Estadística Aplicada 3
Proyecto final
ITAM 2017

Federico Garza Ramírez. CU: 143949.

1 Introducción

El presente trabajo pone en práctica dos técnicas estadísticas estudiadas en el curso Estadística Aplicada 3: Análisis de Componentes Principales y Modelos loglineales. Cada aplicación surge como respuesta a una pregunta específica y lo largo de cada sección se establecen:

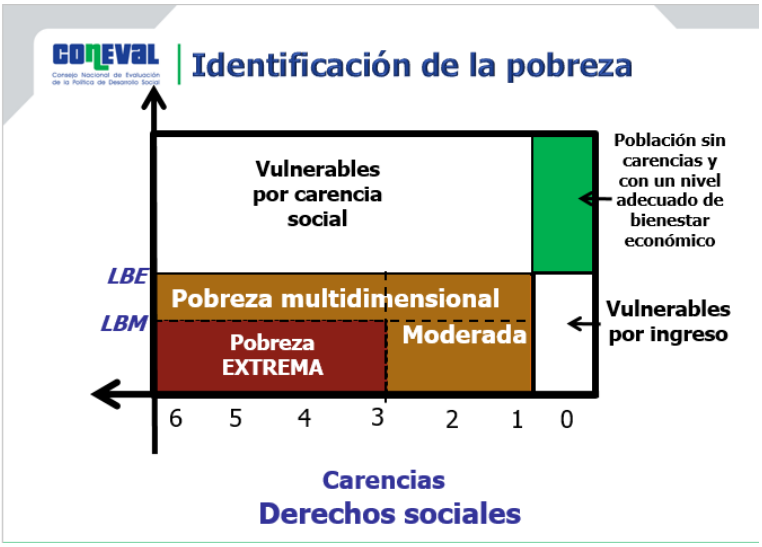
1. Pregunta a responder.
2. Selección y descripción de la base de datos empleada.
3. Empleo del método y resultados.
4. Conclusiones y respuesta a la pregunta inicial.

Se omiten aquí los códigos en **R** empleados pero a menudo se realizan referencias a los mismos; sin embargo, se incluyen *in extenso* en las carpetas adjuntas con explicaciones detalladas de todo lo realizado.

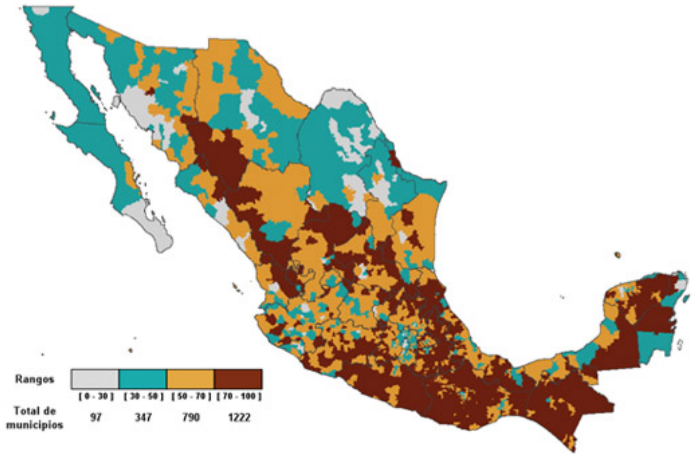
2 Parte: Análisis de Componentes Principales (ACP)

2.1 Motivación

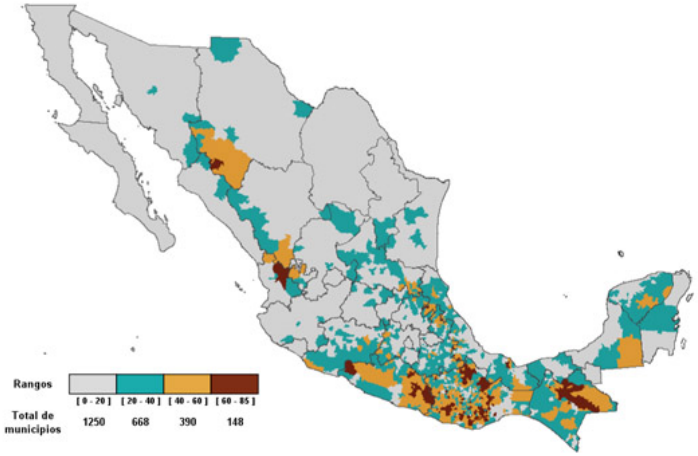
El Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL) mide oficialmente la pobreza en México a través de la siguiente metodología,



