Estimación de la proporción de pobreza por entidad federativa en México 2022, con el enfoque Bayesiano - García Luis

2022-11-10

Introducción.

En este documento se realizará una estimación de la proporción de población en situación de pobreza para cada entidad federativa de la republica Mexicana en el año 2022.

Para ello, usaremos como marco teorico, las normativas metodologicas para la medicion de la pobreza multidimensional en el pais, que son estipuladas por el Consejo Nacional de Evaluacion de la Politica de Desarrollo Social (CONEVAL) segun sus atribuciones establecidas en la Ley General de Desarrollo Social de 2004.

Recordemos que desde 2009, el CONEVAL adoptó un enfoque multidimensional en la metodología de medición de pobreza, cuyos lineamientos se encuentran en la tercera edición de la Metodología para la medición multidimensional de la pobreza en México. Esta metodología está basada en dos perspectivas que analizan e carácter multidimensional de la pobreza, que son la perspectiva de bienestar y la de derechos. La primera se basa en la premisa de que "el ingreso es fundamental para la adquisición de una variedad de bienes y servicios que son indispensables para la satisfacción de las necesidades esenciales" (CONEVAL,2023) , mientras que la segunda considera el cumplimiento de los derechos sociales establecidos como derechos (garantías) constitucionales, y se concibe a la pobreza como una negación de estos derechos. (CONEVAL,2023)

El CONEVAL, con los resultados de la ENIGH, se dedica a construir indicadores de carencia que son un reflejo de la multidimensionalidad de la pobreza, estipulada tambien en el Art36 de la Ley General de Desarrollo Social. Esos indicadores son los siguintes: plp_e, plp, ic_rezedu, ic_asalud, ic_segsoc, ic_cv, ic_sbv, ic_ali, ic_ali_nc.

Estos indicadores, y la proporcion de personas por entidad federativa encuestadas en la ENIGH que las padecen son las variables independientes que usaremos en el modelo para determinar la proporcion de pobreza de personas encuestadas, por entidad federativa

El objetivo, es identificar las variables mas relevantes dentro de la metodologia de medición multidimensional de la pobreza del Coneval, es decir; encontrar los mejores estimadores de la proporción de pobreza en cada entidad federativa y poder asi tambien determinar los pesos de las carencias mas importantes dentro de la multidimensionalidad de la pobreza.

En este ejercicio práctico, se empleará el enfoque Bayesiano porque, a diferencia del enfoque frecuentista, permite incorporar conocimientos previos y actualizaciones de la información.

El enfoque Bayesiano se basa en el Teorema de Bayes, que ofrece un marco matemático para actualizar la probabilidad de una hipótesis a medida que se obtiene más evidencia en los datos. Esto resulta muy util en situaciones donde la información es incompleta o incierta y se quiere integrar los datos observados y las como creencias apriori.

Datos

Como se menciono anteriormente, las variables explictivas serán las proporciones de población encuestadas en la ENIGH que padecen las sigueintes carencias: plp_e, plp, ic_rezedu, ic_asalud, ic_segsoc, ic_cv, ic_sbv, ic_ali, ic_ali_nc.

El calculo de estas proporciones requirió de un tratamiento a la base de datos original, en el cual se transformaron observaciones individuales de variables dicotomicas, en continuas, al sumar todos los unos y dividir entre el total de observaciones de cada varaible. Donde 1 indica la presencia de la carencia en cuestion.

La base de datos original pertenece al archivo pobreza22.csv y esta disponible en: https://www.coneval.org.mx/Medicion/MP/Paginas/Programas_BD_2022.aspx

La variable dependiente *pobreza* es la proporcion de población en situación de pobreza por entidad federativa en 2022.

La variable independiente *plpe* es la proporción de poblacion con ingresos menor a la linea de pobreza extrema por ingresos.

La variable independiente *plp* es la proporción de poblacion con ingresos menor a la linea de pobreza por ingresos.

La variable independiente $ic - rezedu$ es la proporción de poblacion con carencia por rezago educativo.

La variable independiente $ic - asalud$ es la proporción de poblacion con carencia por acceso a servicios de salud.

La variable independiente $ic - segsoc$ es la proporción de poblacion con carencia por acceso a seguridad social.

La variable independiente$ic - cv$ es la proporción de poblacion con carencia por calidad de vivienda.

La variable independiente $ic - sbv$ es la proporción de poblacion con carencia por servicios de vivienda basica.

