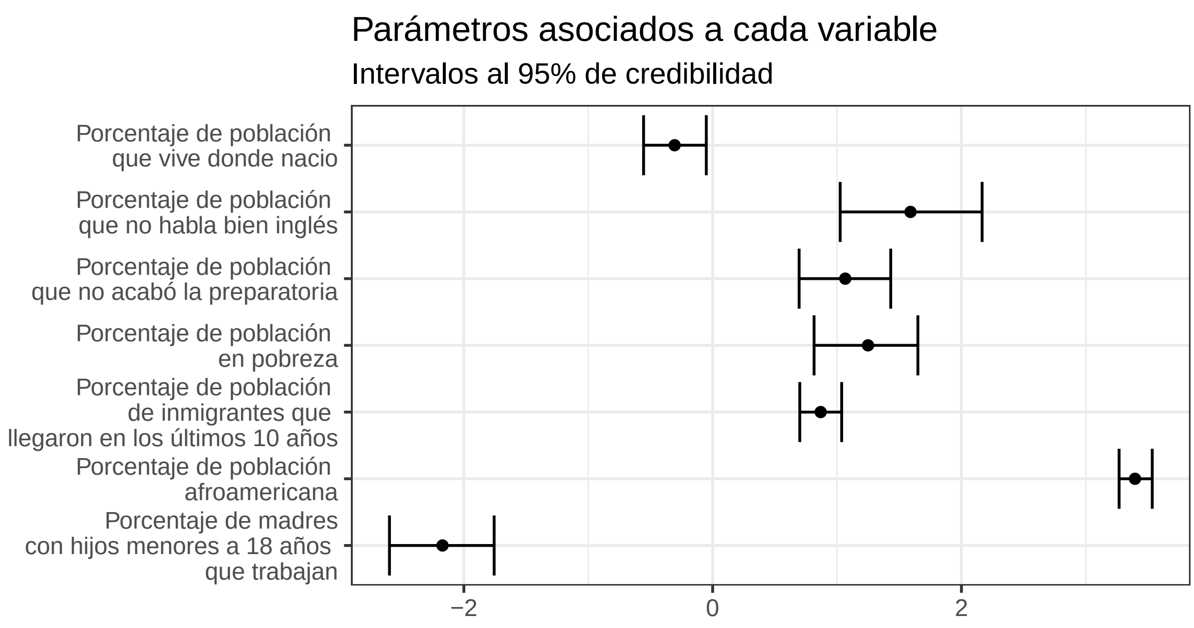
Es importante recalcar que tanto en este modelo como en el siguiente, los parámetros asociados a las desviaciones estándar tienen una inicial impropia sobre todos los reales positivos. Esta decisión fue tomada porque Stan converge mejor a la posterior con este tipo de inciales [2].

El desempeño del modelo es el siguiente:

[2]https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations



Se puede observar una reducción drástrica 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 (originalmente habímos considerado usar además las variables del porcentaje de población hispana, porcentaje de población caucásica y porcentaje de niños entre 12 y 17 años que viven con ambos padres). 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.

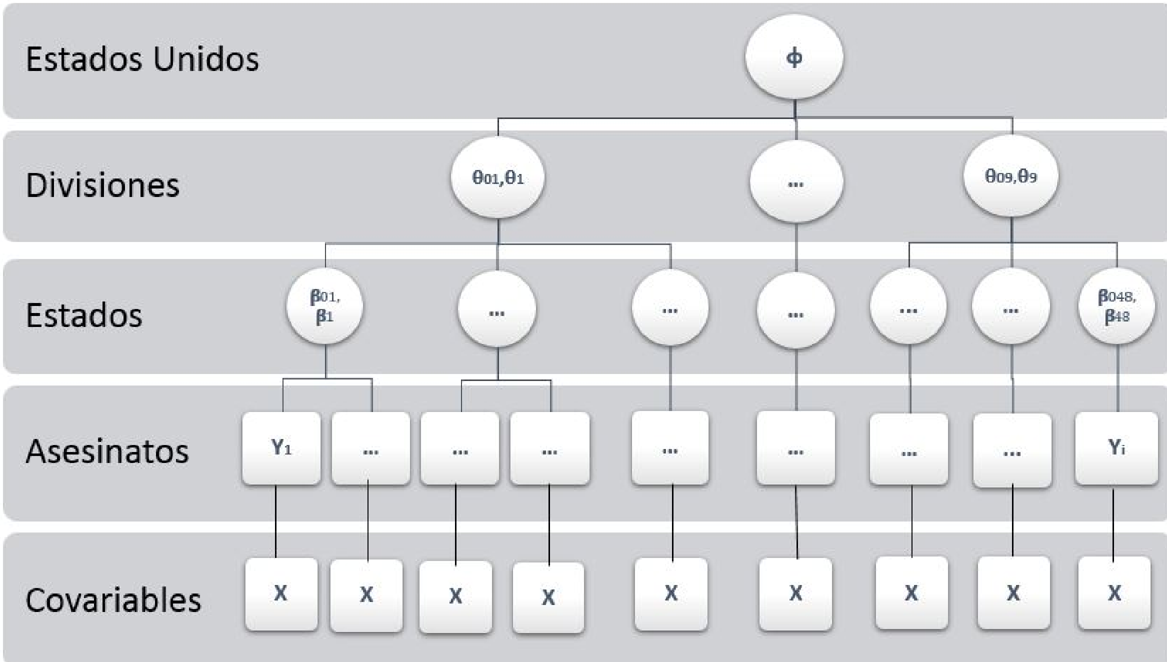
4.4 Modelo de intercepto variante por estado y división y pendientes variables por estado y división

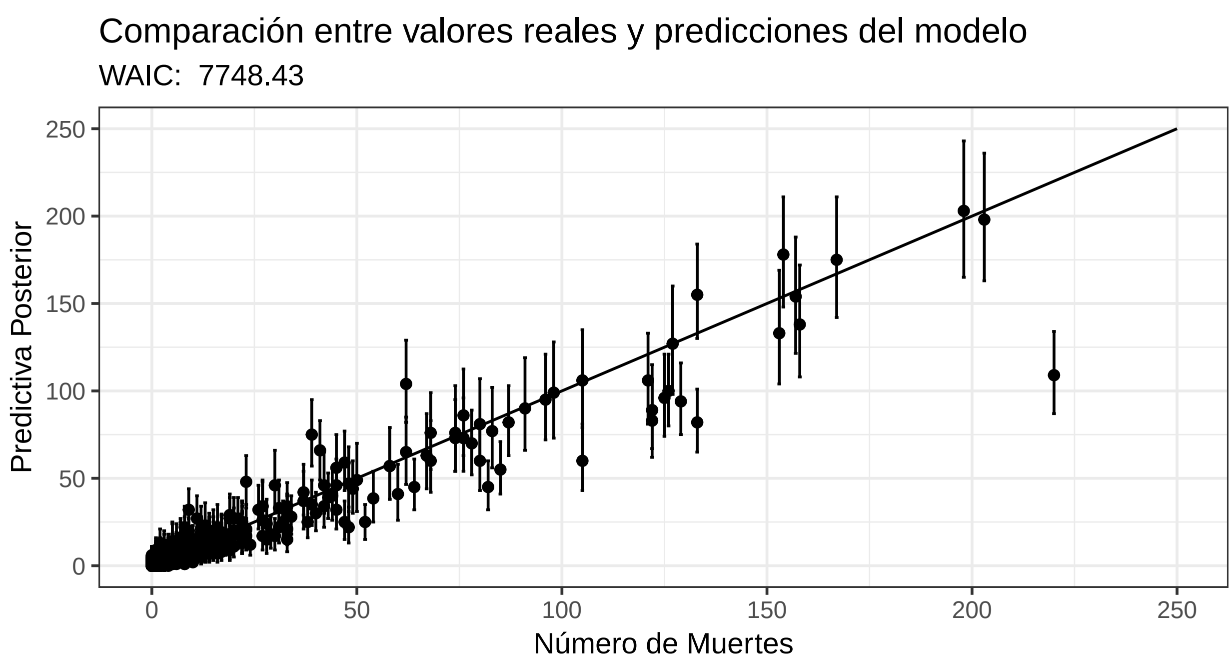
Sea X_i un conjunto de k covariables para la observación i, el modelo se define de la siguiente forma:

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

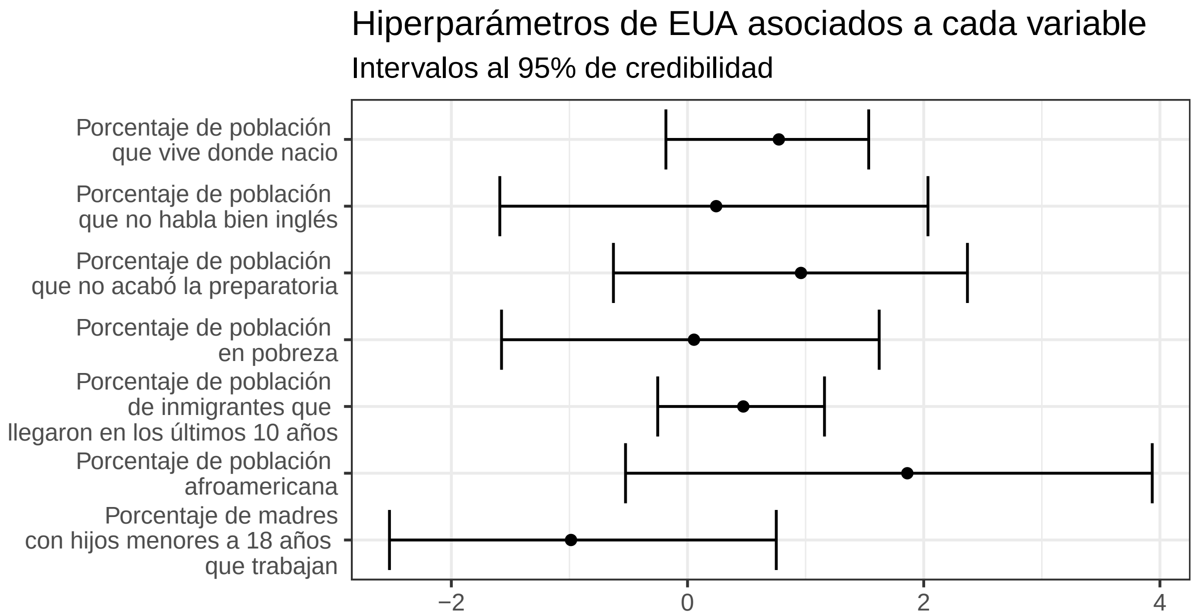
A continuación se muestra la representación de la jerarquía:

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 2 y el modelo 3. A continuación se muestran los efectos globales para EUA de cada variable:

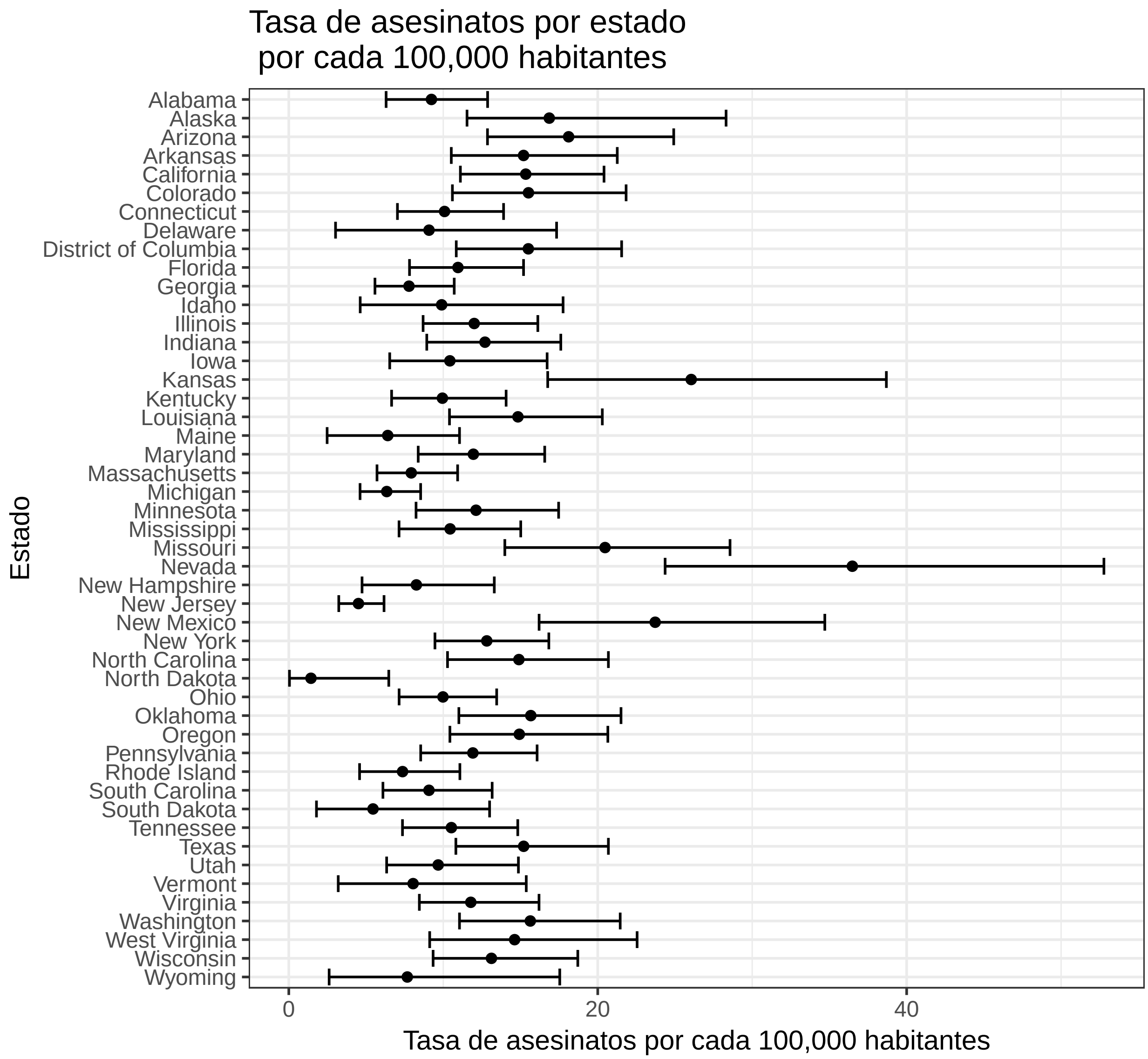


A partir de esta gráfica podemos ver que cuando se dividen los efectos por división y estado de todas las variables no hay contribuciones que no contengan al 0 dentro de su intervalo de credibilidad, por lo que este modelo es poco interpretable y por lo tanto será descartado a pesar que su WAIC sea menor.

4.5 Modelo seleccionado

Se seleccionó el modelo 3 (Modelo de intercepto variante por estado y división y pendientes fijas) para modelar la tasa de asesinatos en EUA. En este modelo se agregan las 7 variables explicativas que muestran tener un impacto sobre la tasa de asesinatos.

En el primer esquema, se ubican las tasas base de asesinatos por cada $100,000$ habitantes, i.e. $100,000e^{\beta_{0s}}$, para cada uno de los 48 estados. Sin embargo, para estados de Kansas, Missouri, Nevada y New Mexico, los parámetros parecen tener un valor mayor al de la mayoría, esto es bueno pues quiere decir que es probable que falten variables para explicar el fénomeno y lo está capturando el intercepto del estado, algo que no podría hacer un modelo no jerárquico.



