

# Modelación del vino a través de los modelos lineales generalizados

Alejandro Gómez Montoya\*  
Yeisson Alexis Acevedo Agudelo†

---

*Keywords:* Vino, Regresión logística, Regresión Poisson, Análisis estadístico.

---

## 1 Introducción

Identificar las propiedades fisicoquímicas y sensoriales del vino resulta de gran utilidad a la hora de asignar una ponderación de calidad para el vino tinto y blanco. En la literatura se encuentran varios trabajos en los que se ha intentado modelar la calidad de este tipo de vinos en relación a pruebas analíticas disponibles en la etapa de certificación, tanto, a saber: características sensoriales [1] y propiedades fisicoquímicas [3]. La mayoría de los autores han usado las redes neuronales, minería de datos, regresión, lineal, logística, multinomial y máquinas de soporte vectores para tal fin [5]. Sin embargo, resulta más provechoso para la industria tener modelos no precisamente de tipo caja negra sino que permita una mejor descripción de la calidad en relación a los componentes del vino y la percepción que genera en sus consumidores.

En este mismo sentido, conviene resaltar que un modelamiento en relación a la multiplicidad categórica de las características del vino tinto y blanco, y en pro de brindar una estimación de su calidad, sería de gran utilidad para aportar las evaluaciones de enología y mejorar su producción [7], o mejor aún, estimar su viabilidad en la caracterización y nivel preferencial en el gusto del consumidor, minimizar el costo en producción [9] e inclusive articular aspectos propios de su comercialización. Claro está, un modelamiento con tales alcances no resulta fácil de estimar si se considera de entrada las interacciones altamente no lineales que en las pruebas de calidad se llevan a cabo. Cortez menciona que en la certificación del vino en relación a las pruebas fisicoquímicas se determinan densidad, alcohol y valores PH [3] mientras que autores como Mei-Yi, proponen la clasificación del vino como toda una ardua tarea y de difícil modelamiento ya que el gusto es uno de los sentidos más incomprendidos [10] ni qué decir si se considera el hecho que el tipo de vino, el tipo de azúcar y el pH conllevan hacia una acidez diversa en el vino preparado.

En relación a lo anterior, el presente trabajo tiene la intención de estimar un modelo de regresión logística con salida factor de calidad y un modelo regresión Poisson con salida factor acidez volátil. Respecto al primero se busca relacionar las variables entradas o explicativas (acidez fija AF, acidez volátil AV, Ácido cítrico AC, Azúcar residual AR, cloruros C, dióxido de azufre libre SO2L, dióxido de azufre total SO2T, densidad D, pH, sulfatos S, grado alcohólico GL) con la variable respuesta Calidad; respecto al segundo modelo, se busca estimar este de manera que relacione las entradas (calidad, tipo de vino, tipo de azúcar y pH) con la variable respuesta AV.

De esta manera resulta claro que el objetivo es aplicar técnicas para el modelado de variables respuestas categóricas, modelando la calidad del vino usando los resultados de las pruebas físico-químicas. La calidad se abordará entonces desde dos puntos de vistas: (i) a partir de una calificación sensorial, y (ii) verificando el número de vinos que satisfacen ciertos rangos preestablecidas para alguna de las variables físico-químicas.

El presente trabajo contiene, en la sección 2 el análisis exploratorio de los datos, tanto para los datos correspondientes a la regresión logística, como los datos correspondientes para la regresión poisson; en

---

\*agomez13@eafit.edu.co

†yaceved2@eafit.edu.co

Table 1: Características de la Base de Datos

	Media	Desviacion	Mediana	Minimo	Maximo	Coef.Variacion	Rango	Rango.Intercuartil
AF	7.22	1.30	7.00	3.80	15.90	0.18	12.10	1.30
AV	0.34	0.16	0.29	0.08	1.58	0.48	1.50	0.17
AC	0.32	0.15	0.31	0.00	1.66	0.46	1.66	0.14
AR	5.44	4.76	3.00	0.60	65.80	0.87	65.20	6.30
C	0.06	0.04	0.05	0.01	0.61	0.63	0.60	0.03
SO2L	30.53	17.75	29.00	1.00	289.00	0.58	288.00	24.00
SO2T	115.74	56.52	118.00	6.00	440.00	0.49	434.00	79.00
D	0.99	0.00	0.99	0.99	1.04	0.00	0.05	0.00
pH	3.22	0.16	3.21	2.72	4.01	0.05	1.29	0.21
S	0.53	0.15	0.51	0.22	2.00	0.28	1.78	0.17
GL	10.49	1.19	10.30	8.00	14.90	0.11	6.90	1.80

la sección 3 se realiza el ajuste, simplificación y análisis del modelo de regresión logística y el mismo procedimiento para la regresión poisson; en la sección 4 se encuentran las conclusiones obtenidas y por último en la sección 5 se presenta la bibliografía usada.