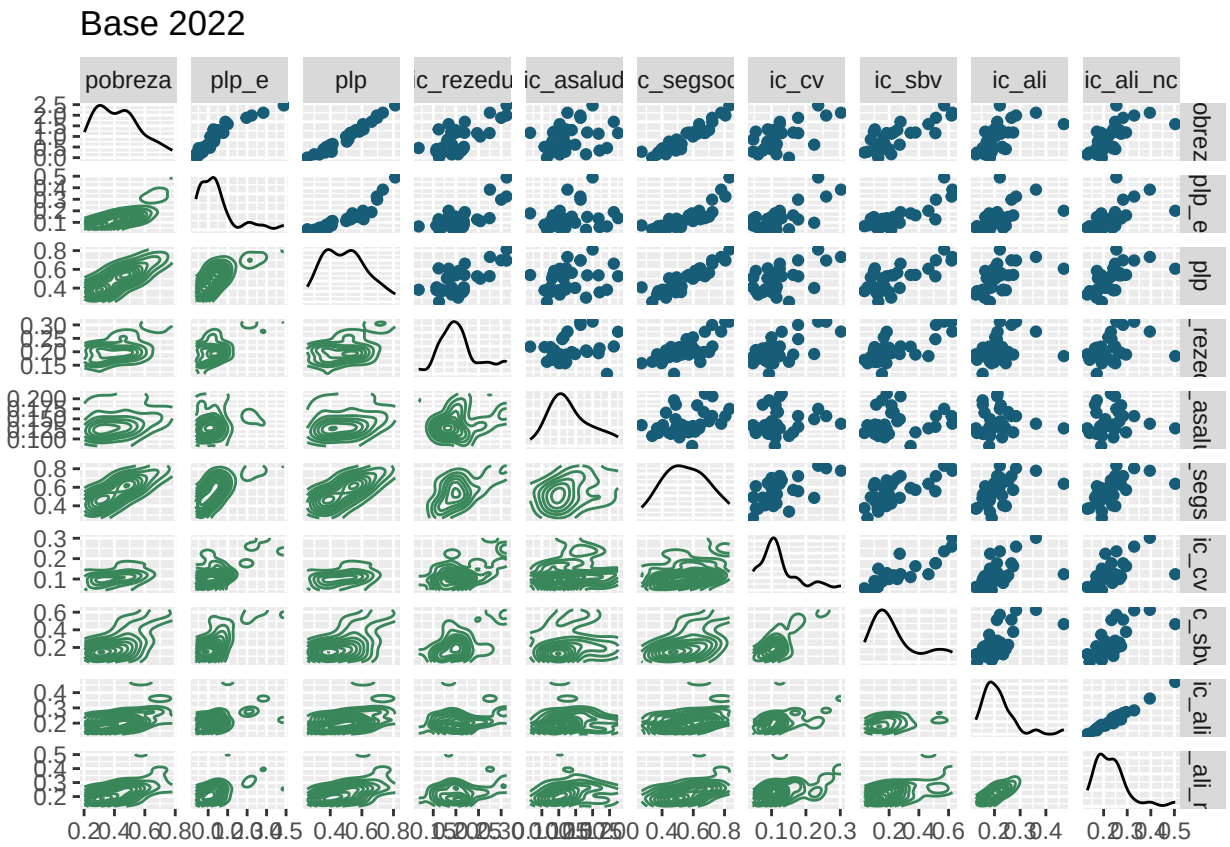
La variable independiente $ic - ali$ es la proporción de poblacion con carencia por alimentacion.

y La variable independiente $ic - ali_nc$ es la proporción de poblacion con carencia por alimentacion nutritiva y de calidad.

A continuación vamos a hacer un pequeño analisis exploratorio de los datos.

```
ggpairs(BP22,
        upper = list(continuous = wrap("points", colour = "#185D77")),
        lower = list(continuous = wrap("density", colour = "#388659")),
        diag = list(continuous = "densityDiag"),
        ggtheme = theme_light() +
          theme(axis.text = element_blank(),
                axis.ticks = element_blank(),
                axis.title = element_blank())) + ggtitle("Base 2022")
```

```
## Warning in warn_if_args_exist(list(...)): Extra arguments: "ggtheme" are being
## ignored.  If these are meant to be aesthetics, submit them using the 'mapping'
## variable within ggpairs with ggplot2::aes or ggplot2::aes_string.
```



En la diagonal principal del gráfico, vemos estimaciones de densidad de cada variable. Estos gráficos muestran cómo están distribuidos los valores de cada variable. Algunas variables parecen tener una distribución bimodal, mientras que otras tienen una distribución más uniforme o sesgada.

Los gráficos por arriba de la diagonal principal muestran la dispersion y relación entre pares de variables. Si los puntos forman una línea clara, hay una fuerte correlación lineal. Si los puntos están dispersos de forma más uniforme, la correlación es más débil.