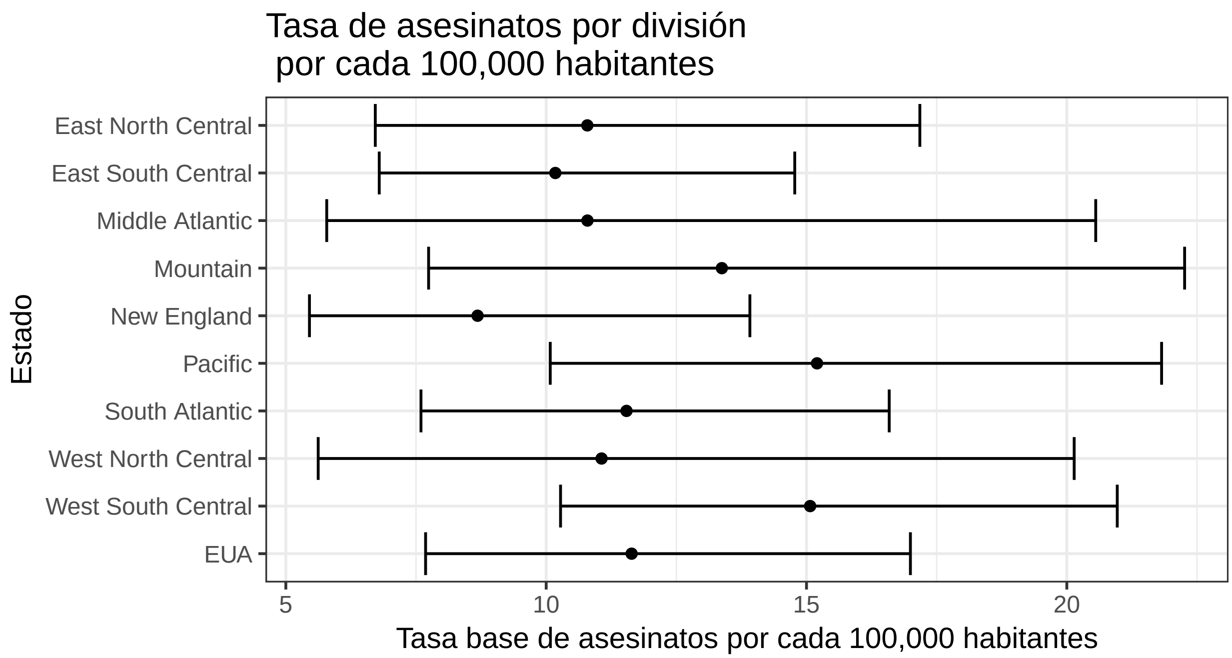
A continuación se muestran los valores de los interceptos por estado con un intervalo al 95% de credibilidad. Además se muestra la Rhat que es un criterio de convergencia de las cadenas de Markov y el número efectivo de observaciones para cada parámetro:

Parámetros	Promedio	2.5%	97.5%	Tamaño de muestra efectiva	Rhat
$\beta_{0,\text{Alaska}}$	-8.67	-9.07	-8.17	665	1
$\beta_{0,\text{Alabama}}$	-9.30	-9.67	-8.96	407	1
$\beta_{0,\text{Arkansas}}$	-8.80	-9.16	-8.46	422	1
$\beta_{0,\text{Arizona}}$	-8.62	-8.96	-8.30	417	1
$\beta_{0,\text{California}}$	-8.79	-9.11	-8.50	391	1
$\beta_{0,\text{Colorado}}$	-8.77	-9.15	-8.43	439	1
$\beta_{0,\text{Connecticut}}$	-9.21	-9.56	-8.88	461	1
$\beta_{0,\text{District of Columbia}}$	-8.77	-9.13	-8.44	419	1
$\beta_{0,\text{Delaware}}$	-9.35	-10.41	-8.66	891	1

Parámetros	Promedio	2.5%	97.5%	Tamaño de muestra efectiva	Rhat
$\beta_{0,\text{Florida}}$	-9.12	-9.46	-8.79	421	1
$\beta_{0,\text{Georgia}}$	-9.46	-9.79	-9.14	397	1
$\beta_{0,\text{Iowa}}$	-9.17	-9.64	-8.70	663	1
$\beta_{0,\text{Idaho}}$	-9.24	-9.98	-8.64	1159	1
$\beta_{0,\text{Illinois}}$	-9.03	-9.35	-8.73	383	1
$\beta_{0,\text{Indiana}}$	-8.98	-9.32	-8.64	430	1
$\beta_{0,\text{Kansas}}$	-8.26	-8.69	-7.86	525	1
$\beta_{0,\text{Kentucky}}$	-9.22	-9.62	-8.87	446	1
$\beta_{0,\text{Louisiana}}$	-8.82	-9.17	-8.50	390	1
$\beta_{0,\text{Massachusetts}}$	-9.44	-9.77	-9.12	477	1
$\beta_{0,\text{Maryland}}$	-9.03	-9.39	-8.71	442	1
$\beta_{0,\text{Maine}}$	-9.70	-10.60	-9.11	1166	1
$\beta_{0,\text{Michigan}}$	-9.67	-9.98	-9.37	429	1
$\beta_{0,\text{Minnesota}}$	-9.02	-9.40	-8.65	564	1
$\beta_{0,\text{Missouri}}$	-8.50	-8.87	-8.16	464	1
$\beta_{0,\text{Mississippi}}$	-9.17	-9.55	-8.80	380	1
$\beta_{0,\text{North Carolina}}$	-8.82	-9.18	-8.48	408	1
$\beta_{0,\text{North Dakota}}$	-11.43	-14.67	-9.65	1325	1
$\beta_{0,\text{New Hampshire}}$	-9.41	-9.96	-8.92	727	1
$\beta_{0,\text{New Jersey}}$	-10.01	-10.34	-9.69	404	1
$\beta_{0,\text{New Mexico}}$	-8.35	-8.73	-7.97	601	1
$\beta_{0,\text{Nevada}}$	-7.91	-8.32	-7.55	449	1
$\beta_{0,\text{New York}}$	-8.97	-9.27	-8.69	385	1
$\beta_{0,\text{Ohio}}$	-9.22	-9.55	-8.91	406	1
$\beta_{0,\text{Oklahoma}}$	-8.77	-9.11	-8.44	424	1
$\beta_{0,\text{Oregon}}$	-8.82	-9.17	-8.48	506	1
$\beta_{0,\text{Pennsylvania}}$	-9.04	-9.37	-8.74	398	1
$\beta_{0,\text{Rhode Island}}$	-9.52	-9.99	-9.11	534	1
$\beta_{0,\text{South Carolina}}$	-9.31	-9.71	-8.94	507	1
$\beta_{0,\text{South Dakota}}$	-9.87	-10.93	-8.95	1664	1
$\beta_{0,\text{Tennessee}}$	-9.16	-9.52	-8.82	405	1
$\beta_{0,\text{Texas}}$	-8.80	-9.13	-8.48	396	1
$\beta_{0,\text{Utah}}$	-9.24	-9.67	-8.81	616	1
$\beta_{0,\text{Virginia}}$	-9.05	-9.38	-8.73	440	1
$\beta_{0,\text{Vermont}}$	-9.45	-10.35	-8.78	931	1
$\beta_{0,\text{Washington}}$	-8.76	-9.11	-8.45	439	1
$\beta_{0,\text{Wisconsin}}$	-8.94	-9.28	-8.58	481	1
$\beta_{0,\text{West Virginia}}$	-8.84	-9.30	-8.40	544	1
$\beta_{0,\text{Wyoming}}$	-9.50	-10.55	-8.65	1155	1

De toda la tabla se aprecia que todas las cadenas de Markov convergen bajo el criterio de Rhat, pues todas son 1. El tamaño mínimo de muestra efectiva fue de 380 lo cual tiende a ocurrir en la estimación de modelos jerarquicos y es lo suficiente para poder hacer inferencia.

Para este segundo esquema, se graficaron los parámetros $\theta_{0,d}$resultantes para cada una de las 9 divisiones y el parámetro Φ global para capturar el efecto sobre los Estados Unidos de manera completa. Los intervalos obtenidos muestran poco movimiento de la tasa base ya que todas las medias fluctúan en un mismo rango de valores entre aproximadamente 10 y 15 sobre la tasa base de asesinatos.



Parámetros	Promedio	2.5%	97.5%	Tamaño de muestra efectiva	Rhat
$\theta_{0,\text{East North Central}}$	-9.13	-9.61	-8.67	741	1
$\theta_{0,\text{East South Central}}$	-9.20	-9.60	-8.82	414	1
$\theta_{0,\text{Middle Atlantic}}$	-9.13	-9.76	-8.49	910	1
$\theta_{0,\text{Mountain}}$	-8.92	-9.47	-8.41	844	1
$\theta_{0,\text{New England}}$	-9.35	-9.82	-8.88	598	1
$\theta_{0,\text{Pacific}}$	-8.80	-9.20	-8.43	617	1
$\theta_{0,\text{South Atlantic}}$	-9.08	-9.49	-8.70	431	1
$\theta_{0,\text{West North Central}}$	-9.12	-9.79	-8.51	1020	1
$\theta_{0,\text{West South Central}}$	-8.81	-9.18	-8.47	437	1

Cabe mencionar que estos parámetros no tienen mucha interpretación pues corresponden al caso donde todas las covariables $x = 0$ y como esto no es posible no es correcto interpretar los parámetros.