Es decir, el CONEVAL **cuenta** el número de pobres de acuerdo a los anteriores criterios (decididos por el Congreso). Con los datos del censo poblacional del 2010 este organismo determinó la distribución de los municipios con mayor pobreza en el país, resumiéndola en el siguiente mapa,



En tanto que la pobreza extrema se ve,



Con esto resulta relevante preguntar, *¿el índice de marginación construido por el ACP (entendido como la primer componente del ACP) arroja resultados similares que los establecidos por el CONEVAL?* Veremos que la respuesta a esta pregunta es afirmativa.

2.2 Selección de la base de datos

Para poder establecer la relación entre el índice de marginación del ACP (IMACP) y los resultados de pobreza por municipio obtenidos por el CONEVAL es necesario utilizar **la misma base de datos** por municipio; es decir, la base de datos ***Principales resultados por localidad (ITER)*** del Instituto Nacional de Estadística y Geografía (INEGI).

La base de datos ***ITER*** disponible en internet **agrupa los datos por estado** por lo que en primer término se agruparon **todos los municipios** de la República Mexicana en un solo archivo `datos_agrupados.xls` a través de los métodos en `agrupamientodatos.R`. Asimismo se limpió esta base para que pudiera utilizarse fácilmente.

Con la nueva base de datos se elaboró el IMACP.

2.3 Empleo del método y resultados

Una vez obtenida la base de datos por municipio (recordemos `datos_agrupados.xls`) se seleccionaron las variables relevantas para la construcción del índice en los métodos `creacionindice.R`. Originalmente `datos_agrupados.xls` cuenta con 190 variables provenientes de las encuestas realizadas por el INEGI en el 2010. Debemos ser cuidadosos en la selección de variables: muchas de ellas están altamente correlacionadas por lo que los resultados del ACP pueden no ser relevantes. Considérese, por ejemplo, las variables `POBTOT` y `POBFEM` que denotan a la población total del municipio y la población femenina. Obtenemos que la correlación entre ambas variables es de 0.9998034; es decir, están altamente correlacionadas. Sin embargo, este resultado no debe parecernos particularmente sorprendente: la población femenina pertenece a la población total y por lo tanto están intimamente relacionadas. Es por esto que debe hacerse *ex ante* un análisis cualitativo que considere relaciones entre las variables por cómo están construidas.

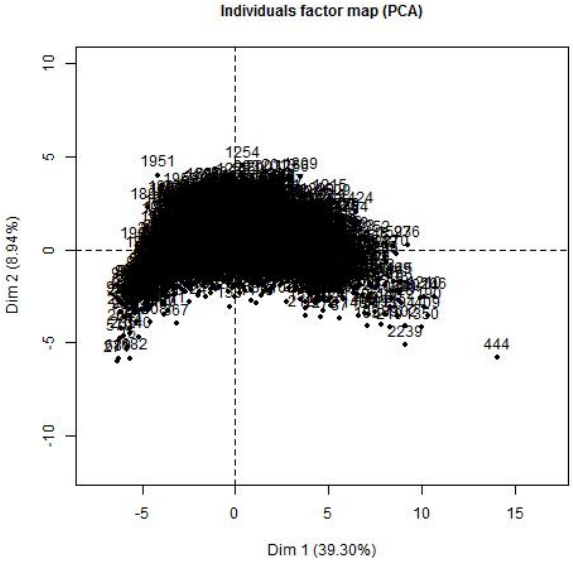
Se seleccionaron, entonces, las variables relevantes de acuerdo a dos criterios:

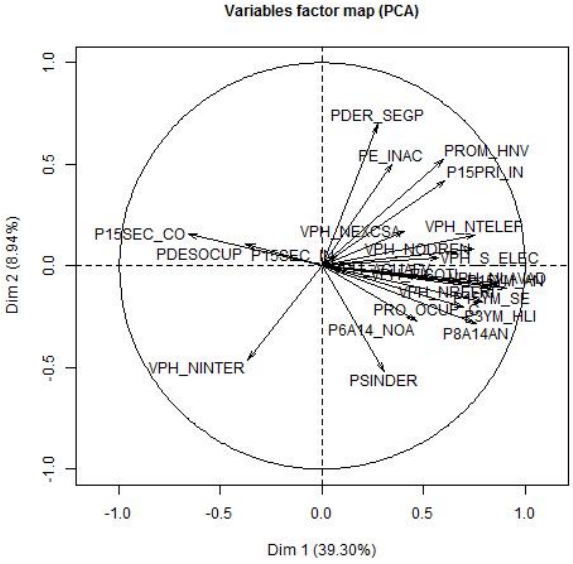
- Las variables que utiliza el CONEVAL para su medición de la pobreza.
- Inexistencia de relación cualitativa en las variables.

Con esto las variables seleccionadas fueron (convertidas a porcentaje),

Variable	Detalle
PROM_HNV	Promedio de hijos nacidos vivos
P3YM_HLI	Personas de 3 y más años que hablan alguna lengua indígena
P8A14AN	Personas de 8 a 14 años analfabetas
P15YM_AN	Personas de 15 y más años analfabetas
P6A14_NOA	Personas de 6 a 14 años que no asisten a la escuela
P15YM_SE	Personas de 15 y más años solo con nivel preescolar
P15PRI_IN	Personas de 15 y más años con primaria incompleta
P15SEC_IN	Personas de 15 y más años con secundaria incompleta
P15SEC_CO	Personas de 15 y más años con secundaria completa
PE_INAC	Población no activa económicamente
PDESOCUP	Población desocupada
PSINDER	Población sin servicios médicos
PDER_SEGP	Población con seguro popular
PRO_OCUP_C	Porporción de ocupantes por cuarto
VPH_PISOTI	Viviendas con piso de tierra
VPH_S_ELEC	Viviendas sin electricidad
VPH_AGUAFV	Viviendas sin agua entubada
VPH_NODREN	Viviendas sin drenaje
VPH_NINTER	Viviendas sin internet
VPH_NREFRI	Viviendas sin refrigerador
VPH_NLAVAD	Viviendas sin lavadora
VPH_NTELEF	Viviendas sin televisión
VPH_NEXCSA	viviendas sin excusado o retrete

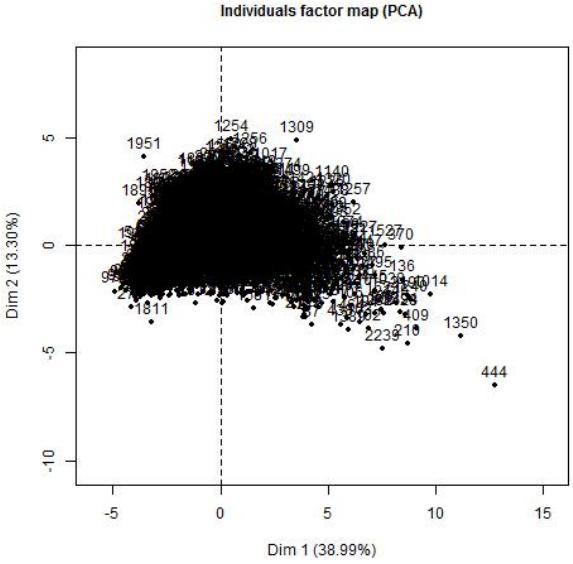
Tenemos un total de 23 variables. Si consideramos todas las variables en primera instancia obtenemos del ACP,