## 2 Análisis exploratorio de datos

### 2.1 Regresión Logística

El presente trabajo emplea una base de datos real que contiene 6497 observaciones [3] relacionadas con dos tipos de vino: tinto y blanco. En estas se obtuvo el indicador de calidad para el vino en un rango de referencia de 3 a 9, ésta medida es la que se considerá como salida de los modelos estimados, y esta dada por expertos en el area [3]. La prueba de calidad del vino se estructura en función de los análisis fisico-químicos, a saber: acidez fija, volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre y total, densidad, pH, sulfatos y grado alcohólico. Estos últimos serán considerados como entradas en los modelos estimados.

En la tabla 1, se muestran las principales medidas de tendencia central y dispersión de las variables involucradas en la modelación. Puede notarse en dicha tabla, que la variable de Densidad tiene una variación minima, esto puede indicar que esta variable no tiene cambios significativos en la base de datos, por lo cual está no contenga información relevante para la construcción del modelo de Calidad del Vino, esto puede contrastarse con la variable de Cloruros, dadoque esta tiene una variación pequeña, sin embargo su coeficiente de variación es significativo. Puede ser resaltado que el alto coeficiente de variación de la Azúcar Residual, puede ser un factor que introduzca errores al modelo.

Es importante estudiar el comportamiento de los datos, para determinar posibles multicolinealidades, por lo cual se analiza la dispersión de los datos, que se muestra en la Figura 1, en esta se representa con cada color una categoría de la calidad de vino, en el eje central el histograma para cada variable explicativa y con una linea negra en cada gráfico de dispersión se representa el valor medio de los datos. De este gráfico puede identificarse correlación entre las variables D-GL, SO2L - SO2T, SO2L - D, D - pH, lo cual puede indicar que puede encontrarse multicolinealidad y puede afectar el resultado del modelo.

Una exploración mas profunda se presenta en la figura ??, en la cual se calcula las densidades para cada variable en función de cada categoria, en color rojo se presenta el vino Blanco y en color Azul se presenta el vino Tinto, esta división se presenta dado que las propiedades de ambos tipos de vino pueden diferir significativamente [4].



