Macrolocalizacion de regiones con K-means

Índice


```
## Joining, by = c("CODUSU", "ANO4", "TRIMESTRE", "NRO_HOGAR", "REGION", "MAS_500", "AGLOMERADO", "PONDI
```

```
## Joining, by = "AGLOMERADO"
```

```
## Joining, by = "SECTOR"
```

Carga de datos

Para comenzar, cargamos los datos de las variables sociales y economicas que definen a las distintas provincias de la Argentina.

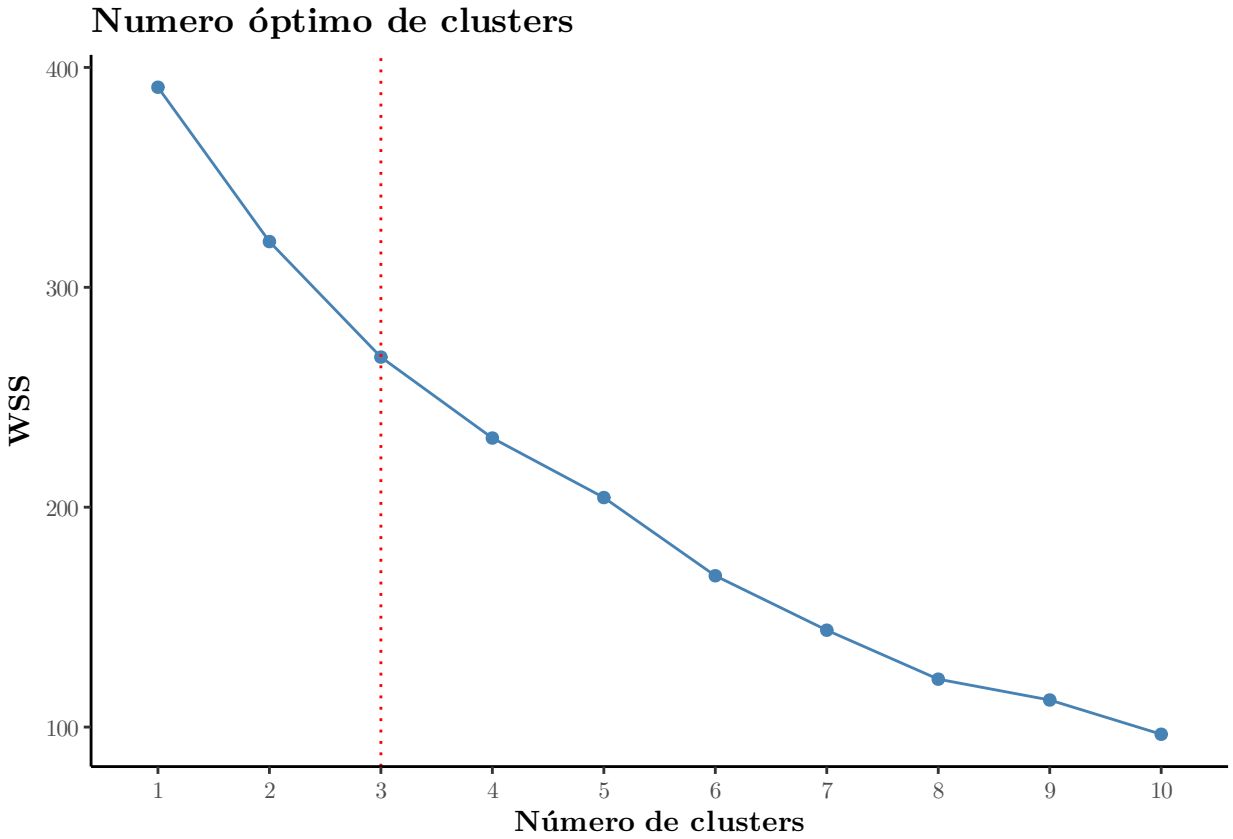
Normalización de la base de datos

Ahora normalizamos la base de datos para lograr coherencia en la aplicación de K-means, debido a quen trabajamos con distintas escalas de datos.