En la parte inferior de la diagonal observamos graficos de desidad bivaraida, donde los contornos mas elipticos son de las densidades bivaraidas mayormente correlacionadas, y los grupos fuera de ellos, podrian ser posibles outliers.

Para visualizar outliers usaremos un boxplot.

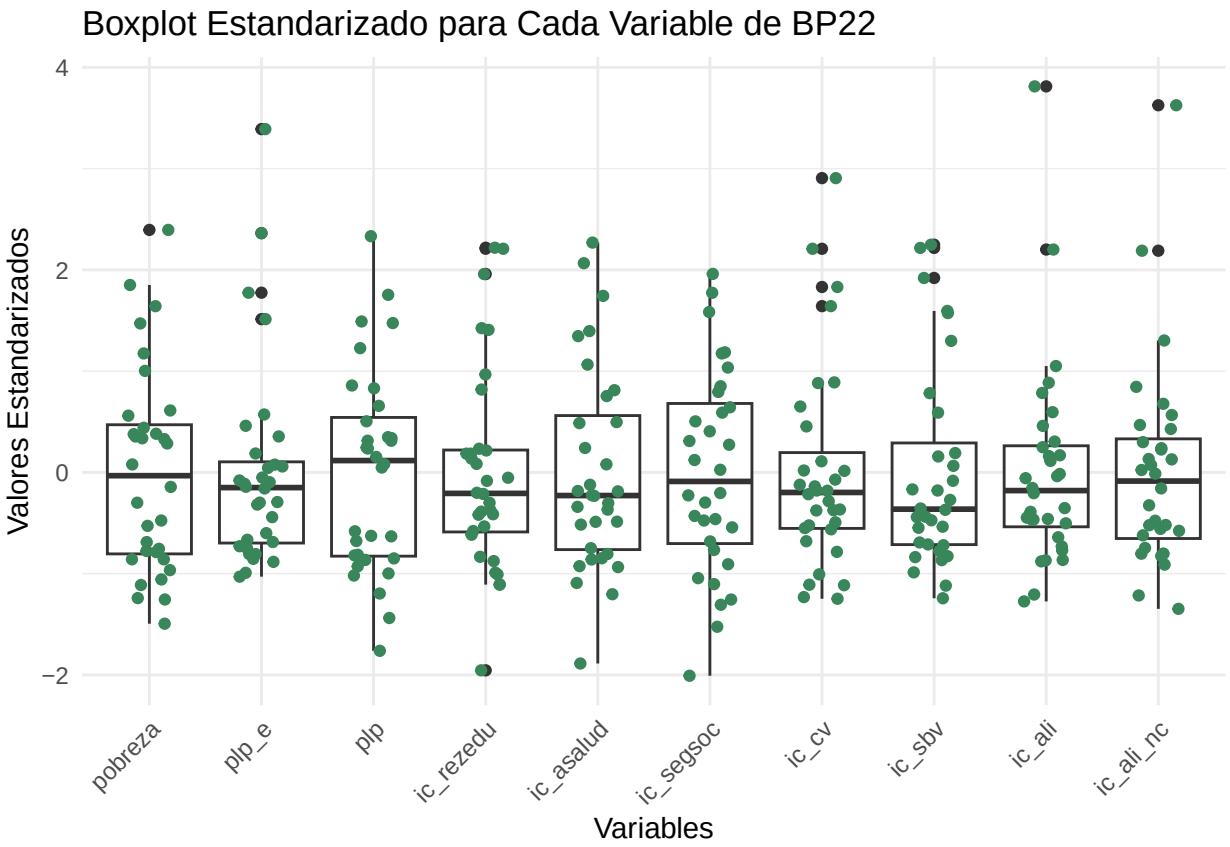
```
BP22_scaled <- BP22 %>%
  mutate(across(everything(), scale))

BP22_long_scaled <- melt(BP22_scaled)
```

```
## No id variables; using all as measure variables
```

```
## Warning: attributes are not identical across measure variables; they will be
## dropped
```