A continuación se muestra los diagnósticos de convergencia de los parámetros asociados a la pendiente de cada variable:

Parámetro asociado al porcentaje de...	Tamaño de muestra efectiva	Rhat
Población afroamericana	1811	1
Población en pobreza	1391	1
Población que no habla bien inglés	1975	1
Población que vive donde nacio	1856	1
Población que no acabó la preparatoria	1811	1
Madres con hijos menores a 18 años que trabajan	558	1
Población de inmigrantes que llegaron en los últimos 10 años	1179	1

Se observa que para todos los parámetros se tiene una Rhat igual a 1, entonces bajo este criterio tenemos convergencia de las cadenas. Además muestra que el tamaño efectivo de muestra mínimo es de 558 por lo tanto se puede realizar la inferencia con los parámetros estimados.

Ahora se presentan los efectos de los covariables y se realizará su interpretación. Para esto es necesario considerar que en el modelo las covariables se introdujeron en el rango del $[0, 1]$ por lo tanto la interpretación será la siguiente: ante un aumento del 1% de una variable explicativa x_i se obtendrá un cambio en la tasa y del $exp(.01 * \beta_i)$.

Parámetro asociado al porcentaje de...	Promedio	2.5%	97.5%
Población afroamericana	3.40	3.27	3.53
Población en pobreza	1.24	0.82	1.65
Población que no habla bien inglés	1.59	1.03	2.17
Población que vive donde nacio	-0.31	-0.56	-0.05
Población que no acabó la preparatoria	1.07	0.69	1.43
Madres con hijos menores a 18 años que trabajan	-2.17	-2.60	-1.76
Población de inmigrantes que llegaron en los últimos 10 años	0.87	0.70	1.04

- **Porcentaje de población afroamericana**: ante un aumento del 1% de población afroamericana dentro de un condado se tiene un aumento en la tasa de asesinatos de 3.46%.
- **Porcentaje de población en pobreza**: ante un aumento del 1% de población en pobreza dentro de un condado se tiene un aumento en la tasa de asesinatos del 1.25%.
- **Porcentaje de población que no habla bien inglés**: ante un aumento del 1% de población que no habla bien inglés dentro de un condado se tiene un aumento en la tasa de asesinatos de 1.6%.
- **Porcentaje de población que vive donde nació**: ante un aumento del 1% de población que vive en el mismo estado donde nació dentro de un condado se tiene una disminución en la tasa de asesinatos de 0.31%.
- **Porcentaje de población que no acabó la preparatoria**: ante un aumento del 1% de población que no acabó la preparatoria dentro de un condado se tiene un aumento en la tasa de asesinatos de 1.08%.
- **Porcentaje de madres con hijos menores a 18 años que trabajan**: ante un au-

mento del 1% de población afroamericana dentro de un condado se tiene una disminución en la tasa de asesinatos de 2.15%

- **Porcentaje de población inmigrante que llegaron en los últimos 10 años**: ante un aumento del 1% de población migrante que llegaron en los últimos 10 años dentro de un condado se tiene un aumento en la tasa de asesinatos de 0.87%.

5 Conclusiones

En este trabajo se presentan 4 tipos de modelos: en el primero, un modelo muy sencillo del cual no se pueden obtener conclusiones distintas a lo que haría un resumen de una tabla, sin embargo, funcionó para elegir la liga y la verosimilitud; el segundo, sirvió para agregar un nivel a la jerarquía y ver que era el camino correcto para mejorar el modelo; el tercer modelo, nos sirivió para observar efectos de las covariables que se usaron; el último modelo, nos sirvió para ver si mayor complejidad dentro de los efectos de las covariables mejoraba la forma de explicar la tasa de asesinatos en EUA.

Un hallazgo importante fue que a pesar de que el modelo 4 mostraba un mejor desempeño ante la métrica de WAIC, los resultados generados por este modelo pueden ser considerados como ruido estadístico y por lo tanto no poder concluir nada. Es por esto que se seleccionó un modelo más simple, pero con mayor interpretación.

Finalmente, el trabajo presente podría mejorarse si se agregan más covariables, si se usa una incial equivalente al problema de Ridge o se usa un modelo Poisson inflado hacia el 0. Sin embargo, se consideró que se tiene un modelo que explica las tasas de asesinatos en EUA y que además controla por divisones y estados dentro del país.

6 Código

6.1 Prueba_modelo_inicial.R

```
# Funciones ayuda ----------------------------------------------------------

predice <- function(modelo, nombre, liga){
  y <- tibble(
    media = extract(modelo, "yn")$yn %>%
      apply(
        MARGIN = 2,
        FUN    = mean
      ),
    mediana = extract(modelo, "yn")$yn %>%
      apply(
        MARGIN = 2,
        FUN    = median
      ),
    modelo  = nombre,
    liga    = liga,
    muertes =  x$murders
  )
  return(y)
}

# Lectura de datos --------------------------------------------------------
library(dplyr)
library(readr)
library(rstan)
library(bayesplot)
library(loo)
library(purrr)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())

df <- read_csv(
  "https://archive.ics.uci.edu/ml/machine-learning-databases/00211/CommViolPredUnnormali
  col_names = FALSE,
  na = "?"
)

names(df) <- read_table(
  here::here("Datos/nombres.txt"),
  col_names = FALSE
```

```
) %>%
  mutate(
    var_names = gsub(
      "(.*) (.*)",
      "\\1",
      X2
    )
  ) %>%
  pull(var_names) %>%
  make.names()

estados <- read_csv(
  here::here("Datos/estados_regiones")
) %>%
  select(`State Code`, Division) %>%
  rename(State = `State Code`)


x <- df %>%
  left_join(estados, by = "State") %>%
  mutate(
    State = State %>% as.factor,
    Division = Division %>% as.factor
  ) %>%
  select(State, murders, pop, Division) %>%
  na.omit()

# Modelo Stan -------------------------------------------------------------

intercepto_variante <- stan_model(
  here::here("Modelos Stan/modelo1.stan")
)

# Distribución Binomial ---------------------------------------------------

bin_logit <- sampling(
  intercepto_variante,
  data = list(
    N            = nrow(x),
    y            = x$murders,
    n            = x$pop,
    division     = as.numeric(x$Division),
    division_no  = 9L,
    liga         = 1L, # Liga Logit
    distribucion = 1L  # Distribución binomial
```

```
  ),
  warmup = 500,
    seed = 984984
)

bin_probit <- sampling(
  intercepto_variante,
  data = list(
    N            = nrow(x),
    y            = x$murders,
    n            = x$pop,
    division     = as.numeric(x$Division),
    division_no  = 9L,
    liga         = 2L, # Liga Probit
    distribucion = 1L  # Distribución binomial
  ),
  warmup = 500,
    seed = 984984
)

bin_cloglog <- sampling(
  intercepto_variante,
  data = list(
    N            = nrow(x),
    y            = x$murders,
    n            = x$pop,
    division     = as.numeric(x$Division),
    division_no  = 9L,
    liga         = 3L, # Liga Complementaria log-log
    distribucion = 1L  # Distribución binomial
  ),
  warmup = 500,
    seed = 984984
)

bin_loglog <- sampling(
  intercepto_variante,
  data = list(
    N            = nrow(x),
    y            = x$murders,
    n            = x$pop,
    division     = as.numeric(x$Division),
    division_no  = 9L,
    liga         = 4L, # Liga log-log
    distribucion = 1L  # Distribución binomial
```

```
  ),
  warmup = 500,
    seed = 984984
)

# Distribución Poisson ----------------------------------------------------

pois <- sampling(
  intercepto_variante,
  data = list(
    N            = nrow(x),
    y            = x$murders,
    n            = x$pop,
    division     = as.numeric(x$Division),
    division_no  = 9L,
    liga         = 5L, # Liga Exponencial
    distribucion = 2L  # Distribución poisson
  ),
  warmup = 500,
    seed = 984984
)

# Comparación de modelos --------------------------------------------------

predicciones <- predice(
  bin_probit,
  "Binomial",
  "Probit"
) %>%
  full_join(
    predice(bin_logit, "Binomial", "Logit"),
    by = c("media", "mediana", "modelo", "liga", "muertes")
  ) %>%
  full_join(
    predice(bin_loglog, "Binomial", "Log-Log"),
    by = c("media", "mediana", "modelo", "liga", "muertes")
  ) %>%
  full_join(
    predice(bin_cloglog, "Binomial", "CLog-Log"),
    by = c("media", "mediana", "modelo", "liga", "muertes")
  ) %>%
  full_join(
    predice(pois, "Poisson", "Exponencial"),
    by = c("media", "mediana", "modelo", "liga", "muertes")
  )
```

```
predicciones %>%
  ggplot(
    aes(
      x = muertes,
      y = mediana,
      colour = liga
    )
  ) +
  geom_point() +
  facet_wrap(liga~modelo)

performance <- tibble(
  modelo = c(
    bin_cloglog,
    bin_logit,
    bin_loglog,
    bin_probit,
    pois
  )
) %>%
  mutate(
    nombre   = c(rep("Binomial", 4), "Poisson"),
    liga     = c("Cloglog", "Logit", "Loglog", "Probit", "Exponencial"),
    log_lik  = map(modelo, extract_log_lik),
    waic_obj = map(log_lik, waic),
    waic     = map(waic_obj, ~.$waic) %>% flatten_dbl()
  ) %>%
  select(
    nombre, liga, waic
  )

saveRDS(
  performance,
  file = here::here("Resultados/desempeño_primeros_modelos.rds")
)
```

