Modelos lineales generalizados para datos categóricos

Carlos Iván Canto Varela

La visualización inicial de las frecuencias relativas para los niveles de satisfacción se puede apreciar con una pared de ladrillos (gráfico de mosaico).

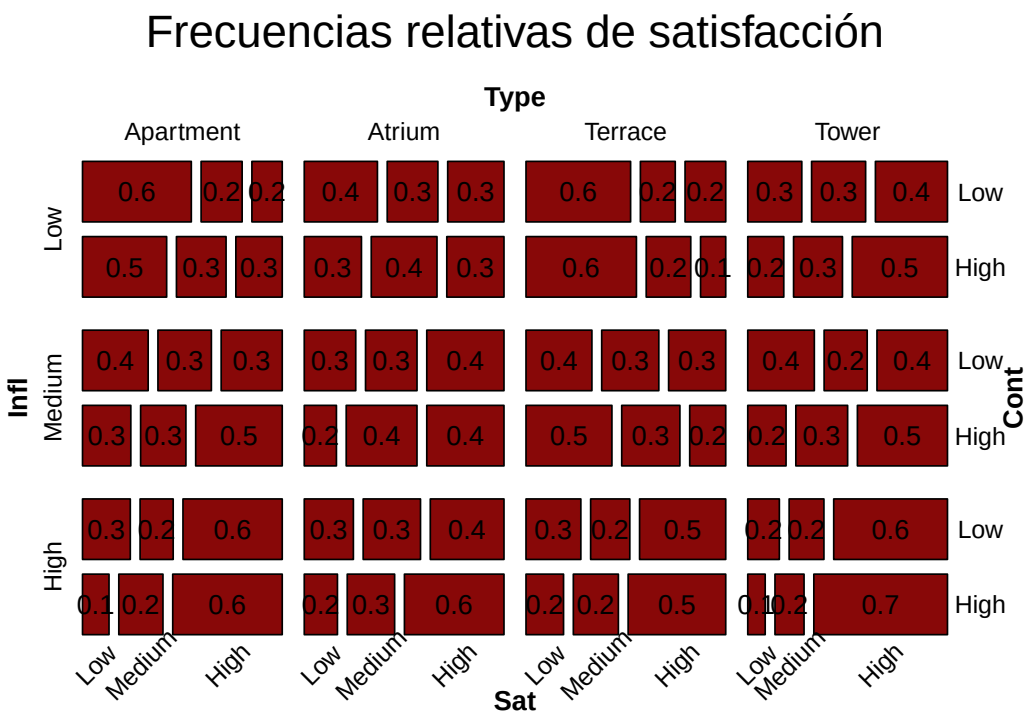

Figura 1: Gráfico de mosaico para las frecuencias relativas en los niveles de satisfacción por combinación

El grupo que muestra más satisfacción es el de las torres, mientras que los más miserables son los que escogieron vivir en terraza. Se observa una tendencia donde a mayor nivel de influencia sobre el mantenimiento, mayor satisfacción y viceversa. También parece darse para el contacto con inquilinos; hasta en aquellos de la categoría *Tower*.

Uno asumiría que si alguien se muda a una torre es porque preferiría estar solo; sé que si yo lo hiciera y tuviera vecinos estaría profundamente insatisfecho con el arrendador.

Con estos datos, el interés es ajustar un modelo logístico multinomial.

En primera instancia, un modelo sin interacciones tiene mejores criterios que uno con todas las interacciones entre categorías:

Puntajes por modelo		
Modelo	**Puntaje AIC**	**Puntaje BIC**
Con interacciones	3527.422	3787.925
Sin interacciones	**3498.084**	**3574.064**

Similarmente, la prueba de hipótesis para comprobar la plausibilidad del modelo señaló una ausencia de evidencia para rechazar que el modelo reducido sea inútil con un valor p mayor a 0.05.

```
## Analysis of Deviance Table
##
## Model 1: Sat ~ Infl * Cont * Type
## Model 2: Sat ~ Infl + Cont + Type
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1      3314     3431.4
## 2      3348     3470.1 -34  -38.662   0.2671
```

Este par de aseveraciones es suficiente para trabajar únicamente con el modelo reducido. Sin embargo, se puede hallar un modelo logístico acumulativo al considerar ordinal a la variable de satisfactibilidad.

Similarmente, se verá la factibilidad de dos modelos distintos:

Puntajes por modelo		
Modelo	**Puntaje AIC**	**Puntaje BIC**
Regular	3498.579	3574.559
Con supuesto de proporcionalidad	**3495.149**	**3538.566**

y con la prueba para el modelo anidado (con supuesto de probabilidad):

```
## Analysis of Deviance Table
##
## Model 1: Sat ~ Infl + Cont + Type
## Model 2: Sat ~ Infl + Cont + Type
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3348     3470.6
## 2      3354     3479.1 -6  -8.5706   0.1992
```

se concluye -con una significancia del 5 %- que el modelo con probabilidades proporcionales es de mayor utilidad entre estos dos. Este es inclusive mejor que el logístico multinomial al reducir por 3 puntos el criterio AIC y por 36 al BIC; lo que es de gran conveniencia porque tiene menos parámetros.