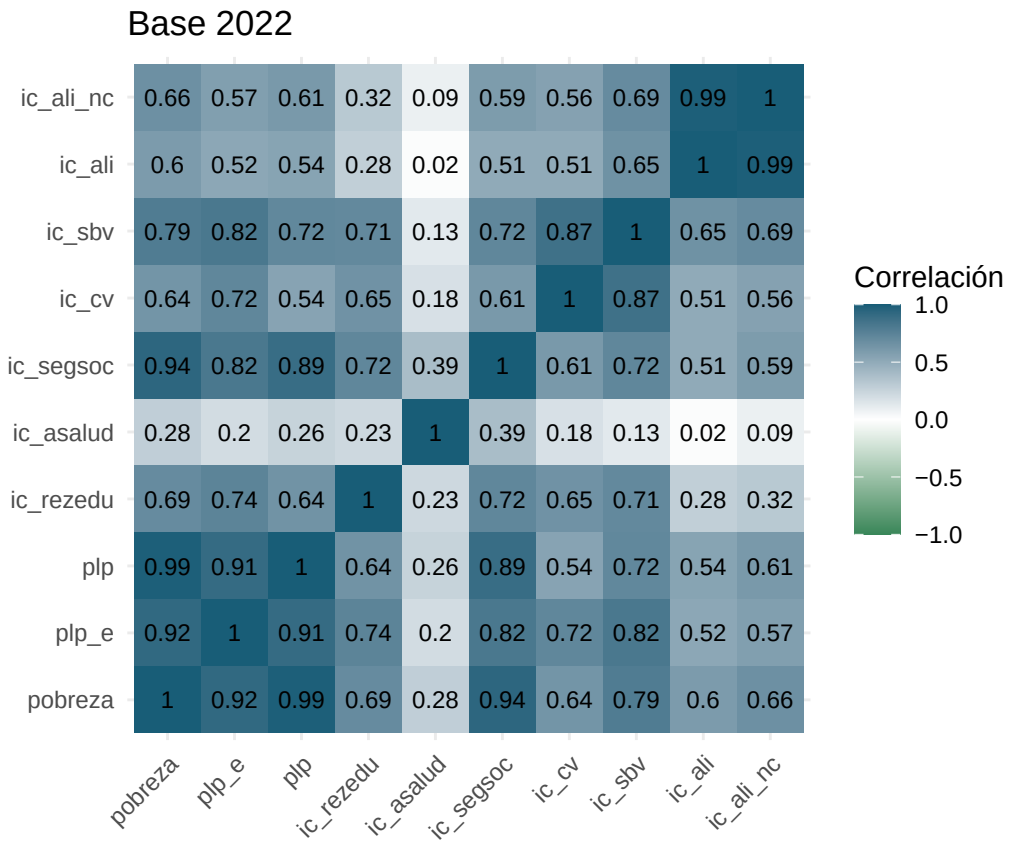
```
ggplot(BP22_long_scaled, aes(x = variable, y = value)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, color = "#388659") +
  theme_minimal() +
  labs(x = "Variables", y = "Valores Estandarizados", title = "Boxplot Estandarizado para Cada Variable de
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



En el grafico podemos visualizar la presencia de outliers en casi todas las varaibles, a exepción de plp, ic_asalud, e ic_segsoc. Aqui cada punto representa una entidad federativa.

```
cor_matriz_22 <- cor(BP22)
cor_larga_22 <- melt(cor_matriz_22)

ggplot(cor_larga_22, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2)), color = "black", size = 3) +
  scale_fill_gradient2(low = "#388659", high = "#185D77", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Correlación") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title = element_blank()) +
  coord_fixed() +
  ggtitle("Base 2022")
```



Con este heatmap de la matriz de correlaciones observamos una asociacion lineal positiva alta entre todas las variables, especialmente nos interesan las mas correlacionadas con la pobreza, que en este caso son plp, plp_e, ic_segsoc, ic_sbv, ic_rezedu, etc.

Una vez realizado el analisis descriptivo pasaremos a especificar el modelo

Planteamiento del modelo econométrico.

Para lo cual se plantea el siguiente modelo econométrico

$$pobreza = b0+b1(plpe)+b2(plp)+b3(ic-rezedu)+b4(ic-asalud)+b5(ic-segsoc)+b6(ic-cv)+b7(ic-sbv)+b8(ic-ali)+b9(icali-n$$

Donde: $pobreza$ es la proporcion de poblacion en situacion de pobreza

$plpe$ es la proporción de poblacion con ingresos menor a la linea de pobreza extrema por ingresos

plp es la proporción de poblacion con ingresos menor a la linea de pobreza por ingresos

$ic - rezedu$ es la proporción de poblacion con carencia por rezago educativo

$ic - asalud$ es la proporción de poblacion con carencia por acceso a servicios de salud

$ic - segsoc$ es la proporción de poblacion con carencia por acceso a seguridad social

$ic - cv$ es la proporción de poblacion con carencia por calidad de vivienda

$ic - sbv$ es la proporción de poblacion con carencia por servicios de vivienda basica

$ic - ali$ es la proporción de poblacion con carencia por alimentacion

$ic - ali_nc$ es la proporción de poblacion con carencia por alimentacion nutritiva y de calidad

Estimación del modelo

Estimación de los parámetros del modelo con todas las variables.

A continuación estimamos el modelo con todas las variables.