Figure 1: Gráfico de Dispersion

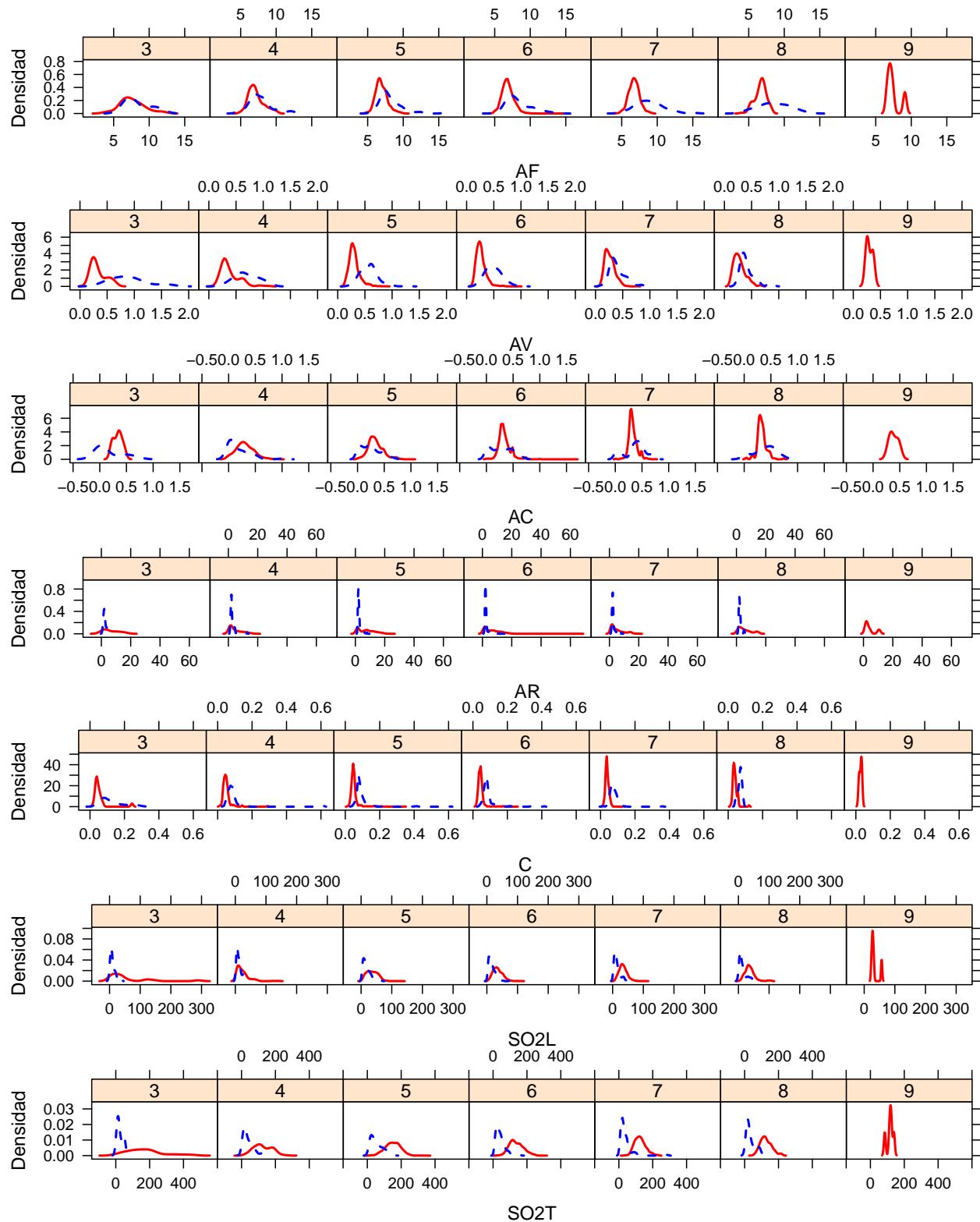


Table 2: Test de Medidas

	3	4	5	6	7	8	9
AF	7.853	7.289	7.327	7.177	7.129	6.835	7.420
AV	0.517	0.458	0.390	0.314	0.289	0.291	0.298
AC	0.281	0.272	0.308	0.324	0.335	0.333	0.386
AR	5.140	4.154	5.804	5.550	4.732	5.383	4.120
C	0.077	0.060	0.065	0.054	0.045	0.041	0.027
SO2L	39.217	20.637	30.237	31.165	30.422	34.534	33.400
SO2T	122.033	103.433	120.839	115.411	108.499	117.518	116.000
D	0.996	0.995	0.996	0.995	0.993	0.993	0.991
pH	3.258	3.232	3.212	3.218	3.228	3.223	3.308
S	0.506	0.506	0.526	0.533	0.547	0.512	0.466
GL	10.215	10.180	9.838	10.588	11.386	11.679	12.180