Número de clusters óptimo

Verificamos el numero de clusters optimos a través de el Within-cluster sum of square, el cual es un indicador que mide la suma de las distancias entre las variables dentro de los cluster y sus centroides, la idea es minimizar esta discrepancia dentro de cada grupo, teniendo en cuenta el trade-off que implica cuando este es mínimo, el cual es el caso en que el numero de clusters es igual al número de variables a clusterizar, en donde si tenemos j observaciones, el numero de clusters sería tal que j=k, y no se estaría dando ninguna información relevante a los efectos de poder resumir caracteristicas comunes entre los grupos.

$$WCSS = \sum_{i=1}^{N_C} \sum_{\mathbf{x} \in Ci} distancia(\mathbf{x}, \bar{\mathbf{x}}_{\mathbf{Ci}})$$



Obtenemos que el numero óptimo de clusters se encuentra en torno al **número 3**, debido a que la contribución marginal de aumentar el numero de clusters a 4, no aportaría una reducción muy elevada al **WSS**, y seguiriamos perdiendo generalidad en la clusterización de los grupos sin una ganancia significativa de similitud dentro de los grupos.

Cómputo de las K-means

Setearemos una semilla para el cálculo de los centroides iniciales del algoritmo de K-means y procederemos a realizar el cálculo de los clusters a través de 1000 iteraciones, eligiendo la que de menor numero de WSS.

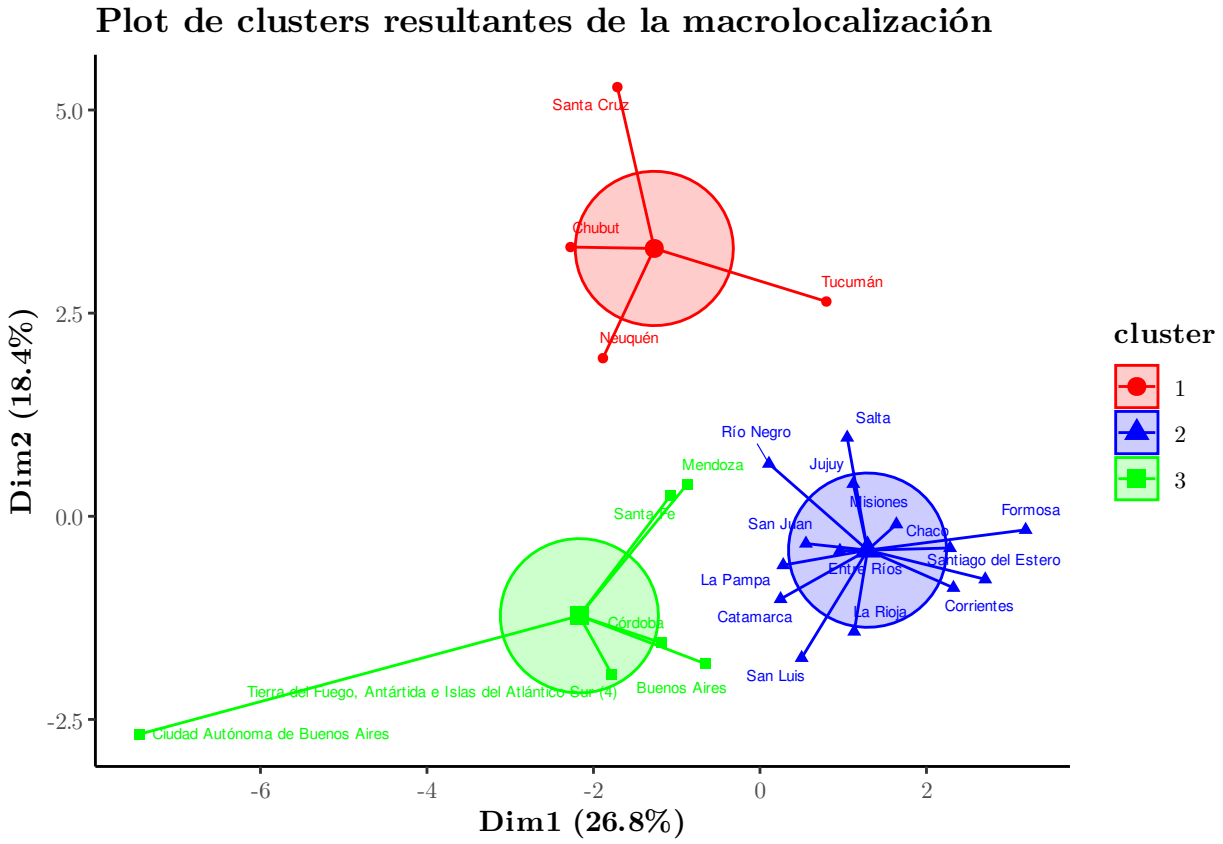
Podemos ver la media de los parámetros para cada uno de los clusters