```
pob_bas = bas.lm(pobreza ~ ., data = BP22, prior = "BIC",
                 modelprior = uniform(),
                 include.always = ~ .,
                 n.models = 1)

pob_coef <- coef(pob_bas)
pob_coef
```

```
##
##  Marginal Posterior Summaries of Coefficients:
##
##  Using  BMA
##
##  Based on the top  1 models
##            post mean  post SD    post p(B != 0)
## Intercept   0.424676   0.001498   1.000000
## plp_e       0.061355   0.053489   1.000000
## plp         0.709919   0.043177   1.000000
## ic_rezedu  -0.041062   0.064553   1.000000
## ic_asalud   0.015885   0.056644   1.000000
## ic_segsoc   0.251540   0.033358   1.000000
## ic_cv       0.046640   0.060615   1.000000
## ic_sbv      0.078128   0.025617   1.000000
## ic_ali      0.138144   0.198404   1.000000
## ic_ali_nc  -0.052588   0.196279   1.000000
```

Con el cuadro resumen podemos identificar varias coasas.

La primer columna es la media posterior de cada coeficiente. Este es el coeficiente beta asociado a variable, que indica el cambio promedio en la variable pobreza asociado con un aumento de una unidad en la variable independiente, manteniendo constantes las otras variables.

La segunda columna es la desviación estándar posterior de cada coeficiente, que muestra la variabilidad y dispersion en la estimación de ese coeficiente.

La columna p(B != 0) nos indica la probabilidad posterior de que el coeficiente sea diferente de cero. Un valor de 1.000000 indica que, según el modelo, hay una certeza muy alta de que la variable tiene un efecto no nulo en la pobreza.

Pruebas para elegir el mejor modelo.

Para elegir el mejor modelo podemos hacer varias pruebas

```
#Eleccion de modelo

n = nrow(BP22)
pob_lm = lm(pobreza ~ ., data=BP22)
pob_step = step(pob_lm, k=log(n))
```

```
## Start:  AIC=-282.68
## pobreza ~ plp_e + plp + ic_rezedu + ic_asalud + ic_segsoc + ic_cv +
##     ic_sbv + ic_ali + ic_ali_nc
##
##             Df Sum of Sq       RSS     AIC
## - ic_ali_nc  1 0.0000052 0.0015842 -286.04
## - ic_asalud  1 0.0000056 0.0015847 -286.03
## - ic_rezedu  1 0.0000290 0.0016081 -285.56
## - ic_ali     1 0.0000348 0.0016138 -285.44
## - ic_cv      1 0.0000425 0.0016215 -285.29
## - plp_e      1 0.0000944 0.0016735 -284.28
## <none>                   0.0015790 -282.68
## - ic_sbv     1 0.0006676 0.0022467 -274.86
## - ic_segsoc  1 0.0040811 0.0056601 -245.29
## - plp        1 0.0194035 0.0209826 -203.36
##
## Step:  AIC=-286.04
## pobreza ~ plp_e + plp + ic_rezedu + ic_asalud + ic_segsoc + ic_cv +
##     ic_sbv + ic_ali
##
##             Df Sum of Sq       RSS     AIC
## - ic_asalud  1 0.0000042 0.0015884 -289.42
## - ic_rezedu  1 0.0000239 0.0016081 -289.02
## - ic_cv      1 0.0000374 0.0016216 -288.76
## - plp_e      1 0.0001178 0.0017020 -287.21
## <none>                   0.0015842 -286.04
## - ic_ali     1 0.0004712 0.0020554 -281.17
## - ic_sbv     1 0.0006625 0.0022467 -278.32
## - ic_segsoc  1 0.0048533 0.0064375 -244.64
## - plp        1 0.0201585 0.0217427 -205.69
##
## Step:  AIC=-289.42
## pobreza ~ plp_e + plp + ic_rezedu + ic_segsoc + ic_cv + ic_sbv +
##     ic_ali
##
##             Df Sum of Sq       RSS     AIC
## - ic_rezedu  1 0.0000249 0.0016133 -292.39
## - ic_cv      1 0.0000416 0.0016300 -292.06
## - plp_e      1 0.0001157 0.0017041 -290.63
## <none>                   0.0015884 -289.42
## - ic_ali     1 0.0004672 0.0020556 -284.63
## - ic_sbv     1 0.0006607 0.0022491 -281.75
## - ic_segsoc  1 0.0055904 0.0071788 -244.62
## - plp        1 0.0201601 0.0217485 -209.15
##
## Step:  AIC=-292.39
## pobreza ~ plp_e + plp + ic_segsoc + ic_cv + ic_sbv + ic_ali
##
##             Df Sum of Sq       RSS     AIC
## - ic_cv      1 0.0000461 0.0016595 -294.95
## - plp_e      1 0.0000935 0.0017068 -294.05
## <none>                   0.0016133 -292.39
## - ic_ali     1 0.0006205 0.0022338 -285.44
## - ic_sbv     1 0.0006410 0.0022543 -285.15
## - ic_segsoc  1 0.0063783 0.0079916 -244.65
## - plp        1 0.0220927 0.0237060 -209.85
```