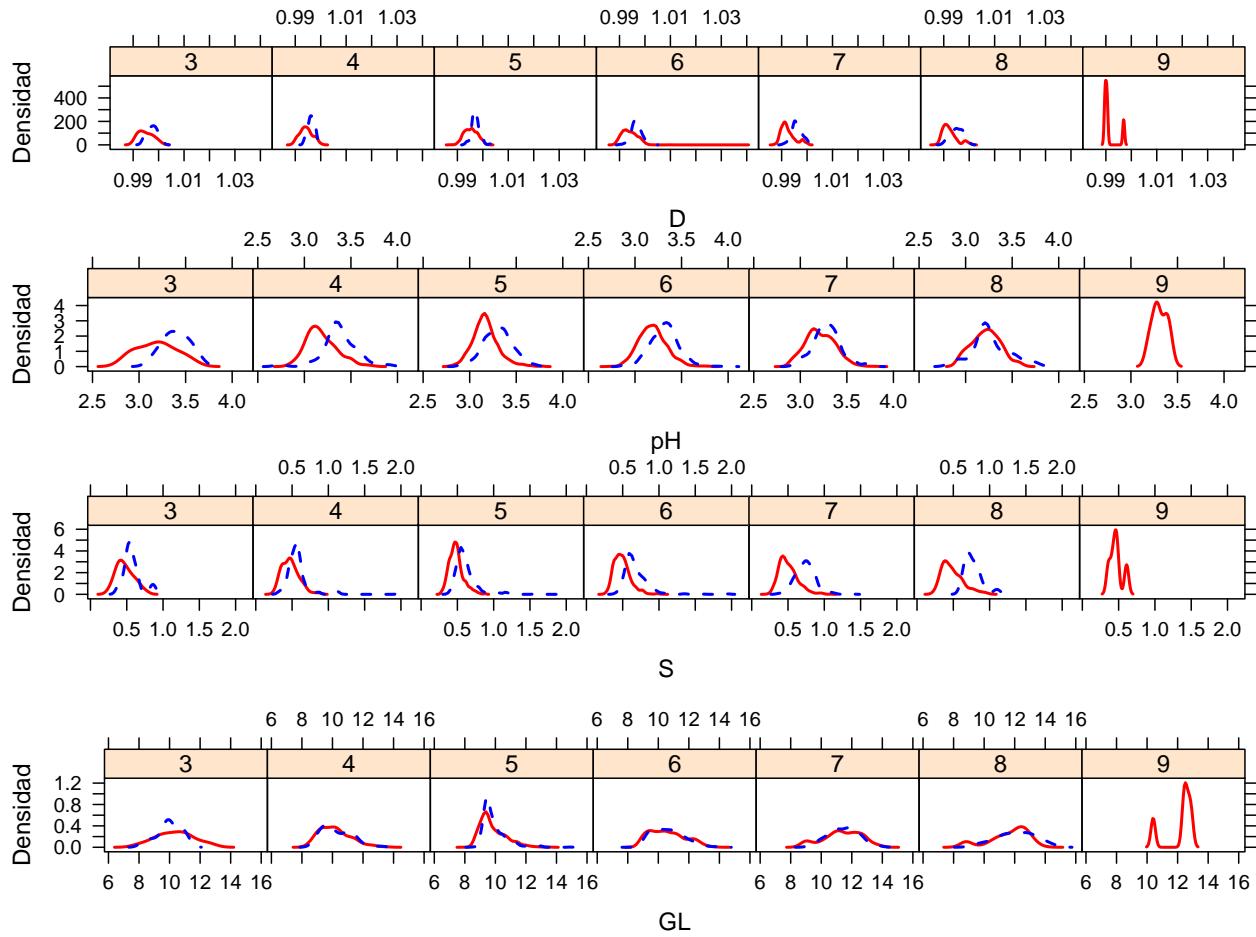


Figure 2: Gráfico de Densidades para cada Variable

Puede notarse que no se encuentran cambios significativos para las variables S, D, también se encuentran comportamientos semejantes de las variables SO2L y SO2T, esto puede indicar que se encuentran variables explicativas que no son realmente relevantes para el modelo de la calidad. Otra herramienta que se utilizó para analizar los datos, corresponde a un test de medias, para cada categoría, que se presenta en la Tabla 2.

Dicha tabla permite identificar que la variable D tiene una minima variación de la media, por lo cual no es recomendable utilizarla en el modelo, de igual forma se encuentran equivalencias para algunas categorias con la variable S.

## 2.2 Regresión Poisson

Para esta estimación, se considerará la misma base de datos real que contiene 6497 observaciones [3]. En este caso se busca estimar si el vino cumple con la Ácidez Volatil deseada, en función de la Calidad del Vino, el Tipo de Vino, el Tipo de Azúcar y si cumple con los requisitos de pH. Cada una de las variables utilizadas son categoricas, de igual forma se gráfica la dispersión de los datos, y se presenta en la Figura 3.