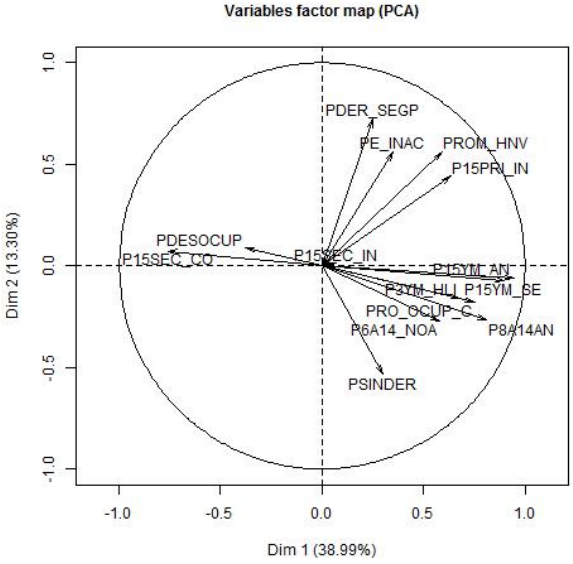




Es decir, se explica el 39.30% de la variabilidad. Observamos que la componente horizontal posee más variables de rezago en la vivienda; en cambio, la componente horizontal posee más variables de desarrollo posterior.

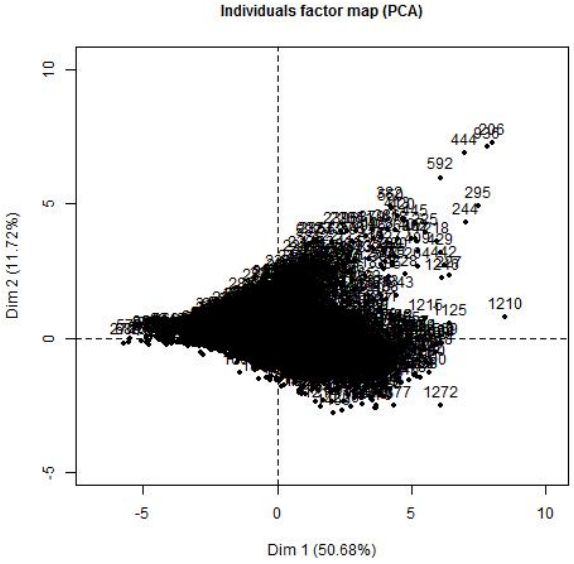
Si quitamos las variables de vivienda obtenemos,

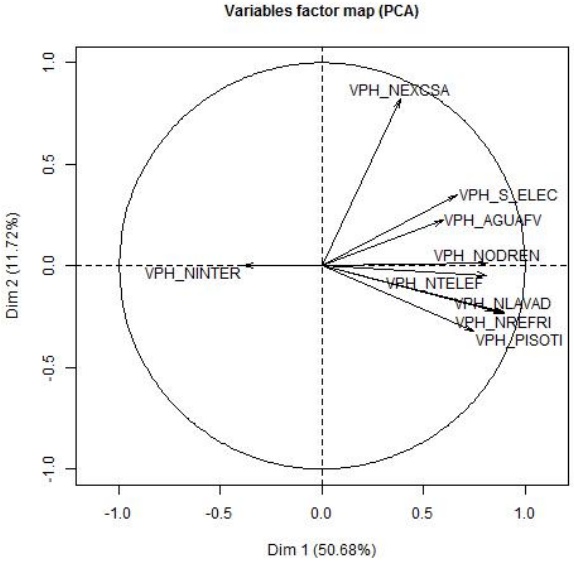




Es decir, se explica el 38.99% de la variabilidad: baja la varianza explicada. Observamos que las componentes horizontal y vertical conservan las mismas propiedades.