```
##
## Step:  AIC=-294.95
## pobreza ~ plp_e + plp + ic_segsoc + ic_sbv + ic_ali
##
##             Df Sum of Sq       RSS     AIC
## - plp_e      1 0.0001831 0.0018426 -295.06
## <none>                   0.0016595 -294.95
## - ic_ali     1 0.0006128 0.0022723 -288.36
## - ic_sbv     1 0.0014661 0.0031256 -278.15
## - ic_segsoc  1 0.0071527 0.0088122 -244.99
## - plp        1 0.0270954 0.0287548 -207.14
##
## Step:  AIC=-295.07
## pobreza ~ plp + ic_segsoc + ic_sbv + ic_ali
##
##             Df Sum of Sq      RSS     AIC
## <none>                   0.001843 -295.06
## - ic_ali     1  0.000490 0.002333 -290.98
## - ic_sbv     1  0.003409 0.005251 -265.02
## - ic_segsoc  1  0.006977 0.008819 -248.43
## - plp        1  0.058771 0.060613 -186.74
```

Con el codigo anterior encontramos combinaciones de varaibles que mejoran los resultados del modelo teniendo como referencia el criterio de informacion de Akaike. En este caso, el mejor modelo es el de menor AIC, que usa las varaibles ic_ali, plp, ic_sbv, e ic_segsoc

Haremos mas pruebas para encontrar el mejor modelo

```
pob_BIC = bas.lm(pobreza ~ ., data = BP22,
                 prior = "BIC", modelprior = uniform())

best = which.max(pob_BIC$logmarg)
bestmodel = pob_BIC$which[[best]]
bestmodel
```

```
## [1] 0 2 5 7 8
```

```
bestgamma = rep(0, pob_BIC$n.vars)
bestgamma[bestmodel + 1] = 1
bestgamma
```

```
##  [1] 1 0 1 0 0 1 0 1 1 0
```

Con esta prueba,buscamos el modelo con el mayor logaritmo del margen de probabilidad, que es una medida de la evidencia a favor de cada modelo considerando los datos observados. Con ello identificamos las variables que deben estar presentes en dicho modelo y visualizamos que son las variables 0 2 5 7 8, que pueden tambien representarse como 1 0 1 0 0 1 0 1 1 0, donde 1 implica la presencia de esa variable en el modelo y 0 su ausencia.

```
pob_bestBIC = bas.lm(pobreza ~ ., data = BP22,
                     prior = "BIC", n.models = 1,
                     bestmodel = bestgamma,
                     modelprior = uniform())

pob_coef = coef(pob_bestBIC)
out = confint(pob_coef)[, 1:2]
```

```
coef_BIC = cbind(pob_coef$postmean, pob_coef$postsd, out)
names = c("post mean", "post sd", colnames(out))
colnames(coef_BIC) = names
coef_BIC
```

```
##           post mean     post sd       2.5%     97.5%
## Intercept 0.4246762 0.001460361 0.42167978 0.4276726
## plp_e     0.0000000 0.000000000 0.00000000 0.0000000
## plp       0.7384756 0.025164694 0.68684188 0.7901093
## ic_rezedu 0.0000000 0.000000000 0.00000000 0.0000000
## ic_asalud 0.0000000 0.000000000 0.00000000 0.0000000
## ic_segsoc 0.2435248 0.024085568 0.19410526 0.2929443
## ic_cv     0.0000000 0.000000000 0.00000000 0.0000000
## ic_sbv    0.1030591 0.014582661 0.07313796 0.1329803
## ic_ali    0.0790553 0.029489858 0.01854711 0.1395635
## ic_ali_nc 0.0000000 0.000000000 0.00000000 0.0000000
```

Con este codigo, usamos el vector de variables que deberian estar presentes, el cual generamos anteriormente, para indicar las variables a usar, de tal modo que el mejor modelo es el que tiene a las variables plp, ic_segsoc, ic_sbv, e ic_ali, ademas de la constante.

Estimación del modelo óptimo elegido.