6.2 corre_modelos.R

```
# Lectura de datos ---------------------------------------------------------
library(dplyr)
library(readr)
library(rstan)
library(bayesplot)
```

```
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())

df <- read_csv(
  "https://archive.ics.uci.edu/ml/machine-learning-databases/00211/CommViolPredUnnormali
  col_names = FALSE,
  na = "?"
)

names(df) <- read_table(
  here::here("Datos/nombres.txt"),
  col_names = FALSE
) %>%
  mutate(
    var_names = gsub(
      "(.*) (.*)",
      "\\1",
      X2
    )
  ) %>%
  pull(var_names) %>%
  make.names()

estados <- read_csv(
  here::here("Datos/estados_regiones")
  ) %>%
  select(`State Code`, Division) %>%
  rename(State = `State Code`)


x <- df %>%
  left_join(estados, by = "State") %>%
  mutate(
    State = State %>% as.factor,
    Division = Division %>% as.factor
  ) %>%
  select(
    State,
    murders,
    pop,
    Division,
    pctBlack,
    # pctWhite,
    pctPoverty,
    # pct12.17w2Par,
```

```
    pctNotSpeakEng,
    pctBornStateResid,
    pctNotHSgrad,
    pctWorkMom.18,
    pctFgnImmig.10
    # pctPolicWhite,
    # pctPolicBlack,
    # officDrugUnits
  ) %>%
  na.omit()

division <- x %>%
  group_by(State, Division) %>%
  summarise() %>%
  ungroup %>%
  mutate(
      State = as.numeric(State),
      Division = as.numeric(Division)
  ) %>%
  pull(Division)

xcov <- x %>%
  select(-c(State, murders, pop, Division)) %>%
  mutate_all(function(x) x /100) %>%
  as.matrix

# Segundo modelo ----------------------------------------------------------

segundo_modelo <- stan_model(
  here::here("Modelos Stan/modelo2.stan")
  )

modelo_dos <- sampling(
  segundo_modelo,
  list(
    N           = nrow(x),
    y           = x$murders,
    n           = x$pop,
    L           = 48,
    state       = as.numeric(x$State),
    division    = division,
    division_no = 9
  ),
  warmup = 500,
  seed = 123456,
```

```
  chains = 4,
  thin = 2,
  control = list(
    adapt_delta = 0.85
  )
)

saveRDS(
  modelo_dos,
  file = here::here("Resultados/modelo_dos.rds")
)


# Tercer modelo -----------------------------------------------------------

tercer_modelo <- stan_model(
  here::here("Modelos Stan/modelo3.stan")
  )

modelo_tres <- sampling(
  tercer_modelo,
  list(
    N           = nrow(x),
    y           = x$murders,
    n           = x$pop,
    L           = 48,
    state       = as.numeric(x$State),
    division    = division,
    division_no = 9,
    P = ncol(xcov),
    X = as.matrix(xcov)
  ),
  warmup = 500,
  iter = 3000,
  seed = 123456,
  chains = 4,
  thin = 4,
  control = list(
    adapt_delta = 0.85
  )
)

saveRDS(
  modelo_tres,
  file = here::here("Resultados/modelo_tres.rds")
```

```
)

# Cuarto Modelo ------------------------------------------------------------

cuarto_modelo <- stan_model(
  here::here("Modelos Stan/modelo4.stan")
)

modelo_cuatro <- sampling(
  cuarto_modelo,
  list(
    N           = nrow(x),
    y           = x$murders,
    n           = x$pop,
    L           = 48L,
    state       = as.numeric(x$State),
    division    = division,
    division_no = 9L,
    P = ncol(xcov),
    X = as.matrix(xcov)
  ),
  warmup = 500,
  iter = 3000,
  seed = 123456,
  chains = 4,
  thin = 4,
  control = list(
    adapt_delta = 0.85
  )
)

saveRDS(
  modelo_cuatro,
  file = here::here("Resultados/modelo_cuatro.rds")
)
```