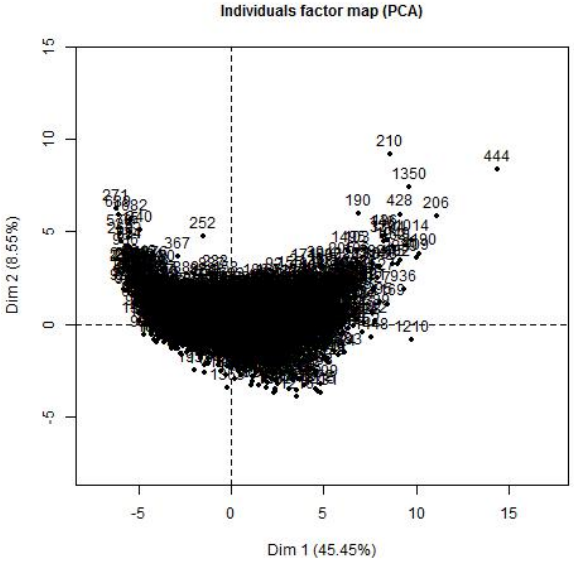
Si dejamos solo las variables de vivienda obtenemos,

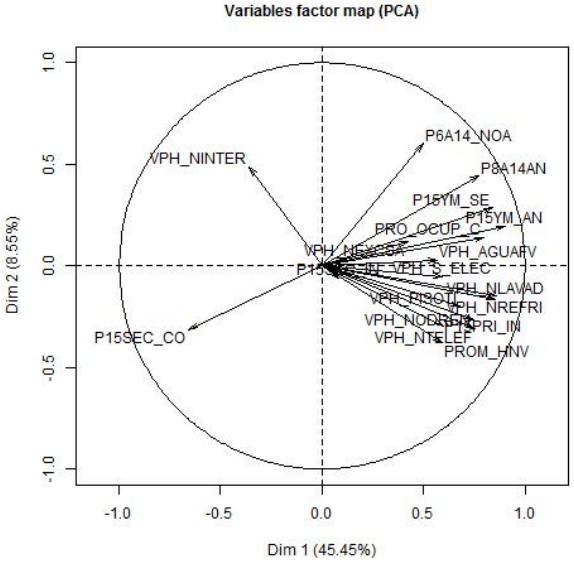




Aumenta la variabilidad explicada a 50.68%. Observamos que las componentes horizontal y vertical conservan las mismas propiedades.

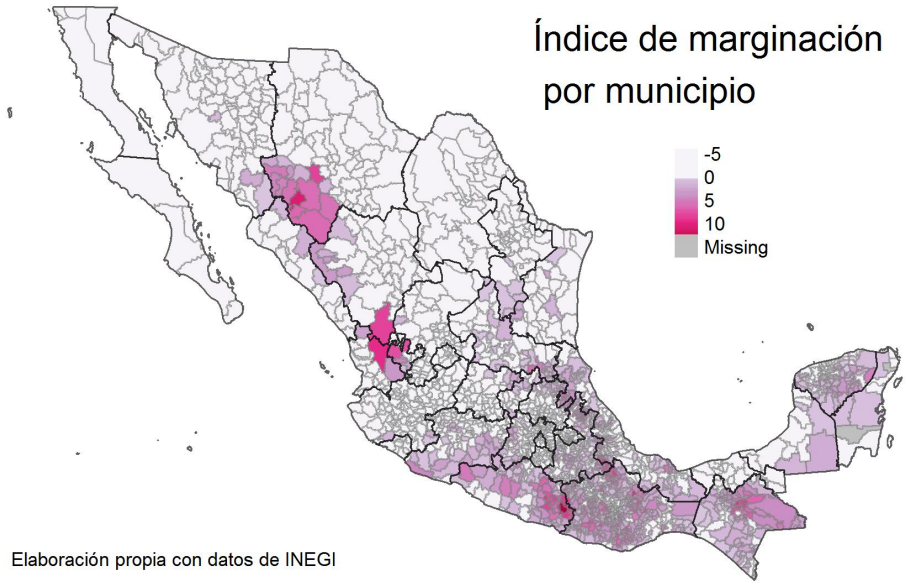
Si quitamos algunas variables *raras* como promedio de hijos nacidos vivos y otros obtenemos,