Cuadro 1: Resumen de indicadores por cluster

Indicadores	Cluster 1	Cluster 2	Cluster 3
Habitantes (2010)	530,187.00	677,181.00	3,042,344.00
Tasa promocion efectiva secundaria (2017)	76.25	78.60	82.05
Mortalidad infantil promedio 2016-2019	8.20	9.50	8.25
Homicidios dolosos C/ 100.000 hab. promedio (2016-2019)	7.25	3.75	5.30
Muertes en Accidentes Viales C/ 100.000 hab. promedio (2016-2019)	12.40	14.05	7.50
Robos (excluye los agravados) C/ 100.000 hab. Promedio (2016-2019)	973.40	654.85	1,566.95
Robos agravados C/ 100.000 hab. Promedio (2016-2019)	27.80	1.70	9.50
Violaciones C/ 100.000 hab. Promedio (2016-2019)	15.45	11.15	8.75
Exportaciones per-cápita en USD promedio (2016-2019)	2,857.20	794.20	1,420.55
Demanda de MWH energía electrica per cápita (2016)	3.80	2.20	3.35
Pobreza Promedio (2017-2019)	26.05	32.55	29.35
Tasa actividad Promedio (2017-2019)	44.55	42.50	46.00

Indicadores	Cluster 1	Cluster 2	Cluster 3
Cantidad empresas C/ 100.000 hab. (2016-2017)	1,569.62	846.96	1,771.95
Remuneracion real de trabajadores registrados del sector privado (2016-2019)	29,995.27	14,997.52	17,149.25
Porcentaje de empleados en Agricultura, ganadería, pesca y actividades extractivas (2016-2019)	0.21	0.13	0.05
Porcentaje de empleados en Comercio, servicios, electricidad, gas, agua y construccion (2016-2019)	0.67	0.69	0.71
Porcentaje de empleados en Industria (2016-2019)	0.10	0.15	0.22

agregamos a la base de datos original la denominación de cada cluster



Se puede ver claramente la diferencia en la similitud de los tres grupos, siendo solamente la Capital Federal la única que presenta mayor disimilitud con respecto a su cluster, lo cual indica que esta capital lleva una dinámica socio-económica muy peculiar con respecto al promedio de las provincias argentinas, inclusive considerablemente distinta que las provincias con la cual ostenta mayor similitud.

Flujos migratorios entre clusters

Dentro de los cluster, existen provincias de las cuales salen gran cantidad de inmigrantes, al igual que existen provincias que son receptoras de estos mismos, en primer lugar caracterizaremos cual es el cluster con mayor nivel de *expulsion* de inmigrantes interprovinciales.

Cuadro 2: Cluster de origen de los migrantes

Cluster	Porcentaje
1	10.41 %
2	61.67 %
3	27.92 %

Se puede notar que las provincias del **cluster Nº2** son las que mayor nivel de expulsión poseen, seguidas por las provincias pertenecientes al cluster Nº3, y por úlitmo se encuentran las pertenecientes al cluster Nº1.

