5. Modelos lineales generalizados para datos categóricos

La base de datos *Preg5.csv* contiene información sobre el nivel de satisfacción (Sat) de un conjunto de individuos que rentan una vivienda. El interés es identificar si entre los factores que definen este nivel están: el tipo de vivienda (Type), la percepción sobre su influencia en las decisiones sobre el mantenimiento de la vivienda (Infl) y el contacto que tienen con el resto de inquilinos (Cont).

i) Gráfica de frecuencias relativas.

Una vez cargados los datos, se agregó una variable adicional en la que se describen las posibles combinaciones de las categorias de Infl *(Low, Medium, High)*, Type *(Apartment, Atrium, Terrace, Tower)* y Cont *(Low, High)*. En el cuadro 13 se presenta una muestra aleatoria de los datos para observar su estructura de acuerdo a las categorias.

Cuadro 13: Muestra aleatoria de los datos Preg5.csv

X	Sat	Infl	Type	Cont	Infl.Type.Cont
187	Low	Low	Atrium	Low	Low.Atrium.Low
1589	High	High	Apartment	High	High.Apartment.High
1286	High	Medium	Apartment	Low	Medium.Apartment.Low
408	Low	Medium	Apartment	High	Medium.Apartment.High
595	Medium	Low	Tower	High	Low.Tower.High

Para representar visualmente esta información, en la figura 18 se muestra un gráfico de barras con las frecuencias relativas correspondientes a cada grupo de categorias.