6.3 analisis_modelos.R

```
# Lectura de datos --------------------------------------------------------

library(bayesplot)
library(rstan)
library(dplyr)
library(ggplot2)
```

```r
library(readr)
library(loo)
library(purrr)
ggplot2::theme_set(theme_bw())

predice <- function(modelo){
  y <- tibble(
    media = extract(modelo, "yn")$yn %>%
      apply(
        MARGIN = 2,
        FUN    = mean
      ),
    mediana = extract(modelo, "yn")$yn %>%
      apply(
        MARGIN = 2,
        FUN    = median
      ),
    intalto = extract(modelo, "yn")$yn %>%
      apply(
        MARGIN = 2,
        FUN    = quantile,
        probs = 0.975
      ),
    intbajo = extract(modelo, "yn")$yn %>%
      apply(
        MARGIN = 2,
        FUN    = quantile,
        probs = 0.025
      ),
    muertes =  x$murders
  )
  return(y)
}

df <- read_csv(
  "https://archive.ics.uci.edu/ml/machine-learning-databases/00211/CommViolPredUnnormali
  col_names = FALSE,
  na = "?"
)

names(df) <- read_table(
  here::here("Datos/nombres.txt"),
  col_names = FALSE
) %>%
  mutate(
```

```
    var_names = gsub(
      "(.*) (.*)",
      "\\1",
      X2
    )
  ) %>%
  pull(var_names) %>%
  make.names()

estados <- read_csv(
  here::here("Datos/estados_regiones")
) %>%
  select(`State Code`, Division, State) %>%
  rename(
    Name  = State,
    State = `State Code`
    )


x <- df %>%
  left_join(estados, by = "State") %>%
  mutate(
    State = State %>% as.factor,
    Division = Division %>% as.factor
  ) %>%
  select(
    State,
    murders,
    pop,
    Division,
    pctBlack,
    # pctWhite,
    pctPoverty,
    # pct12.17w2Par,
    pctNotSpeakEng,
    pctBornStateResid,
    pctNotHSgrad,
    pctWorkMom.18,
    pctFgnImmig.10
    # pctPolicWhite,
    # pctPolicBlack,
    # officDrugUnits
  ) %>%
  na.omit()
```

```
division <- x %>%
  group_by(State, Division) %>%
  summarise() %>%
  ungroup %>%
  mutate(
    State = as.numeric(State),
    Division = as.numeric(Division)
  ) %>%
  pull(Division)

# Modelo dos ------------------------------------------------------------

modelo_dos <- readRDS(
  here::here("Resultados/modelo_dos.rds")
)

mod2_predvsval <- predice(modelo_dos) %>%
  filter(muertes < 250) %>%
  ggplot(aes(x = muertes, y = mediana)) +
  geom_errorbar(aes(ymin = intbajo, ymax = intalto)) +
  geom_point() +
  geom_line(aes(x = seq(1, 250, length.out = 2205), y = seq(1, 250, length.out = 2205)))
  labs(
    title = "Comparación entre valores reales y predicciones del modelo",
    x = "Número de Muertes",
    y = "Predictiva Posterior",
    subtitle = paste(
      "WAIC:\t",
      modelo_dos %>%
        extract_log_lik() %>%
        waic() %>%
        .$waic %>%
        round(2)
    )
  )
saveRDS(mod2_predvsval, file = here::here("Resultados/mod2_predvsval.rds"))

beta0_m1    <- extract(modelo_dos, pars = "beta0")$beta0

beta0_df_m1 <- tibble(
  media   = apply(beta0_m1, MARGIN = 2, FUN = mean),
  mediana = apply(beta0_m1, MARGIN = 2, FUN = median),
  int_baj = apply(beta0_m1, MARGIN = 2, FUN = quantile, probs = 0.025),
  int_al  = apply(beta0_m1, MARGIN = 2, FUN = quantile, probs = 0.975),
  ymin    = apply(beta0_m1, MARGIN = 2, FUN = min),
```

```
  ymax    = apply(beta0_m1, MARGIN = 2, FUN = max),
  state   = filter(estados, State %in% levels(x$State)) %>% pull(Name)
) %>%
  mutate_if(is.numeric, ~(exp(.) * 100))

ggplot(
  data = beta0_df_m1,
  aes(
    y = forcats::fct_rev(reorder(state, state))
  )
) +
  geom_errorbarh(
    aes(
      xmin   = int_baj,
      xmax   = int_al
    )
  ) +
  geom_point(aes(x = mediana)) +
  labs(
    y        = "Estado",
    x        = "Tasa de asesinatos",
    title    = "Tasa de asesinato por estado",
    subtitle = "Efectos por estado"
  )

theta_m1    <- extract(modelo_dos, pars = "theta")$theta

theta_df_m1 <- tibble(
  media   = apply(theta_m1, MARGIN = 2, FUN = mean),
  mediana = apply(theta_m1, MARGIN = 2, FUN = median),
  int_baj = apply(theta_m1, MARGIN = 2, FUN = quantile, probs = 0.025),
  int_al  = apply(theta_m1, MARGIN = 2, FUN = quantile, probs = 0.975),
  ymin    = apply(theta_m1, MARGIN = 2, FUN = min),
  ymax    = apply(theta_m1, MARGIN = 2, FUN = max),
  div     = levels(x$Division)
) %>%
  mutate_if(is.numeric, ~(exp(.) * 100))

phi_m1 <- extract(modelo_dos, pars = "phi_param")$phi_param
phi_df_m1 <- tibble(
  media = mean(phi_m1),
  mediana = median(phi_m1),
  int_baj = quantile(phi_m1, probs = 0.025),
  int_al  = quantile(phi_m1, probs = 0.975),
  div = "EUA"
```

```
) %>%
  mutate_if(is.numeric, ~(exp(.) * 100))

ggplot(
  data = full_join(
    theta_df_m1,
    phi_df_m1,
    by = c("media", "mediana", "int_baj", "int_al", "div")
  ) %>%
    mutate(
      div = factor(div, levels =
                     c(
                       "EUA",
                       "West South Central",
                       "West North Central",
                       "South Atlantic",
                       "Pacific",
                       "New England",
                       "Mountain",
                       "Middle Atlantic",
                       "East South Central",
                       "East North Central"
                     )
      )
    ),
  aes(
    x = div
  )
) +
  geom_errorbar(
    aes(
      ymin   = int_baj,
      ymax   = int_al
    )
  ) +
  geom_point(aes(y = mediana)) +
  coord_flip() +
  theme_bw() +
  labs(
    x        = "División",
    y        = "Tasa de asesinatos",
    title    = "Tasa de asesinato por división",
    subtitle = "Efectos por división"
  )
```

```
# Modelo Tres -------------------------------------------------------------

modelo_tres <- readRDS(
  here::here("Resultados/modelo_tres.rds")
)


mod3_predvsval <- predice(modelo_tres) %>%
  filter(muertes < 250) %>%
  ggplot(aes(x = muertes, y = mediana)) +
  geom_errorbar(aes(ymin = intbajo, ymax = intalto)) +
  geom_point() +
  geom_line(aes(x = seq(1, 250, length.out = 2205), y = seq(1, 250, length.out = 2205)))
  labs(
    title = "Comparación entre valores reales y predicciones del modelo",
    x = "Número de Muertes",
    y = "Predictiva Posterior",
    subtitle = paste(
      "WAIC:\t",
      modelo_tres %>%
        extract_log_lik() %>%
        waic() %>%
        .$waic %>%
        round(2)
    )
  )
saveRDS(mod3_predvsval, file = here::here("Resultados/mod3_predvsval.rds"))


beta0_m3    <- extract(modelo_tres, pars = "beta0")$beta0

beta0_df_m3 <- tibble(
  media   = apply(beta0_m3, MARGIN = 2, FUN = mean),
  mediana = apply(beta0_m3, MARGIN = 2, FUN = median),
  int_baj = apply(beta0_m3, MARGIN = 2, FUN = quantile, probs = 0.025),
  int_al  = apply(beta0_m3, MARGIN = 2, FUN = quantile, probs = 0.975),
  ymin    = apply(beta0_m3, MARGIN = 2, FUN = min),
  ymax    = apply(beta0_m3, MARGIN = 2, FUN = max),
  state   = filter(estados, State %in% levels(x$State)) %>% pull(Name)
) %>%
  mutate_if(is.numeric, ~exp(.) * 100000)

beta0_graph <- ggplot(
  data = beta0_df_m3,
```

```
  aes(
    x = forcats::fct_rev(reorder(state, state))
  )
) +
  geom_errorbar(
    aes(
      ymin   = int_baj,
      ymax   = int_al
    )
  ) +
  geom_point(aes(y = mediana)) +
  coord_flip() +
  theme_bw() +
  labs(
    x        = "Estado",
    y        = "Tasa de asesinatos por cada 100,000 habitantes",
    title    = "Tasa de asesinatos base por estado por cada 100,000 habitantes"
  )
saveRDS(beta0_graph, here::here("Resultados/mod3_tasabase.rds"))
theta_m3    <- extract(modelo_tres, pars = "theta")$theta

theta_df_m3 <- tibble(
  media   = apply(theta_m3, MARGIN = 2, FUN = mean),
  mediana = apply(theta_m3, MARGIN = 2, FUN = median),
  int_baj = apply(theta_m3, MARGIN = 2, FUN = quantile, probs = 0.025),
  int_al  = apply(theta_m3, MARGIN = 2, FUN = quantile, probs = 0.975),
  ymin    = apply(theta_m3, MARGIN = 2, FUN = min),
  ymax    = apply(theta_m3, MARGIN = 2, FUN = max),
  div     = levels(x$Division)
) %>%
  mutate_if(is.numeric, ~exp(.) * 100000)

phi_m3 <- extract(modelo_tres, pars = "phi_param")$phi_param
phi_df_m3 <- tibble(
  media = mean(phi_m3),
  mediana = median(phi_m3),
  int_baj = quantile(phi_m3, probs = 0.025),
  int_al  = quantile(phi_m3, probs = 0.975),
  div = "EUA"
) %>%
  mutate_if(is.numeric, ~exp(.) * 100000)

mod3_tasabase_div <- ggplot(
  data = theta_df_m3 %>%
    full_join(phi_df_m3) %>%
```

```
    mutate(
      div = factor(div, levels =
                     c(
                       "EUA",
                       "West South Central",
                       "West North Central",
                       "South Atlantic",
                       "Pacific",
                       "New England",
                       "Mountain",
                       "Middle Atlantic",
                       "East South Central",
                       "East North Central"
                     )
      )
    ),
  aes(
    x = div
  )

) +
  geom_errorbar(
    aes(
      ymin   = int_baj,
      ymax   = int_al
    )
  ) +
  geom_point(aes(y = mediana)) +
  coord_flip() +
  theme_bw() +
  labs(
    x        = "Estado",
    y        = "Tasa base de asesinatos por cada 100,000 habitantes",
    title    = "Tasa base de asesinato por División por cada 100,000 habitantes"
  )

saveRDS(mod3_tasabase_div, here::here("Resultados/mod3_tasabase_div.rds"))

beta <- extract(modelo_tres, pars = "beta")$beta
cov_names <- x %>%
  select(-c(State, murders, pop, Division)) %>%
  names
beta_df <- tibble(
  media   = apply(beta, MARGIN = 2, FUN = mean),
  mediana = apply(beta, MARGIN = 2, FUN = median),
```

```
  int_baj = apply(beta, MARGIN = 2, FUN = quantile, probs = 0.025),
  int_al  = apply(beta, MARGIN = 2, FUN = quantile, probs = 0.975),
  ymin    = apply(beta, MARGIN = 2, FUN = min),
  ymax    = apply(beta, MARGIN = 2, FUN = max),
  var     = cov_names,
  nombres = case_when(
    var == "pctBlack"           ~ "Porcentaje de población \nafroamericana",
    var == "pctPoverty"         ~ "Porcentaje de población \nen pobreza",
    var == "pctNotSpeakEng"     ~ "Porcentaje de población \nque no habla bien inglés",
    var == "pctBornStateResid"  ~ "Porcentaje de población \nque vive donde nacio",
    var == "pctNotHSgrad"       ~ "Porcentaje de población \nque no acabó la preparatori
    var == "pctWorkMom.18"      ~ "Porcentaje de madres\n con hijos menores a 18 años \n
    var == "pctFgnImmig.10"     ~ "Porcentaje de población \nde inmigrantes que \nllegar
  )
)


p <- ggplot(
  data = beta_df,
  aes(
    x = nombres
  )
) +
  geom_errorbar(
    aes(
      ymin   = int_baj,
      ymax   = int_al
    ),
    stat = "identity"
  ) +
  geom_point(aes(y = mediana)) +
  theme_bw() +
  coord_flip() +
  labs(
    title = "Parámetros asociados a cada variable",
    subtitle = "Intervalos al 95% de credibilidad",
    x = "",
    y = ""
  )

saveRDS(p, file = here::here("Resultados/mod3_efectos.rds"))

# Modelo cuatro ------------------------------------------------------------

modelo_cuatro <- readRDS(
```

```
  here::here("Resultados/modelo_cuatro.rds")
)

mod4_predvsval <- predice(modelo_cuatro) %>%
  filter(muertes < 250) %>%
  ggplot(aes(x = muertes, y = mediana)) +
  geom_errorbar(aes(ymin = intbajo, ymax = intalto)) +
  geom_point() +
  geom_line(aes(x = seq(1, 250, length.out = 2205), y = seq(1, 250, length.out = 2205)))
  labs(
    title = "Comparación entre valores reales y predicciones del modelo",
    x = "Número de Muertes",
    y = "Predictiva Posterior",
    subtitle = paste(
      "WAIC:\t",
      modelo_cuatro %>%
        extract_log_lik() %>%
        waic() %>%
        .$waic %>%
        round(2)
    )
  )
saveRDS(mod4_predvsval, file = here::here("Resultados/mod4_predvsval.rds"))


modelo_cuatro %>%
  extract_log_lik() %>%
  waic() %>%
  .$waic

beta0    <- extract(modelo_cuatro, pars = "beta0")$beta0

beta0_df <- tibble(
  media   = apply(beta0, MARGIN = 2, FUN = mean),
  mediana = apply(beta0, MARGIN = 2, FUN = median),
  int_baj = apply(beta0, MARGIN = 2, FUN = quantile, probs = 0.025),
  int_al  = apply(beta0, MARGIN = 2, FUN = quantile, probs = 0.975),
  ymin    = apply(beta0, MARGIN = 2, FUN = min),
  ymax    = apply(beta0, MARGIN = 2, FUN = max),
  state   = levels(x$State)
)

ggplot(
  data = beta0_df,
  aes(
```

```
    x = state
  )
) +
  geom_errorbar(
    aes(
      ymin   = int_baj,
      ymax   = int_al
    ),
    stat = "identity"
  ) +
  geom_point(aes(y = mediana)) +
  coord_flip() +
  theme_bw() +
  labs(
    x        = "Probabilidad",
    y        = "Estado",
    title    = "Probabilidad base de asesinato",
    subtitle = "Efectos por estado"
  )

theta0    <- extract(modelo_cuatro, pars = "theta0")$theta0

theta0_df <- tibble(
  media   = apply(theta0, MARGIN = 2, FUN = mean),
  mediana = apply(theta0, MARGIN = 2, FUN = median),
  int_baj = apply(theta0, MARGIN = 2, FUN = quantile, probs = 0.025),
  int_al  = apply(theta0, MARGIN = 2, FUN = quantile, probs = 0.975),
  ymin    = apply(theta0, MARGIN = 2, FUN = min),
  ymax    = apply(theta0, MARGIN = 2, FUN = max),
  div     = levels(x$Division)
)

phi <- extract(modelo_cuatro, pars = "phi_param")$phi_param
phi_df <- tibble(
  media = mean(phi),
  mediana = median(phi),
  int_baj = quantile(phi, probs = 0.025),
  int_al  = quantile(phi, probs = 0.975),
  div = "EUA"
)

ggplot(
  data = theta0_df %>%
    full_join(phi_df),
  aes(
```

```
    x = div
  )
) +
  geom_errorbar(
    aes(
      ymin   = int_baj,
      ymax   = int_al
    ),
    stat = "identity"
  ) +
  geom_point(aes(y = mediana)) +
  geom_hline(aes(yintercept = phi_df$int_baj), colour = "blue") +
  geom_hline(aes(yintercept = phi_df$int_al), colour = "blue") +
  coord_flip() +
  theme_bw() +
  labs(
    x        = "Probabilidad",
    y        = "Estado",
    title    = "Probabilidad base de asesinato",
    subtitle = "Efectos por división"
  )

beta <- extract(modelo_cuatro, pars = "beta")$beta
cov_names <- x %>%
  select(-c(State, murders, pop, Division)) %>%
  names

map(
  1:7,
  function(i){
    beta_df <- tibble(
      media   = apply(beta[, , i], MARGIN = 2, FUN = mean),
      mediana = apply(beta[, , i], MARGIN = 2, FUN = median),
      int_baj = apply(beta[, , i], MARGIN = 2, FUN = quantile, probs = 0.025),
      int_al  = apply(beta[, , i], MARGIN = 2, FUN = quantile, probs = 0.975),
      ymin    = apply(beta[, , i], MARGIN = 2, FUN = min),
      ymax    = apply(beta[, , i], MARGIN = 2, FUN = max),
      var     = levels(x$State)
    )
    p <- ggplot(
      data = beta_df,
      aes(
        x = var
      )
    ) +
```

```
      geom_errorbar(
        aes(
          ymin   = int_baj,
          ymax   = int_al
        )
      ) +
      geom_point(aes(y = mediana)) +
      coord_flip() +
      theme_bw() +
      labs(
        title = paste("Efecto por covariable:", cov_names[i]),
        x = ""
      )
    print(p)
  }
)


theta <- extract(modelo_cuatro, pars = "theta")$theta

map(
  1:7,
  function(i){
    theta_df <- tibble(
      media   = apply(theta[, , i], MARGIN = 2, FUN = mean),
      mediana = apply(theta[, , i], MARGIN = 2, FUN = median),
      int_baj = apply(theta[, , i], MARGIN = 2, FUN = quantile, probs = 0.025),
      int_al  = apply(theta[, , i], MARGIN = 2, FUN = quantile, probs = 0.975),
      var     = levels(x$Division)
    )
    p <- ggplot(
      data = theta_df,
      aes(
        x = var
      )
    ) +
      geom_errorbar(
        aes(
          ymin   = int_baj,
          ymax   = int_al
        )
      ) +
      geom_point(aes(y = mediana)) +
      coord_flip() +
      theme_bw() +
```

```
      labs(
        title = paste("Efecto por covariable:", cov_names[i]),
        x = ""
      )
    print(p)
  }
)


cov_hiper <- rstan::extract(modelo_cuatro, pars = "cov_hiper")$cov_hiper

cov_hiper_df <- tibble(
  media   = apply(cov_hiper, MARGIN = 2, FUN = mean),
  mediana = apply(cov_hiper, MARGIN = 2, FUN = median),
  int_baj = apply(cov_hiper, MARGIN = 2, FUN = quantile, probs = 0.025),
  int_al  = apply(cov_hiper, MARGIN = 2, FUN = quantile, probs = 0.975),
  ymin    = apply(cov_hiper, MARGIN = 2, FUN = min),
  ymax    = apply(cov_hiper, MARGIN = 2, FUN = max),
  var     = cov_names,
  nombres = case_when(
    var == "pctBlack"           ~ "Porcentaje de población \nafroamericana",
    var == "pctPoverty"         ~ "Porcentaje de población \nen pobreza",
    var == "pctNotSpeakEng"     ~ "Porcentaje de población \nque no habla bien inglés",
    var == "pctBornStateResid"  ~ "Porcentaje de población \nque vive donde nacio",
    var == "pctNotHSgrad"       ~ "Porcentaje de población \nque no acabó la preparatori
    var == "pctWorkMom.18"      ~ "Porcentaje de madres\n con hijos menores a 18 años \n
    var == "pctFgnImmig.10"     ~ "Porcentaje de población \nde inmigrantes que \nllegar
  )
)


p <- ggplot(
  data = cov_hiper_df,
  aes(
    x = nombres
  )
) +
  geom_errorbar(
    aes(
      ymin   = int_baj,
      ymax   = int_al
    ),
    stat = "identity"
  ) +
  geom_point(aes(y = mediana)) +
```

```
  theme_bw() +
  coord_flip() +
  labs(
    title = "Hiperparámetros de EUA asociados a cada variable",
    subtitle = "Intervalos al 95% de credibilidad",
    x = "",
    y = ""
  )

saveRDS(p, file = here::here("Resultados/mod4_efectos.rds"))
```