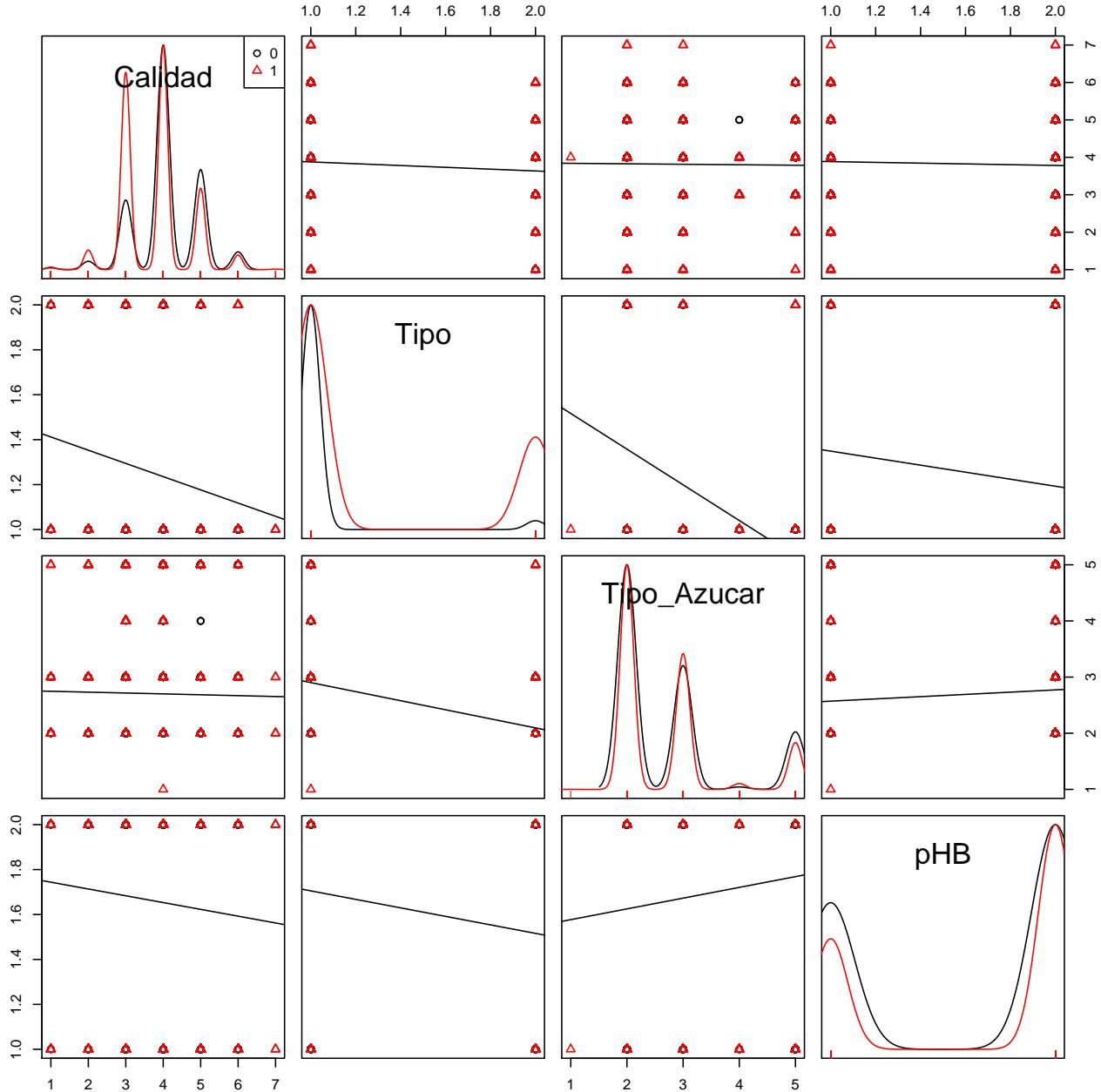


Figure 3: Gráfico de Dispersión

Son notorias las diferencias de cada variable para las variables de Calidad, el indice de pH y el Tipo de Vino, sin embargo no se encuentra un cambio significativo para el Tipo de Azucar, esto puede indicar que dicha variable no es significativa para el modelo. La figura 4 se representan la distribución de los datos para cada tipo de categoría, esta representación de la información corresponde a la tabla de contingencia de los datos y permite concluir que se cuentan con excasas muestras de vinos Dulces y Semidulces, del conjunto de los vinos Tintos. También puede identificarse una mayor proporción de vinos Blancos que Tintos, y permite identificar que variables tienen una mayor predominancia como los vinos extra secos y los vinos secos.

### 3 Ajuste del modelo

#### 3.1 Regresión Logistica

Como primer modelo se ajusta un GLM mediante regresión logística con la siguiente característica:

```
Calidad~AF+AV+AC+AR+C+S02L+S02T+D+pH+S+GL+Tipo
```

```
## Calidad ~ AF + AV + AC + AR + C + S02L + S02T + D + pH + S +
##       GL + Tipo
```

De dicho modelo se obtiene:

```
## [1] "Deviance: 13770.540644"
## [1] "AIC: 13926.540644"
```