Se explica ahora el 45.5% de la variabilidad. Observamos que en los Análisis de Componentes Principales realizados no hay mucho aumento de la variabilidad explicada. Por tanto elegimos donde están todas las variables para que se pueda tener la mayor representatividad posible. Tomando la primer componente como el índice de marginación, los datos obtenidos por municipio se guardaron en `indice_marginacion.xls`

Entonces el índice de marginación por municipios obtenido es,



En los archivos adjuntos puede encontrase este mismo mapa de forma interactiva.

2.4 Conclusiones

En conclusión podemos observar cierta semejanza entre los mapas obtenidos por el CONEVAL y por el IMACP pues zonas como el sureste de México y la Sierra Madre Occidental que son las más pobres, también resultan las más marginadas de acuerdo a IMACP.

3 Modelo loglineal

3.1 Motivación

Una de las peores características de nuestro país es la falta de oportunidades de acceso a niveles elevados de instrucción educativa como licenciatura, especialización o posgrado. Dado que muy pocos alcanzan estos niveles es relevante preguntar, *¿puede decirse que el nivel de satisfacción con la vida* ***depende*** *del nivel de instrucción?*

3.2 Selección de la base de datos

Para contestar la pregunta anterior **untilizaremos un modelo loglineal**. La base de datos elegida para este propósito es el **Módulo de Bienestar Subjetivo, BIARE** desarrollado por el INEGI en 2014.

Comenzamos seleccionando las siguientes variables categóricas,

1. `sexo` = `M` o `H`, femenino y masculino respectivamente.
2. `satisfaccion` = `M_I`, `I`, `S`, `M_S`, muy insatisfecho, insatisfecho, satisfecho y muy satisfecho respectivamente.
3. `instruccion` = `N`, `PRI_CO`, `SEC_CO`, `PREP`, `LIC`, `POSG`, ningún grado de instrucción educativa, primaria completa, secundaria completa. preparatoria, licenciatura y posgrado respectivamente.

La limpieza y selección de estas variables se guardó en `satisfaccion_mexico.xls` y se realizó en `creacion_satisfaccion.R`.

3.3 Empleo del método y resultados

El empleo del modelo se realizó en `loglineal.R`. La muestra de los datos a analizar quedó como sigue (una pequeña visualización),

```
     satisfaccion sexo instruccion   count
1         M_INS    M           N  211117
2           INS    M           N  607890
3           SAT    M           N 1485451
4         M_SAT    M           N 1146652
5         M_INS    H           N   85574
6           INS    H           N  281974
7           SAT    H           N  851682
8         M_SAT    H           N  563930
9         M_INS    M      PRI_CO  580355
10          INS    M      PRI_CO 1166276
```

Se comenzó analizando el **modelo saturado**. Con este se empleó el método `seq` para establecer cuál es el que mejor ajusta. Se obtuvo el siguiente resultado,

```
> ajuste.NS_S_I = loglm(count~ satisfaccion * sexo * instruccion,
+                       data = muestra)
> stepAIC(ajuste.NS_S_I, direction = "backward", test="Chisq",
+         scope = list(upper = ~satisfaccion*sexo*instruccion,
+                      lower = ~satisfaccion+sexo+instruccion),
+         trace = TRUE)
Start:  AIC=96
count ~ satisfaccion * sexo * instruccion

                                Df   AIC   LRT   Pr(Chi)
<none>                               96
- satisfaccion:sexo:instruccion 15 33094 33028 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Call:
loglm(formula = count ~ satisfaccion * sexo * instruccion, data = muestra)

Statistics:
                 X^2 df P(> X^2)
Likelihood Ratio   0  0        1
Pearson            0  0        1
```