Cuadro 3: Cluster de destino de los migrantes

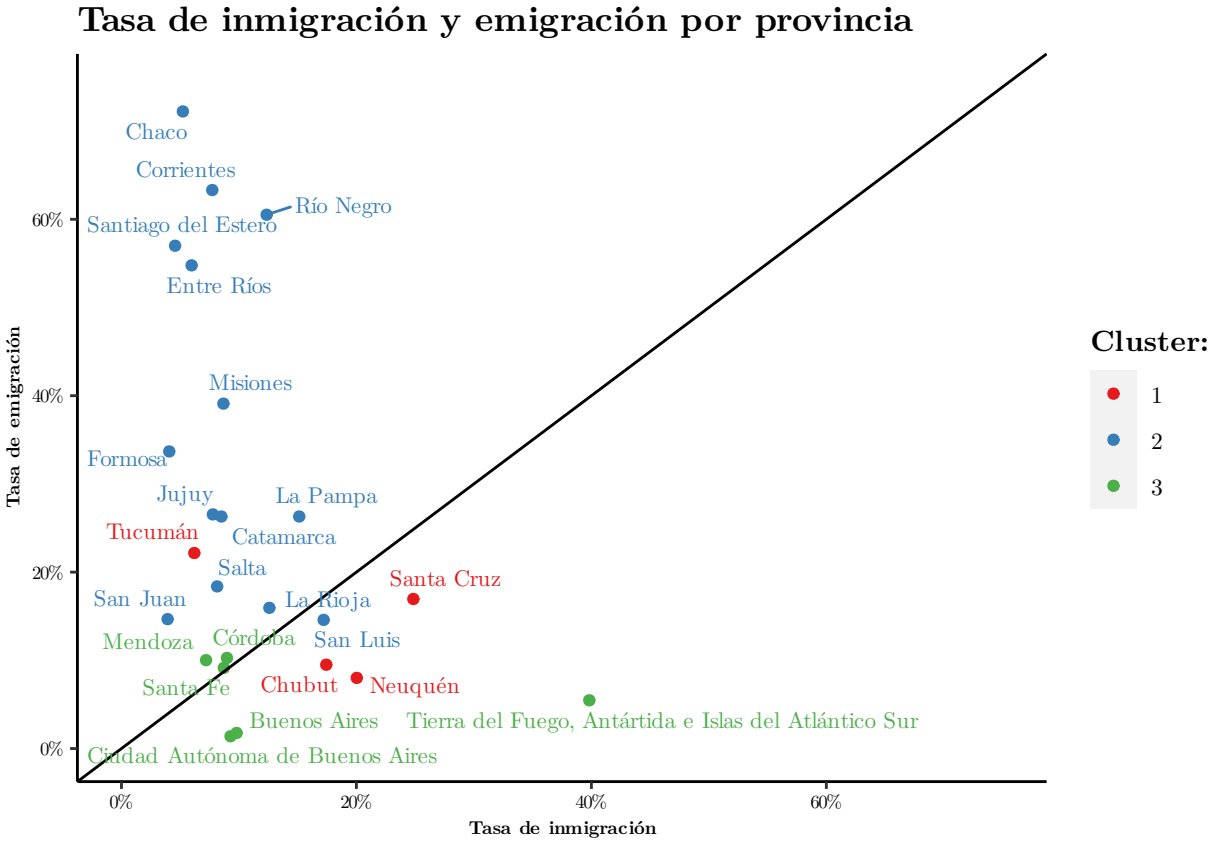
Cluster	Porcentaje
1	7.85 %
2	12.83 %
3	79.32 %

Sin embargo, viendo cuales son los cluster de destino con mayor porcentaje de migrantes, encontramos que el **cluster Nº 3** es el que mayor nivel de *atracción* posee por elevada diferencia (79.32 %)

```
## Joining, by = "Provincia"
```

```
## Joining, by = "AGLOMERADO"
```

```
## Joining, by = "Provincia"
## Joining, by = "Provincia"
```



Definición de migrante según cluster de origen

En primer lugar se analizará si existen microdeterminantes dentro de las características de los migrantes que porovoquen que ciertas personas con tengan una mayor propensión a migrar desde ciertas localizaciones, es decir, se analizaran los factores de expulsión de los distintos cluster; por otro lado, se realizará un análisis diferenciando el destino de los distintos migrantes, con el fin de establecer los determinantes de atracción de las distintas localizaciones.