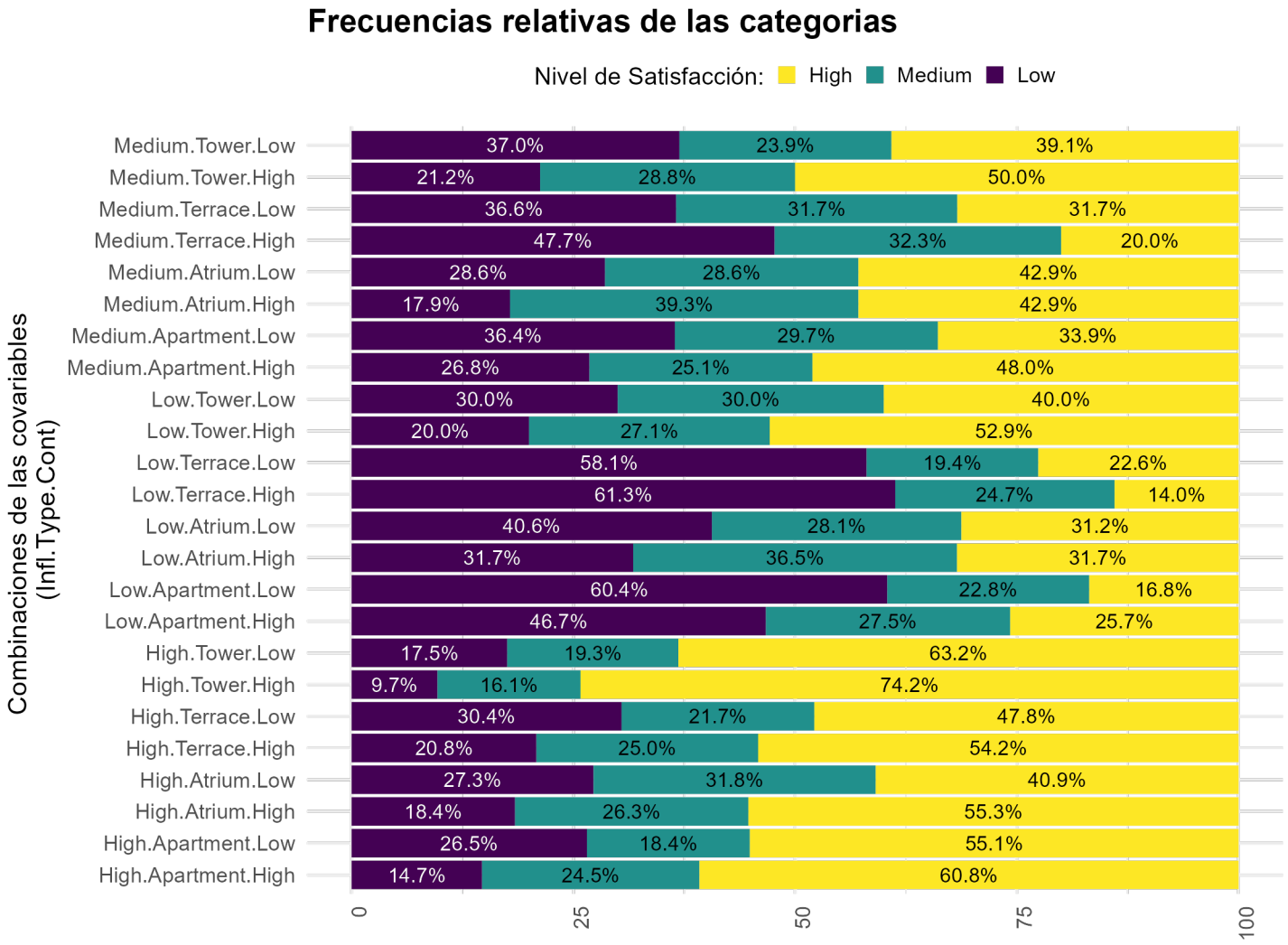

Figura 18: BarChart: Distribución del nivel de satisfacción por cada grupo de individuos.

En esta gráfica se observa que los grupos que tienen un mayor porcentaje de individuos con alta satisfacción, tienen en común una alta influencia en las decisiones del mantenimiento de la vivienda. De igual manera en las categorias donde abundan los individuos con un

nivel de satisfacción bajo, tienen en común un nivel bajo también en su influencia en el mantenimiento. Por lo tanto, se puede deducir que la variable Infl, es considerablemente importante en la satisfacción final del arrendatario.

ii) Ajuste de modelos logísticos multinomiales

Debido a que se está tratando con variables categóricas, se utilizará la función *vglm*() de la paquetería VGAM en la cuál se pueden ajustar modelos lineales generalizados con distribución multinomial.

Sabiendo que los niveles categóricos de la variable Sat están en el orden *Low, Medium, High*, se escogió el nivel de referencia como "Low" al ajustar los modelos.

```
# Modelo con interacciones (nivel de referencia "Low")
fit_int <- VGAM::vglm(formula =  Sat ~ Infl*Type*Cont, family = multinomial(refLevel = "Low"),
                      data = datos5)
# Modelo sin interacciones (nivel de referencia "Low")
fit_red <- VGAM::vglm(formula = Sat ~ Infl+Type+Cont, family = multinomial(refLevel = "Low"),
                      data = datos5)
```

En el Cuadro 14 se muestran los resultados de la prueba de hipótesis (test *LRT*) en la que se analiza si es plausible utilizar el modelo reducido sobre el modelo completo:

$$H_0 : \text{Modelo reducido} \quad \text{vs.} \quad H_a : \text{Modelo completo}$$

Cuadro 14: Analysis of Deviance Table

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
3348	3470.084	NA	NA	NA
3314	3431.422	34	38.6622	0.2671363

Como se observa en el resultado, el p-value ($p = 0.267$) es mayor a un nivel de significancia de $\alpha = 0.05$. Por lo tanto, no se rechaza la hipótesis nula y podemos considerar como bueno el modelo reducido con un 95 % de confianza.

Por otro lado, en el cuadro 15 se muestran los puntajes obtenidos al calcular el AIC y BIC para cada modelo.

Cuadro 15: Puntajes de AIC y BIC para ambos modelos

	AIC	BIC
fit_red	3498.084	3574.064
fit_int	3527.422	3787.925

Esto nos ayuda a confirmar que el modelo reducido describe mejor el comportamiento de los datos, pues en ambas medidas tiene un puntaje menor.

iii) Ajuste de modelos logísticos multinomiales acumulativos

Hasta ahora, el análisis realizado se hizo considerando que la variable dependiente *Sat* es nominal, pero como además es una variable ordinal, se pueden ajustar modelos logísticos acumulativos.

```
# Modelos acumulativo sin supuesto de proporcionalidad
fit_acum <- VGAM::vglm(formula =  Sat ~ Infl+Type+Cont, family = cumulative(parallel = FALSE),
                       data = datos5)
# Modelos acumulativo con supuesto de proporcionalidad
fit_acum_prop <- VGAM::vglm(formula = Sat ~ Infl+Type+Cont, family = cumulative(parallel = TRUE),
                      data = datos5)
```

De la misma forma en que se hizo en el inciso ii, en el cuadro 16 se muestra el resultado de la prueba de hipótesis en la que se comparan ambos modelos.

Cuadro 16: Analysis of Deviance Table

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
3354	3479.149	NA	NA	NA
3348	3470.579	6	8.570599	0.199206

Como se observa en el resultado, el p-value ($p = 0.199$) es mayor a un nivel de significancia de $\alpha = 0.05$. Por lo tanto, no hay evidencia para rechazar la hipótesis nula y podemos considerar como bueno el modelo reducido (bajo el supuesto de proporcionalidad). Al calcular los puntajes de AIC y BIC para cada modelo, se obtiene lo que se presenta en el cuadro 17.

