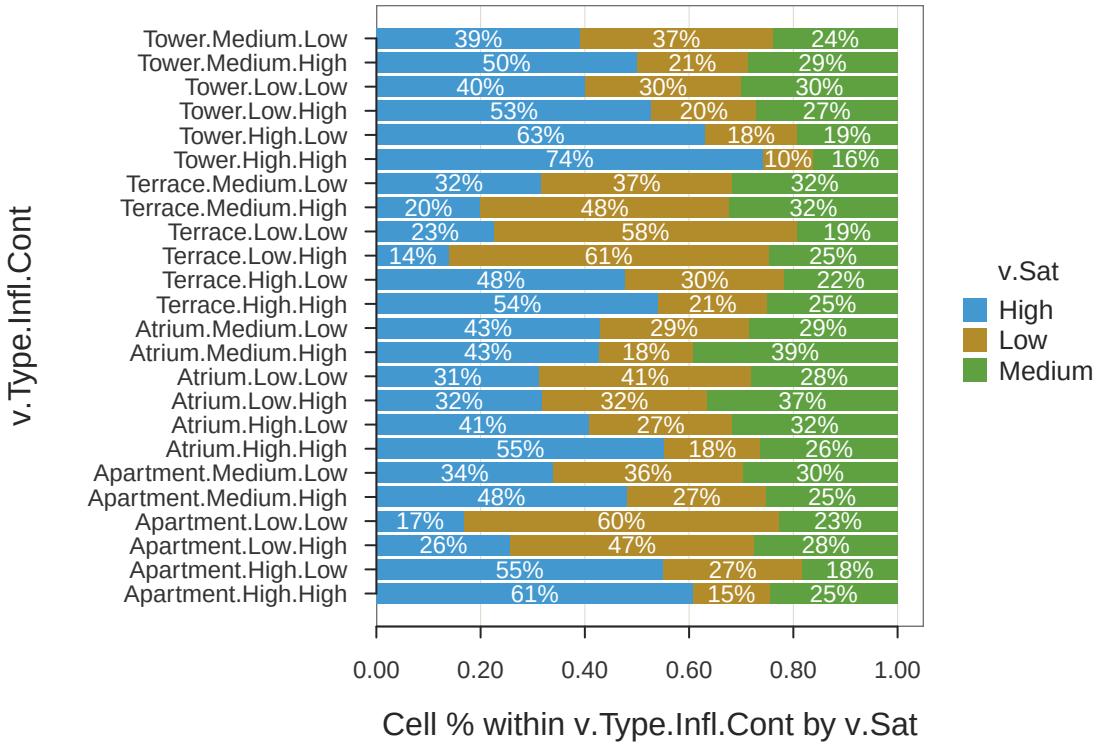
5. Modelos lineales generalizados para datos categóricos

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: apartment, atrium, terrace y tower), la percepción sobre su influencia en las decisiones sobre el mantenimiento de la vivienda (Infl: high, medium y low) y el contacto que tienen con el resto de inquilinos (Cont: high y low).

i) Gráfica de frecuencias relativas

Todas las covariables son categóricas, a continuación mostramos la gráfica que describe las frecuencias relativas para los tres niveles de satisfacción considerando cada cruce Type-Infl-Cont (en ese orden). Podemos observar que la mayor satisfacción (74%) se presenta en Tower.High.High, que se refiere a vivir en una torre, con alta influencia sobre el mantenimiento de la vivienda y con alto contacto con el resto de los inquilinos. Por otro lado, el menor nivel de satisfacción (14%) corresponde a vivir en una terraza, con baja influencia sobre el mantenimiento de la vivienda y con alto nivel de contacto con los demás inquilinos.

```
## >>> Note: v.Type.Infl.Cont is not in a data frame (table)
## >>> Note: v.Sat is not in a data frame (table)
```




```
## >>> Suggestions
## Plot(v.Type.Infl.Cont, v.Sat)  # bubble plot
## BarChart(v.Type.Infl.Cont, by=v.Sat, horiz=TRUE)  # horizontal bar chart
## BarChart(v.Type.Infl.Cont, fill="steelblue")  # steelblue bars
##
## Cramer's V: 0.252
##
## Chi-square Test of Independence:
##      Chisq = 213.070, df = 46, p-value = 0.000
```

ii) Modelo logístico multinomial nominal

Ajustamos varios modelos para la variable dependiente de satisfacción (Sat), considerando en un modelo completo las intereacciones de la influencia sobre mantenimiento (Infl), tipo de vivienda (Type) y contacto con otros vecinos (Cont) y no considerando estas interacciones en un modelo reducido. Luego hacemos uso de la función anova que nos permite realizar análisis de varianza entre los modelos ajustados, planteando las hiótesis H_0 : Podemos utilizar el modelo reducido contra H_a : Debemos utilizar el modelo completo. El p-value asociado a la prueba es menor a 0.05, por lo que a un nivel de confianza de 95%, no tenemos evidencia para rechazar la hipótesis nula, por lo tanto podemos usar el modelo reducido.