Realizando un test para los coeficientes se encuentra un patrón, en el cual las variables AC, AR, SO2T, GL no son significativas para la mayoría de las categorías:

```
coeftest(m1)
```

```
##
## z test of coefficients:
##
##             Estimate Std. Error     z value Pr(>|z|)
## 4:(Intercept) -1.6859e+01 8.0051e-02 -2.1061e+02 < 2.2e-16 ***
## 4:AF          -5.2098e-01 1.1999e-01 -4.3417e+00 1.414e-05 ***
## 4:AV          -1.1453e+00 4.7283e-01 -2.4222e+00 0.0154248 *
## 4:AC          -2.4827e-02 5.1605e-01 -4.8100e-02 0.9616295
## 4:AR          -5.5632e-02 4.9718e-02 -1.1190e+00 0.2631543
## 4:C           -1.0747e+01 7.4314e-02 -1.4462e+02 < 2.2e-16 ***
## 4:S02L         -8.2787e-02 1.2214e-02 -6.7783e+00 1.216e-11 ***
## 4:S02T         7.3023e-03 5.3553e-03 1.3636e+00 0.1727059
## 4:D           3.2072e+01 8.0009e-02 4.0085e+02 < 2.2e-16 ***
## 4:pH          -2.0517e+00 3.7267e-01 -5.5054e+00 3.684e-08 ***
## 4:S            4.7465e+00 5.3094e-01 8.9399e+00 < 2.2e-16 ***
## 4:GL          -1.8619e-01 1.4696e-01 -1.2670e+00 0.2051730
## 4:TipoTinto   -9.5511e-02 3.8084e-01 -2.5080e-01 0.8019764
## 5:(Intercept) -3.1461e+01 3.7389e-01 -8.4144e+01 < 2.2e-16 ***
## 5:AF          -8.2522e-01 1.0912e-01 -7.5622e+00 3.962e-14 ***
## 5:AV          -4.8037e+00 3.1995e-01 -1.5014e+01 < 2.2e-16 ***
## 5:AC          7.2580e-01 2.7107e-01 2.6775e+00 0.0074173 **
## 5:AR          -2.6966e-02 4.5899e-02 -5.8750e-01 0.5568661
## 5:C           -1.3874e+01 5.3611e-01 -2.5879e+01 < 2.2e-16 ***
## 5:S02L         -4.8022e-02 9.8150e-03 -4.8927e+00 9.947e-07 ***
## 5:S02T         1.3200e-02 4.9801e-03 2.6505e+00 0.0080366 **
## 5:D           5.7482e+01 3.6753e-01 1.5640e+02 < 2.2e-16 ***
```