Finalmente, se tienen las probabilidades de nivel de satisfacción para cada caso estimadas por el modelo:

```
##      Infl      Type Cont       Low    Medium      High
## 1     Low Apartment  Low 0.5190446 0.2605076 0.2204478
## 2     Low Apartment High 0.4294564 0.2820628 0.2884808
## 3     Low    Atrium  Low 0.4675586 0.2745382 0.2579032
## 4     Low    Atrium High 0.3798388 0.2875971 0.3325641
## 5     Low   Terrace  Low 0.6444841 0.2114255 0.1440905
## 6     Low   Terrace High 0.5583814 0.2471825 0.1944361
## 7     Low     Tower  Low 0.3784495 0.2876751 0.3338754
## 8     Low     Tower High 0.2980880 0.2837746 0.4181374
## 9  Medium Apartment  Low 0.3798515 0.2875964 0.3325521
## 10 Medium Apartment High 0.2993357 0.2839753 0.4166890
## 11 Medium    Atrium  Low 0.3326238 0.2876008 0.3797754
## 12 Medium    Atrium High 0.2579547 0.2745537 0.4674916
## 13 Medium   Terrace  Low 0.5071211 0.2641195 0.2287594
## 14 Medium   Terrace High 0.4178032 0.2838213 0.2983756
## 15 Medium     Tower  Low 0.2568266 0.2742122 0.4689612
```

```
## 16 Medium     Tower High 0.1942210 0.2470589 0.5587201
## 17   High Apartment  Low 0.2292408 0.2643196 0.5064396
## 18   High Apartment High 0.1718051 0.2328648 0.5953301
## 19   High    Atrium  Low 0.1948550 0.2474227 0.5577223
## 20   High    Atrium High 0.1444204 0.2117081 0.6438715
## 21   High   Terrace  Low 0.3331576 0.2876330 0.3792094
## 22   High   Terrace High 0.2584150 0.2746916 0.4668934
## 23   High     Tower  Low 0.1436926 0.2110837 0.6452237
## 24   High     Tower High 0.1047771 0.1724227 0.7228002
```

Como era de esperarse, las probabilidades encontradas tienen bastante parecido con las frecuencias relativas, de modo que las interpretaciones iniciales siguen siendo mayoritariamente válidas. Sin embargo, los nuevos ajustes implican que el gráfico de mosaico ya no es tan preciso. Esta tabla puede puede ser difícil de interpretar y no agradable a la vista. En vez de eso, es más intuitivo analizar casos concretos.

Por ejemplo, el efecto que se observa al considerar la variable *Infl* cuando se asume que la persona renta una vivienda tipo *Tower* y tiene un nivel de contacto *Low* con otros inquilinos se puede ver de manera clara en el siguiente gráfico.

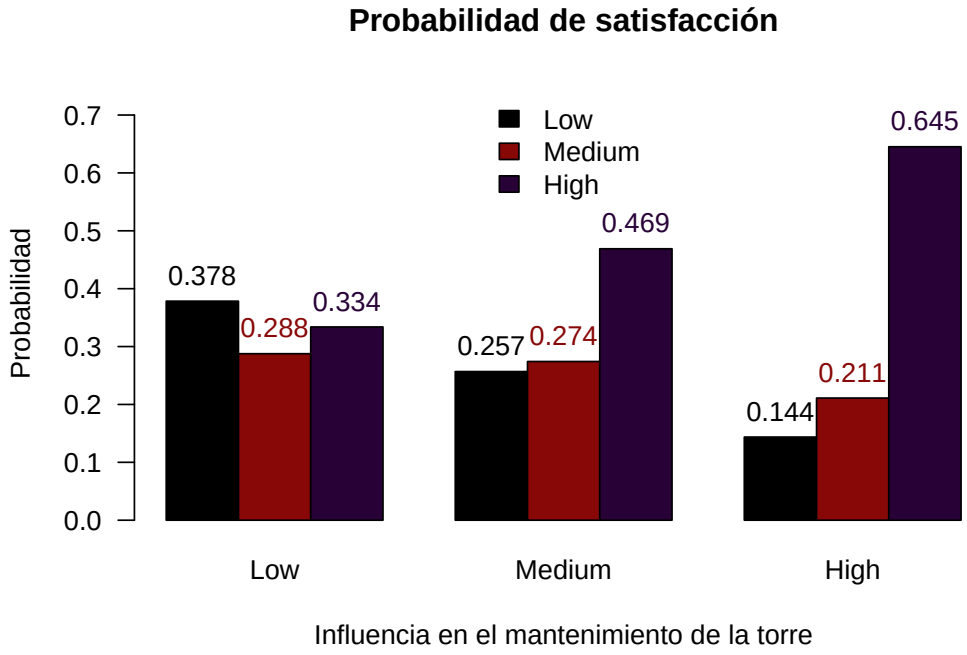

Figura 2: Probabilidades de satisfacción según del modelo al variar la influencia en el mantenimiento

Puede verse que a mayor influencia en el mantenimiento de la torre, la probabilidad de satisfacción alta aumenta, mientras que las de satisfacción baja y media disminuyen. Sin embargo, la de media parece hacerlo de forma mucho más tenue.