Donde observamos que de hecho el modelo saturado es el que mejor ajusta. ¿Podemos ignorar los restantes?

```
> ajuste.NS_S.NS_I.S_I = update(ajuste.NS_S_I,
+                               .~. -satisfaccion:sexo:instruccion)
> ajuste.NS_S.NS_I.S_I
Call:
loglm(formula = count ~ satisfaccion + sexo + instruccion + satisfaccion:sexo +
    satisfaccion:instruccion + sexo:instruccion, data = muestra)

Statistics:
                     X^2 df P(> X^2)
Likelihood Ratio 33028.35 15        0
Pearson          33317.29 15        0
```

Observamos que el ajuste **no es significativo**. Sin embargo, desarrollamos todos los modelos posibles para tres variables categóricas (nueve) y vimos el ajuste que hacen de los datos. Obtuvimos lo siguiente (muestra parcial),

```
satisfaccion sexo instruccion NS_S_I NS_S.NS_I.S_I  NS_I.S_I  NS_S.S_I NS_S.NS_I
48        M_SAT    H        POSG 211117     221563.41 195617.21 172104.84 201172.67
47          SAT    H        POSG 580355     580251.80 501328.89 383540.15 567199.37
46          INS    H        POSG 672423     650940.77 558990.63 614880.23 642275.60
45        M_INS    H        POSG 260993     247772.96 206500.32 380408.25 266722.91
44        M_SAT    M        POSG 149019     166620.74 137051.49 304075.53 187627.85
43          SAT    M        POSG  13773      20530.32  17094.55  32671.00  22681.60
42          INS    M        POSG  85574      75127.59 101073.79  54673.67  95518.33
41        M_INS    M        POSG 256155     256258.20 335181.11 157659.90 269310.63
40        M_SAT    H         LIC 274810     296292.23 388242.37 262568.46 304957.40
39          SAT    H         LIC 132372     145592.04 186864.68 211646.00 126642.09
```

Observamos que a pesar de no ser significativo, el homogéneo obtiene buenos resultados. Por otro lado centremos nuestro análisis restante en los coeficientes del modelo saturado,

```
> ajuste.NS_S_I$param$satisfaccion.sexo.instruccion
, , instruccion = N

            sexo
satisfaccion            M            H
       M_INS -0.001315789  0.001315789
       INS   -0.011598208  0.011598208
       SAT   -0.007959118  0.007959118
       M_SAT  0.020873114 -0.020873114

, , instruccion = PRI_CO

            sexo
satisfaccion           M           H
       M_INS  0.06615605 -0.06615605
       INS   -0.01346050  0.01346050
       SAT   -0.02749116  0.02749116
       M_SAT -0.02520439  0.02520439

, , instruccion = SEC_CO

            sexo
satisfaccion           M           H
       M_INS  0.10473976 -0.10473976
       INS   -0.01348184  0.01348184
       SAT   -0.06942120  0.06942120
       M_SAT -0.02183672  0.02183672

, , instruccion = PREP

            sexo
satisfaccion           M           H
       M_INS  0.13503876 -0.13503876
       INS   -0.08623762  0.08623762
       SAT   -0.03448601  0.03448601
       M_SAT -0.01431514  0.01431514

, , instruccion = LIC

            sexo
satisfaccion            M           H
```

```
       M_INS -0.037512524  0.037512524
       INS    0.026836822 -0.026836822
       SAT    0.002368471 -0.002368471
       M_SAT  0.008307231 -0.008307231

, , instruccion = POSG

            sexo
satisfaccion           M           H
       M_INS -0.26710625  0.26710625
       INS    0.09794134 -0.09794134
       SAT    0.13698901 -0.13698901
       M_SAT  0.03217590 -0.03217590
```

Observemos, por ejemplo, que el logaritmo de las observaciones cuando el factor es posgrado en el caso de las **mujeres aumenta en** 0.0321 en tanto que en el caso de los **hombres disminuye**.

3.4 Conclusiones

En conclusión, el único modelo que satisface la prueba ^{2} es el saturado por lo que no podemos hablar de ausencia en las relaciones entre las variables categóricas selccionadas.