Factores determinantes de expulsión de los migrantes

En el universo de habitantes de una determinada región/cluster, pueden existir caracteristicas determinadas que diferencien de manera significativa entre las personas que nacieron y decidieron quedarse en la misma región, de las que si bien nacieron en la misma región, optaron por migrar hacia otra. A continuación se realizará un análisis diferenciando por cluster de características socioeconómicas de las personas que nacieron en un mismo cluster, con el fin de determinar si existen diferencias entre los migrantes y los no migrantes.

```
## Joining, by = "AGLOMERADO"
```

```
## Joining, by = "CH15_COD"
```

```
## # A tibble: 2 x 2
## # Groups:   inmigrante_oc1 [2]
##   inmigrante_oc1      n
##            <dbl>  <int>
## 1              0 267575
## 2              1   3934
```

```
##
## Call:
## glm(formula = inmigrante_oc1 ~ edad + medio + alto, family = "binomial",
##     data = inmigrantes_cluster1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.1953  -0.1764  -0.1705  -0.1645   2.9816
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.4381686  0.0615388 -72.120  < 2e-16 ***
## edad         0.0045576  0.0009905   4.601  4.2e-06 ***
## medio       -0.0725633  0.0375778  -1.931   0.0535 .
## alto         0.0232331  0.0433435   0.536   0.5919
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41126  on 271508  degrees of freedom
## Residual deviance: 41094  on 271505  degrees of freedom
## AIC: 41102
##
## Number of Fisher Scoring iterations: 7
```

```
## Joining, by = "AGLOMERADO"
```

```
## Joining, by = "CH15_COD"
```

```
## # A tibble: 2 x 2
```

```
## # Groups:   inmigrante_oc2 [2]
##   inmigrante_oc2      n
##            <dbl>  <int>
## 1              0 248590
## 2              1  22919
```

```
##
## Call:
## glm(formula = inmigrante_oc2 ~ hombre + POBRE + OCUPADO_BAJO +
##     OCUPADO_ALTO + CASADO + PROPIETARIO + HIJO_DUMMY + edad +
##     medio + alto, family = "binomial", data = inmigrantes_cluster2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6262  -0.4493  -0.3983  -0.3495   2.5384
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.087480   0.038247 -54.579  < 2e-16 ***
## hombre        0.048456   0.016539   2.930  0.00339 **
## POBRE        -0.134935   0.016046  -8.410  < 2e-16 ***
## OCUPADO_BAJO -0.213737   0.018044 -11.845  < 2e-16 ***
## OCUPADO_ALTO -0.288424   0.025426 -11.344  < 2e-16 ***
## CASADO       -0.047061   0.016586  -2.837  0.00455 **
## PROPIETARIO  -0.177697   0.015443 -11.507  < 2e-16 ***
## HIJO_DUMMY   -0.163162   0.015581 -10.472  < 2e-16 ***
## edad          0.005252   0.000502  10.462  < 2e-16 ***
## medio        -0.350255   0.016482 -21.251  < 2e-16 ***
## alto         -0.481734   0.023288 -20.686  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 157159  on 271508  degrees of freedom
## Residual deviance: 154910  on 271498  degrees of freedom
## AIC: 154932
##
## Number of Fisher Scoring iterations: 5
```

```
## Joining, by = "AGLOMERADO"
```

```
## Joining, by = "CH15_COD"
```

```
##
## Call:
## glm(formula = inmigrante ~ hombre + POBRE + OCUPADO_BAJO + OCUPADO_ALTO +
##     CASADO + PROPIETARIO + HIJO_DUMMY + edad + medio + alto,
##     family = "binomial", data = cluster1_expulsion)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0901  -0.6169  -0.5313  -0.4581   2.3327
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  -2.063033   0.097078 -21.251  < 2e-16 ***
## hombre       -0.027128   0.041833  -0.648 0.516667
## POBRE         0.147847   0.040731   3.630 0.000284 ***
## OCUPADO_BAJO -0.142579   0.046547  -3.063 0.002191 **
## OCUPADO_ALTO -0.226096   0.060588  -3.732 0.000190 ***
## CASADO        0.067918   0.042616   1.594 0.110998
## PROPIETARIO  -0.266583   0.038793  -6.872 6.33e-12 ***
## HIJO_DUMMY   -0.411228   0.039150 -10.504  < 2e-16 ***
## edad          0.012819   0.001307   9.809  < 2e-16 ***
## medio         0.208129   0.041572   5.006 5.54e-07 ***
## alto          0.539903   0.055158   9.788  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21778  on 25043  degrees of freedom
## Residual deviance: 21282  on 25033  degrees of freedom
## AIC: 21304
##
## Number of Fisher Scoring iterations: 4

## Joining, by = "AGLOMERADO"

## Joining, by = "CH15_COD"

##
## Call:
## glm(formula = inmigrante ~ hombre + POBRE + OCUPADO_BAJO + OCUPADO_ALTO +
##     CASADO + PROPIETARIO + HIJO_DUMMY + edad + medio + alto,
##     family = "binomial", data = cluster2_expulsion)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0864  -0.6962  -0.5989  -0.4861   2.2244
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2287446  0.0409076 -30.037  < 2e-16 ***
## hombre        0.0727253  0.0176336   4.124 3.72e-05 ***
## POBRE        -0.0160034  0.0170623  -0.938 0.348277
## OCUPADO_BAJO -0.1272368  0.0192257  -6.618 3.64e-11 ***
## OCUPADO_ALTO -0.1273920  0.0273878  -4.651 3.30e-06 ***
## CASADO        0.0642979  0.0177376   3.625 0.000289 ***
## PROPIETARIO  -0.4925933  0.0165084 -29.839  < 2e-16 ***
## HIJO_DUMMY   -0.3881879  0.0166631 -23.296  < 2e-16 ***
## edad          0.0096384  0.0005472  17.613  < 2e-16 ***
## medio        -0.3180610  0.0178557 -17.813  < 2e-16 ***
## alto         -0.3709773  0.0252563 -14.689  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 115535  on 116511  degrees of freedom
```

```
## Residual deviance: 112691  on 116501  degrees of freedom
## AIC: 112713
##
## Number of Fisher Scoring iterations: 4
```

```
## Joining, by = "AGLOMERADO"
```

```
## Joining, by = "CH15_COD"
```

```
##
## Call:
## glm(formula = inmigrante ~ hombre + POBRE + OCUPADO_BAJO + OCUPADO_ALTO +
##     CASADO + PROPIETARIO + HIJO_DUMMY + edad + medio + alto,
##     family = "binomial", data = cluster3_expulsion)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8209  -0.6197  -0.5416  -0.4555   2.2661
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.0691249  0.0466048 -44.397  < 2e-16 ***
## hombre        0.0444512  0.0195821   2.270  0.02321 *
## POBRE        -0.4402895  0.0196645 -22.390  < 2e-16 ***
## OCUPADO_BAJO -0.0973594  0.0225800  -4.312 1.62e-05 ***
## OCUPADO_ALTO -0.0664447  0.0266759  -2.491  0.01275 *
## CASADO       -0.0543753  0.0197562  -2.752  0.00592 **
## PROPIETARIO   0.0099968  0.0186541   0.536  0.59202
## HIJO_DUMMY    0.0084777  0.0185618   0.457  0.64787
## edad          0.0071722  0.0006026  11.901  < 2e-16 ***
## medio         0.2625130  0.0197308  13.305  < 2e-16 ***
## alto          0.3684715  0.0241858  15.235  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 101060  on 118753  degrees of freedom
## Residual deviance:  99595  on 118743  degrees of freedom
## AIC: 99617
##
## Number of Fisher Scoring iterations: 4
```