```
pob_bas2 = bas.lm(pobreza ~ plp + ic_segsoc + ic_sbv + ic_ali,
                  data = BP22, prior = "BIC",
                  modelprior = uniform())

round(summary(pob_bas2), 3)
```

```
##           P(B != 0 | Y) model 1 model 2 model 3 model 4 model 5
## Intercept         1.000   1.000   1.000   1.000   1.000   1.000
## plp               1.000   1.000   1.000   1.000   1.000   1.000
## ic_segsoc         1.000   1.000   1.000   1.000   1.000   0.000
## ic_sbv            1.000   1.000   1.000   0.000   0.000   1.000
## ic_ali            0.885   1.000   0.000   1.000   0.000   0.000
## BF                   NA   1.000   0.130   0.000   0.000   0.000
## PostProbs            NA   0.885   0.115   0.000   0.000   0.000
## R2                   NA   0.997   0.997   0.992   0.988   0.987
## dim                  NA   5.000   4.000   4.000   3.000   3.000
## logmarg              NA  92.081  90.038  77.057  71.619  70.031
```

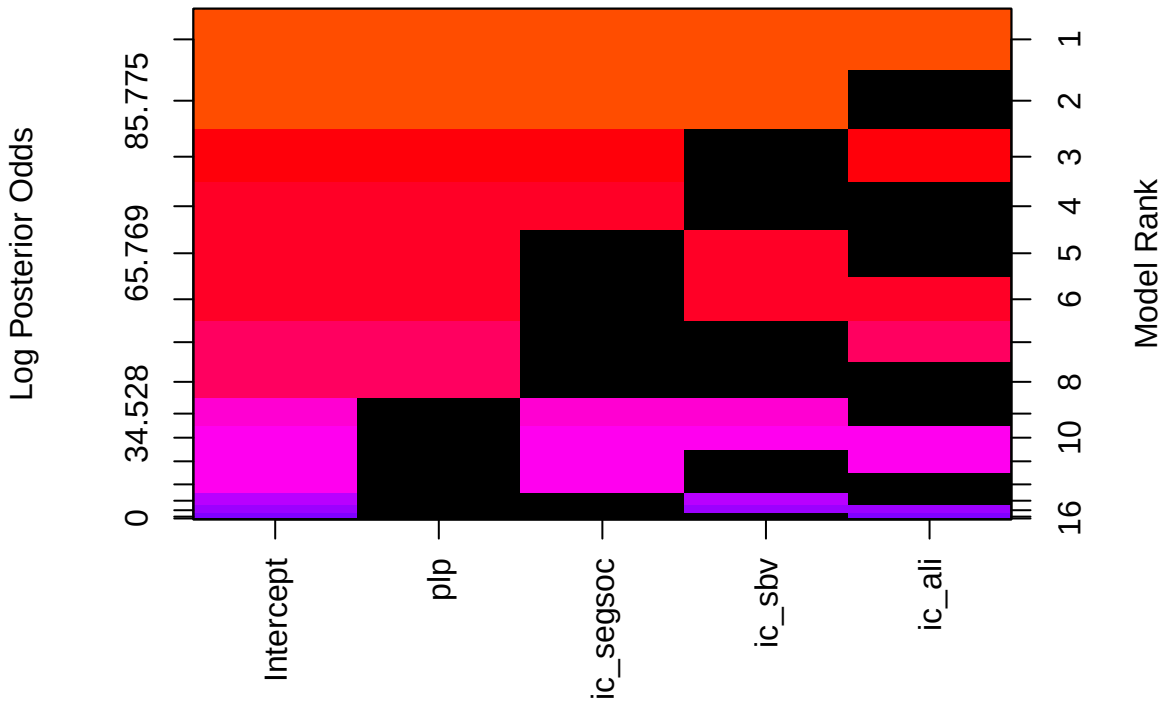
```
print(pob_bas2)
```

```
##
## Call:
## bas.lm(formula = pobreza ~ plp + ic_segsoc + ic_sbv + ic_ali,
##     data = BP22, prior = "BIC", modelprior = uniform())
##
##
##  Marginal Posterior Inclusion Probabilities:
## Intercept        plp  ic_segsoc     ic_sbv     ic_ali
##    1.0000     1.0000     1.0000     1.0000     0.8852
```

Aqui podemos comparar otros modelos alternativos que podriamos usar, comparando su R2 y su logmarg. Ademas las probabilidades de incluir las varaibles en la marginal posterior, es alta en todos los casos.

Tambien podemos visualizar otros posible modelos de la siguiente forma:

```
par(mfrow = c(1, 1))
image(pob_bas2)
```



Finalmente se presentan los resultados de las estimaciones del mejor modelo:

```
pob_coef = coef(pob_bas2)
out = confint(pob_coef)[, 1:2]
coef_BIC = cbind(pob_coef$postmean, pob_coef$postsd, out)
names = c("post mean", "post sd", colnames(out))
colnames(coef_BIC) = names
coef_BIC
```

```
##            post mean    post sd      2.5%     97.5%
## Intercept 0.42467619 0.00147876 0.4216672 0.4277009
## plp       0.73959869 0.02563556 0.6861990 0.7904046
## ic_segsoc 0.24298418 0.02442634 0.1923042 0.2920829
## ic_sbv    0.10509078 0.01561344 0.0727537 0.1361510
## ic_ali    0.06998305 0.03748013 0.0000000 0.1288410
```

1. El intercepto en post mean es de : 0.424676. Este valor representa la proporción estimada de pobreza cuando todas las variables independientes (plp, ic_segsoc, ic_sbv, ic_ali) son iguales a cero. Esto puede interpretarse como el nivel medio de pobreza en México en 2022, dada la ausencia de las otras variables contempladas en el modelo.

2. plp (proporción de la población con ingresos menores a la línea de pobreza por ingresos) en su post mean tiene un valor de: 0.739599. Esto indica que manteniendo constantes las demás variables, un aumento de un punto porcentual en la proporción de la población bajo la línea de pobreza por ingresos está asociado con un aumento de aproximadamente 0.74 puntos porcentuales en la proporción de pobreza total. Su post SD es de 0.025636. Esta es una medida de la incertidumbre o dispersion en la estimación del coeficiente. Mientras que su post p(B != 0): 1.000000. Esto implica una certeza total de que esta variable tiene un efecto significativo sobre la proporción de pobreza.

3.ic_segsoc (proporción de población con carencia por seguridad social): Aqui un aumento de un punto porcentual está asociado con un aumento de 0.242984 puntos porcentuales en la proporción de pobreza.

4.ic_sbv (proporción de población con carencia por servicios básicos de vivienda):Aquí un aumento de un punto porcentual en `ic_sbv` se asocia con un aumento de 0.105091 puntos porcentuales en la proporción de pobreza.

5.ic_ali (proporción de población con carencia por alimentación): Tiene una post mean de: 0.069983. Este coeficiente es más bajo en magnitud comparado con los anteriores, sugiriendo un impacto menor sobre la proporción de pobreza.Y su post p(B != 0) es de: 0.885241. Que a diferencia de las otras variables, aquí hay una menor certeza (88.52%) de que `ic_ali` tenga un impacto significativo sobre la proporción de pobreza.

En cuanto a los intervalos de credibilidad, podemos afirmar que con un 95% de probabilidad la beta intercepto o constante esta en el intervalo de (0.42160273 - 0.4277656); la beta asociada a la varaible plp esta entre (0.68959369 - 0.7963372); la beta asociada a ic_segsoc esta entre (0.19109903 - 0.2925648), la de ic_sbv esta entre (0.07413911 0.1385078); y la de ic_ali entre 0.00000000 0.1312207). Esto ultimo podria ser un problema, dado que el hecho de que el 0 este incluido en este intervalo podria afectar la significancia de este coeficiente. Sin embargo estamos considerando la P(B != 0 | Y) que es de 0.88 para este coeficiente asociado a variable.

Conclusiones

Como conclusion, podemos decir que el modelo indica que las variables plp, ic_segsoc, ic_sbv, ic_ali, asi como el intercepto o constante, tienen un impacto positivo en la proporción de pobreza, con la mayor influencia proveniente de la proporción de la población bajo la línea de pobreza por ingresos (`plp`), reflejando de esta forma la gran importancia y correlacion que tiene la perspectiva de bienestar economico en la pobreza multidimensional. La variable de carencia por alimentación (`ic_ali`), aunque también parece influir en la pobreza, tiene una menor certeza estadística en comparación con las otras variables elegidas en el modelo.

Referencias:

DOF. Diario Oficial de la Federación (2005 a, 24 de agosto). DECRETO por el que se regula el Consejo Nacional de Evaluación de la Política de Desarrollo Social. Art 3o, fracción II. Recuperado de: https://www.coneval.org.mx/rw/resource/coneval/normateca/2343.pdf

CONEVAL (2019). Metodología para la medición multidimensional de la pobreza en México. Recuperado de: https://www.coneval.org.mx/InformesPublicaciones/InformesPublicaciones/Documents/Metodologia-medicion-multidimensional-3er-edicion.pdf

CONEVAL. (2023). Medición de pobreza 2022: Resumen Ejecutivo. pp 11. Recuperado de https://www.coneval.org.mx/Medicion/MP/Documents/MMP_2022/Pobreza_multidimensional_2022.pdf