Cuadro 17: Puntajes de AIC y BIC para ambos modelos

	AIC	BIC
fit_acum_prop	3495.149	3538.566
fit_acum	3498.579	3574.559

Esto nos ayuda a confirmar que el modelo bajo el supuesto de proporcionalidad (fit_acum_prop) describe mejor el comportamiento de los datos pues sus puntajes fueron menores al del modelo fit_acum.

iv) Selección de modelo e interpretación de resultados

Como se observa en los cuadros 15 y 17, tanto el AIC como el BIC indican que el modelo fit_acum_prop es más adecuado a los datos que el modelo fit_red.

El modelo acumulativo proporcional se describe explícitamente de la siguiente manera:

$$logit\left[\mathbb{P}(Sat \leq j)\right] = \beta_0^{(j)} + \beta_1 I_{Medium} + \beta_2 I_{High} + \beta_3 T_{Atrium} + \beta_4 T_{Terrace} + \beta_5 T_{Tower} + \beta_6 C_{High}$$

Para $j = 1, 2$, donde 1 corresponde al nivel "Low", y 2 a "Medium". El valor estimado para cada coeficiente β_i se encuentra en el cuadro 18, donde sólo el asociado a β_0 (Intercept) cuenta con dos valores distintos, según la categoria correspondiente.

Cuadro 18: Estimación de los coeficientes betas del modelo.

	logitlink(P[Y<=1])	logitlink(P[Y<=2])
(Intercept)	0.0762152	1.2630586
InflMedium	-0.5663937	-0.5663937
InflHigh	-1.2888184	-1.2888184
TypeAtrium	-0.2061633	-0.2061633
TypeTerrace	0.5186648	0.5186648
TypeTower	-0.5723499	-0.5723499
ContHigh	-0.3602845	-0.3602845

En la figura 19 se muestran las probabilidades de que el modelo prediga cada nivel de satisfacción (de forma acumulativa) únicamente para los grupos de individuos bajo la restricción $Type$="Apartment" y $Cont$="High".

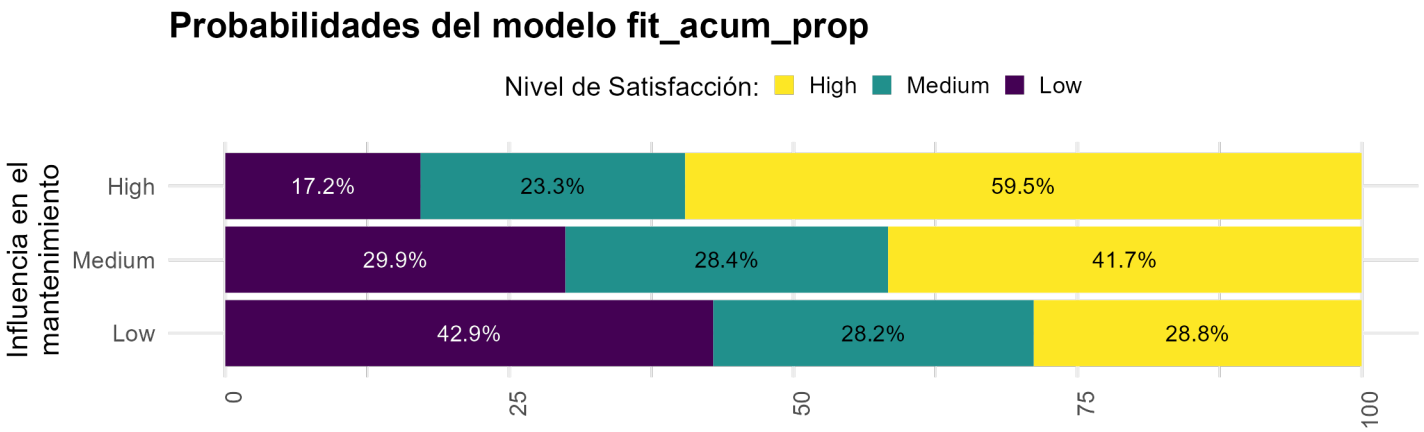

Figura 19: Resultados del modelo para los grupos con $Type$="Apartment" y $Cont$="High".

Haciendo una comparación directa entre estos resultados del modelo y el comportamiento de los datos observado en la figura 18 es posible notar que particularmente para este grupo de individuos las probabilidades predichas por el modelo son bastante similares a las frecuencias relativas de los datos. Esto indica que la suposición que se tenía de que la variable Infl es bastante influyente en el nivel de satisfacción era acertada, ya que como se observa, a mayor nivel en la variable de influencia en el mantenimiento, la probabilidad de satisfacción alta es mayor.