6.4 modelo1.stan

```
data {
  int<lower=0> N;  // Tamaño de los datos
  int y[N];        // Número de asesinados
  int n[N];        // Offset
  int division[N]; // Divisiones
  int division_no; // Número de division
  int liga;
  int distribucion;
}

parameters {
  vector[division_no] theta;
  real phi_param;
  real<lower = 0> lambda;
}

transformed parameters {
  vector[N] prob;
  for(i in 1:N){
    if(liga == 1){
      prob[i] = inv_logit(theta[division[i]]);
    } else if(liga == 2){
      prob[i] = Phi(theta[division[i]]);
    } else if(liga == 3){
      prob[i] = 1 - exp(-exp(theta[division[i]]));
    } else if(liga == 4){
      prob[i] = exp(-exp(theta[division[i]]));
    } else if(liga == 5){
      prob[i] = exp(theta[division[i]]);
    }
  }
```

```
}

model {
  if(distribucion == 1){
      y ~ binomial(n, prob);
  } else {
    for(i in 1:N){
      y[i] ~ poisson(n[i] * prob[i]);
    }
  }
  theta     ~ normal(phi_param, lambda);
  phi_param ~ normal(0, 10);
  lambda    ~ gamma(0.001, 0.001);
}

generated quantities{
  int yn[N];
  vector[N] log_lik;

  if(distribucion == 1){
    yn = binomial_rng(n, prob);
    for(i in 1:N){
      log_lik[i] = binomial_lpmf(y[i] | n[i], prob[i]);
    }
  } else {
    for(i in 1:N){
      yn[i] = poisson_rng(n[i] * prob[i]);
      log_lik[i] = poisson_lpmf(y[i] | n[i] * prob[i]);
    }
  }

}
```