Cuadro 4:

	Dependent variable:		
	inmigrante		
	(1)	(2)	(3)
hombre	−0.028	0.073***	0.044**
	(0.042)	(0.018)	(0.020)
POBRE	0.145***	−0.016	−0.440***
	(0.041)	(0.017)	(0.020)
OCUPADO_BAJO	−0.147***	−0.127***	−0.097***
	(0.047)	(0.019)	(0.023)
OCUPADO_MEDIO	−0.329***		
	(0.068)		
OCUPADO_ALTO	−0.012	−0.127***	−0.066**
	(0.083)	(0.027)	(0.027)
CASADO	0.065	0.064***	−0.054***
	(0.043)	(0.018)	(0.020)
PROPIETARIO	−0.268***	−0.493***	0.010
	(0.039)	(0.017)	(0.019)
HIJO_DUMMY	−0.409***	−0.388***	0.008
	(0.039)	(0.017)	(0.019)
edad	0.013***	0.010***	0.007***
	(0.001)	(0.001)	(0.001)
medio	0.208***	−0.318***	0.263***
	(0.042)	(0.018)	(0.020)
alto	0.511***	−0.371***	0.368***
	(0.056)	(0.025)	(0.024)
Constant	−2.046***	−1.229***	−2.069***
	(0.097)	(0.041)	(0.047)
Observations	25,044	116,512	118,754
Log Likelihood	−10,634.160	−56,345.440	−49,797.470
Akaike Inf. Crit.	21,292.310	112,712.900	99,616.950
Note:			*p<0.1; **p<0.05; ***p<0.01