### Gráfico de Mosaico Regresión Poisson

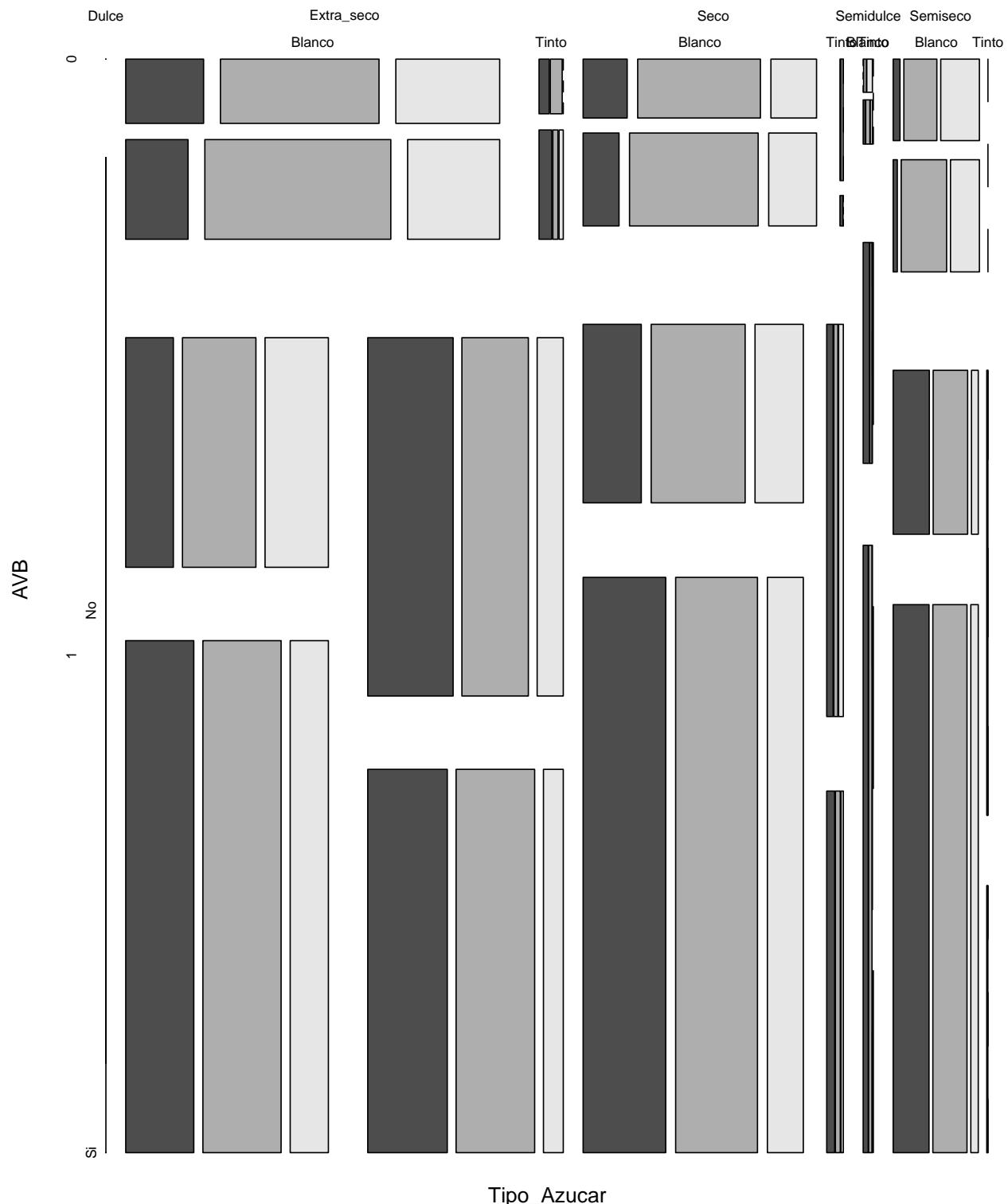


Figure 4: Gráfico de Mosaico

```

## 5:pH      -3.8636e+00 3.4934e-01 -1.1060e+01 < 2.2e-16 ***
## 5:S       4.4031e+00 2.6726e-01 1.6475e+01 < 2.2e-16 ***
## 5:GL      -3.1893e-01 1.3358e-01 -2.3875e+00 0.0169644 *
## 5:TipoTinto 3.5629e+00 2.7674e-01 1.2874e+01 < 2.2e-16 ***
## 6:(Intercept) -5.2933e+01 3.1151e-01 -1.6992e+02 < 2.2e-16 ***
## 6:AF      -8.4027e-01 1.0872e-01 -7.7286e+00 1.087e-14 ***
## 6:AV      -8.7961e+00 3.0896e-01 -2.8470e+01 < 2.2e-16 ***
## 6:AC      1.9357e-01 2.5197e-01 7.6820e-01 0.4423424
## 6:AR      2.3756e-02 4.5844e-02 5.1820e-01 0.6043263
## 6:C       -1.4876e+01 5.3103e-01 -2.8014e+01 < 2.2e-16 ***
## 6:S02L    -3.5395e-02 9.6874e-03 -3.6537e+00 0.0002585 ***
## 6:S02T    6.8334e-03 4.9745e-03 1.3737e+00 0.1695357
## 6:D       7.1138e+01 3.0635e-01 2.3222e+02 < 2.2e-16 ***
## 6:pH      -3.6483e+00 3.3432e-01 -1.0913e+01 < 2.2e-16 ***
## 6:S       5.9752e+00 2.2906e-01 2.6086e+01 < 2.2e-16 ***
## 6:GL      5.0855e-01 1.3193e-01 3.8548e+00 0.0001158 ***
## 6:TipoTinto 3.6147e+00 2.6524e-01 1.3628e+01 < 2.2e-16 ***
## 7:(Intercept) 9.8628e+01 4.0448e-01 2.4384e+02 < 2.2e-16 ***
## 7:AF      -5.4054e-01 1.1205e-01 -4.8243e+00 1.405e-06 ***
## 7:AV      -1.1333e+01 4.0247e-01 -2.8157e+01 < 2.2e-16 ***
## 7:AC      -1.6282e-01 3.3314e-01 -4.8870e-01 0.6250231
## 7:AR      1.2482e-01 4.6599e-02 2.6786e+00 0.0073922 **
## 7:C       -2.7290e+01 3.1979e-02 -8.5336e+02 < 2.2e-16 ***
## 7:S02L    -2.9581e-02 9.9163e-03 -2.9830e+00 0.0028540 **
## 7:S02T    4.4673e-03 5.0763e-03 8.8000e-01 0.3788411
## 7:D       -9.4464e+01 3.9832e-01 -2.3716e+02 < 2.2e-16 ***
## 7:pH      -2.1279e+00 3.5592e-01 -5.9785e+00 2.252e-09 ***
## 7:S       7.8059e+00 2.7067e-01 2.8840e+01 < 2.2e-16 ***
## 7:GL      9.6402e-01 1.3458e-01 7.1633e+00 7.877e-13 ***
## 7:TipoTinto 3.9392e+00 2.8483e-01 1.3830e+01 < 2.2e-16 ***
## 8:(Intercept) 2.9018e+01 6.7096e-02 4.3249e+02 < 2.2e-16 ***
## 8:AF      -6.7797e-01 1.2551e-01 -5.4015e+00 6.609e-08 ***
## 8:AV      -1.0565e+01 6.6774e-01 -1.5821e+01 < 2.2e-16 ***
## 8:AC      4.0026e-01 5.6559e-01 7.0770e-01 0.4791362
## 8:AR      1.4623e-01 4.8839e-02 2.9941e+00 0.0027529 **
## 8:C       -3.0718e+01 1.3719e-02 -2.2392e+03 < 2.2e-16 ***
## 8:S02L    -1.4735e-02 1.0516e-02 -1.4012e+00 0.1611599
## 8:S02T    2.2055e-03 5.6146e-03 3.9280e-01 0.6944584
## 8:D       -2.8293e+01 6.7343e-02 -4.2014e+02 < 2.2e-16 ***
## 8:pH      -2.3529e+00 3.8878e-01 -6.0520e+00 1.431e-09 ***
## 8:S       7.0574e+00 4.7444e-01 1.4875e+01 < 2.2e-16 ***
## 8:GL      1.2884e+00 1.4548e-01 8.8561e+00 < 2.2e-16 ***
## 8:TipoTinto 3.5532e+00 3.8680e-01 9.1861e+00 < 2.2e-16 ***
## 9:(Intercept) -2.6050e+01 4.4244e-02 -5.8877e+02 < 2.2e-16 ***
## 9:AF      8.4079e-01 2.4999e-01 3.3632e+00 0.0007704 ***
## 9:AV      -5.5668e+00 9.3393e-03 -5.9606e+02 < 2.2e-16 ***
## 9:AC      -9.7352e-01 1.3570e-02 -7.1739e+01 < 2.2e-16 ***
## 9:AR      1.7277e-01 8.3738e-02 2.0632e+00 0.0390908 *
## 9:C       -1.1916e+02 3.6683e-03 -3.2483e+04 < 2.2e-16 ***
## 9:S02L    -1.8613e-02 4.0568e-02 -4.5880e-01 0.6463652
## 9:S02T    -5.7108e-03 1.7155e-02 -3.3290e-01 0.7392196
## 9:D       -3.0323e+01 4.4765e-02 -6.7737e+02 < 2.2e-16 ***
## 9:pH      1.1054e+01 2.5371e-01 4.3568e+01 < 2.2e-16 ***
## 9:S       4.4692e-01 4.7228e-02 9.4630e+00 < 2.2e-16 ***