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

Por otra parte, podemos calcular las AIC de ambos modelos, el resultado es que el AIC del modelo completo o con interacciones es de 3527.4216616 y para el modelo reducido es de 3498.0838663. Por lo tanto, se tiene una mayor evidencia de que el modelo reducido es mejor por tener un menor AIC.

```
##
## Call:
## vglm(formula = Sat ~ Infl + Type + Cont, family = multinomial(refLevel = "Low"),
##     data = Datos)
##
## Coefficients:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept):1   1.2201     0.1585   7.699   0.0000000000000137 ***
## (Intercept):2   0.1709     0.1825   0.936             0.349200
## InflLow:1      -1.6126     0.1671  -9.649 < 0.0000000000000002 ***
## InflLow:2      -0.6649     0.1863  -3.568             0.000359 ***
## InflMedium:1   -0.8778     0.1641  -5.348   0.0000000890670893 ***
## InflMedium:2   -0.2185     0.1872  -1.167             0.243151
## TypeAtrium:1    0.3277     0.1886   1.737             0.082391 .
## TypeAtrium:2    0.5671     0.1947   2.913             0.003577 **
## TypeTerrace:1  -0.6767     0.1756  -3.854             0.000116 ***
## TypeTerrace:2  -0.2309     0.1748  -1.321             0.186653
## TypeTower:1     0.7356     0.1553   4.738   0.0000021614222276 ***
## TypeTower:2     0.4357     0.1725   2.525             0.011562 *
## ContLow:1      -0.4818     0.1241  -3.881             0.000104 ***
## ContLow:2      -0.3609     0.1324  -2.726             0.006420 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,2]), log(mu[,3]/mu[,2])
##
## Residual deviance: 3470.084 on 3348 degrees of freedom
##
## Log-likelihood: -1735.042 on 3348 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
```

```
##
## Reference group is level  2  of the response
```

iii) Modelo logístico acumulativo (cumulative logit) ordinal

Considerando las covariables del modelo reducido (sin interacciones) y la variable Sat como ordinal, ajustaremos un modelo logístico acumulativo (cumulative logit) sin considerar el supuesto de proporcionalidad (parallel) y otro asumiendo este supuesto. Dado que este último está anidado en el primero, realizaremos una prueba de hipótesis con la función anova para analizar si es plausible asumir este modelo más sencillo, donde planteamos las hipótesis H_0 : Podemos utilizar el modelo reducido contra H_a : Debemos utilizar el modelo completo. El modelo reducido es aquel que tiene probabilidades proporcionales, así que nos quedaremos con ese modelo, pues no hay suficiente evidencia para rechazar la hipótesis nula.

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

Por otro lado, el AIC del modelo logístico acumulativo sin supuesto de proporcionalidad es de 3498.5787001 y el AIC para el modelo logístico acumulativo con el supuesto de proporcionalidad es de 3495.1492991. Este menor AIC apoya la elección del modelo reducido logístico acumulativo con el supuesto de proporcionalidad.

```
##
## Call:
## vglm(formula = Sat ~ Infl + Type + Cont, family = cumulative(parallel = TRUE),
##     data = Datos)
##
## Coefficients:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept):1 -1.57289    0.12519 -12.564 < 0.0000000000000002 ***
## (Intercept):2 -0.38604    0.11938  -3.234             0.001222 **
## InflLow        1.28882    0.12670  10.172 < 0.0000000000000002 ***
## InflMedium     0.72242    0.12372   5.839        0.00000000524 ***
## TypeAtrium    -0.20616    0.13993  -1.473             0.140675
## TypeTerrace    0.51866    0.13358   3.883             0.000103 ***
## TypeTower     -0.57235    0.11875  -4.820        0.00000143628 ***
## ContLow        0.36028    0.09536   3.778             0.000158 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 3479.149 on 3354 degrees of freedom
##
## Log-likelihood: -1739.575 on 3354 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##     InflLow  InflMedium  TypeAtrium TypeTerrace   TypeTower     ContLow
```

```
##    3.6284964    2.0594206    0.8137002    1.6797833    0.5641981    1.4337373
```

iv) Selección del modelo e interpretación de resultados

Comparando los AIC de todos los modelos, elegimos el modelo logístico multinomial ordinal reducido acumulativo con proporcionalidad, que tiene el menor AIC de 3495.1492991. A continuación se presenta en una gráfica las probabilidades estimadas para cada nivel de satisfacción (Sat: low, medium y high) al considerar la variable de influencia sobre el mantenimiento (Infl) y el nivel de contacto con otros inquilinos (Cont: high y low), cuando se asume que la persona renta una vivienda tipo Apartment.

Podemos observar que la Gráfica de la columna izquierda (Cont=low), cuando se tiene contacto bajo con el resto de inquilinos, muestra las probabilidades de baja, media y alta satisfacción (Sat), considerando únicamente Apartment, y la influencia sobre el mantenimiento (Infl) bajo, medio y alto. En este caso, la probabilidad de baja satisfacción se asocia con mayor probailidad (52%) a Apartments con baja influencia sobre el mantenimiento y bajo contacto con los otras personas que habitan el lugar, y la mayor probabilidad (51%) se asocia con Apartments con alta influencia sobre el mantenimiento, a pesar del bajo contacto. Por otra parte, el la Gráfica de la columna derecha, se observa que hay una probabilidad de satisfacción muy alta (60%) asociada a Apartments donde hay una alta influencia en mantenimiento y alto contacto con los demás inquilinos del lugar, la probabilidad que la satisfacción sea baja en un Apartment de estas características es baja (17%).