6.5 modelo2.stan

```
data {
  int<lower=0> N;  // Tamaño de los datos
  int y[N];        // Número de asesinados
  int n[N];        // Offset
  int L;           // Número de estados
  int state[N];    // Estados
  int division[L]; // Divisiones
  int division_no; // Número de division
}
```

```
parameters {
  vector[L] beta0;
  vector[division_no] theta;
  vector<lower = 0>[division_no] theta_sd;
  real phi_param;
  real<lower = 0> lambda;
}

transformed parameters {
  vector[N] prob;
  for(i in 1:N){
    prob[i] = exp(beta0[state[i]]);
  }
}

model {
  // Verosimilitud
  for(i in 1:N){
    y[i] ~ poisson(n[i] * prob[i]);
  }
  // Cambio de estado a división
  for(j in 1:L){
    beta0[j] ~ normal(theta[division[j]],theta_sd[division[j]]);
  }
  // Cambio de división  a hiperparámetros
  theta    ~ normal(phi_param, lambda);
  theta_sd ~ gamma(0.001, .001);
  // Priors vagas
  phi_param ~ normal(0, 10);
  lambda    ~ gamma(0.001, 0.001);
}

generated quantities{
  int yn[N];
  vector[N] log_lik;
  for(i in 1:N){
    yn[i]      = poisson_rng(n[i] * prob[i]);
    log_lik[i] = poisson_lpmf(y[i] | n[i] * prob[i]);
  }
}
```

6.6 modelo3.stan

```
data {
  int<lower=0> N;  // Tamaño de los datos
  int y[N];        // Número de asesinados
  int n[N];        // Offset
  int L;           // Número de estados
  int state[N];    // Estados
  int division[L]; // Divisiones
  int division_no; // Número de division
  int P;           // Número de covariables
  matrix[N,P] X;   // Matriz de covariables
}

parameters {
  vector[L] beta0;
  vector[division_no] theta;
  vector<lower = 0>[division_no] theta_sd;
  real phi_param;
  real<lower = 0> lambda;
  vector[P] beta;
}

transformed parameters {
  vector[N] prob;
  for(i in 1:N){
    prob[i] = exp(beta0[state[i]] + row(X, i) * beta);
  }
}

model {
  // Verosimilitud
  for(i in 1:N){
    y[i]~ poisson(n[i] * prob[i]);
  }
  // Cambio de estado a division
  for(j in 1:L){
    beta0[j] ~ normal(
      theta[division[j]],
      theta_sd[division[j]]
      );
  }
  // Cambio de division a hiperparámetros
  theta    ~ normal(phi_param, lambda);
```

```
  // theta_sd ~ gamma(0.001, 0.001);
  // Priors vagas
  phi_param ~ normal(0, 10);
  // lambda    ~ gamma(0.001, 0.001);
  beta      ~ normal(0, 1);
}

generated quantities{
  int yn[N];
  vector[N] log_lik;
  for(i in 1:N){
    yn[i]      = poisson_rng(n[i] * prob[i]);
    log_lik[i] = poisson_lpmf(y[i] | n[i] * prob[i]);
  }
}
```

6.7 modelo4.stan

```
data {
  int<lower=0> N;  // Tamaño de los datos
  int y[N];        // Número de asesinados
  int n[N];        // Offset
  int L;           // Número de estados
  int state[N];    // Estados
  int division[L]; // Divisiones
  int division_no; // Número de division
  int P;           // Número de covariables
  matrix[N,P] X;   // Matriz de covariables
}

parameters {
  vector[L] beta0;
  vector[division_no] theta0;
  vector<lower = 0>[division_no] theta0_sd;
  real phi_param;
  real<lower = 0> lambda;
  matrix[L, P] beta;
  matrix[division_no, P] theta;
  matrix<lower = 0>[division_no, P] theta_sd;
  vector[P] cov_hiper;
  vector<lower = 0>[P] cov_sd_hiper;
}

transformed parameters {
```

```
  vector[N] prob;
  for(i in 1:N){
    prob[i] = exp(beta0[state[i]] + dot_product(row(X, i), row(beta, state[i])));
  }
}

model {
  // Verosimilitud
  for(i in 1:N){
    y[i]~ poisson(n[i] * prob[i]);
  }
  // Cambio de estado a division
  for(j in 1:L){
    beta0[j] ~ normal(
      theta0[division[j]],
      theta0_sd[division[j]]
      );
      for(p in 1:P){
        beta[j, p] ~ normal(
          theta[division[j], p],
          theta_sd[division[j], p]
        );
      }
  }
  // Cambio de division a hiperparámetros
  theta0    ~ normal(phi_param, lambda);
  for(p in 1:P){
    for(i in 1:division_no){
      theta[i, p] ~ normal(cov_hiper[p], cov_sd_hiper[p]);
    }
  }
  // Priors vagas
  phi_param ~ normal(0, 10);
  cov_hiper ~ normal(0, 1);

}

generated quantities{
  int yn[N];
  vector[N] log_lik;
  for(i in 1:N){
    yn[i]      = poisson_rng(n[i] * prob[i]);
    log_lik[i] = poisson_lpmf(y[i] | n[i] * prob[i]);
  }
}
```

Referencias

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Vol. Analytical methods for social research. New York: Cambridge University Press.

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