```

```

## 9:GL          1.7191e+00  1.5751e-01  1.0914e+01 < 2.2e-16 ***
## 9:TipoTinto   -2.3032e+01  9.4965e-11 -2.4253e+11 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Luego de simplificar el modelo, se plantea el modelo final como:

```
Calidad ~ AF + AV + C + SO2L + pH + S + Tipo
```

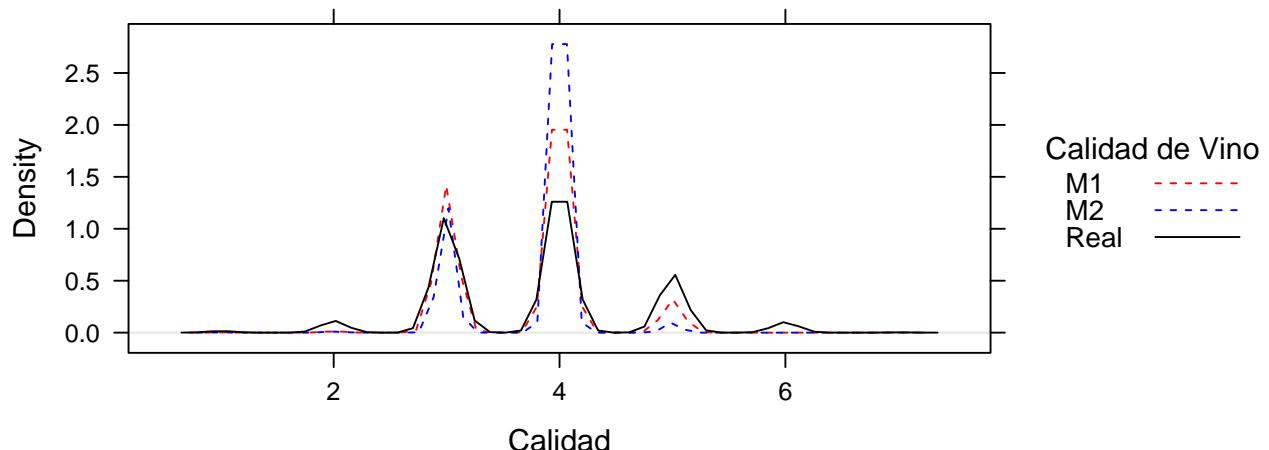
```
## Calidad ~ AF + AV + C + SO2L + pH + S + Tipo
```

De dicho modelo se obtiene:

```
## [1] "Deviance: 15256.372197"
```

```
## [1] "AIC: 15352.372197"
```

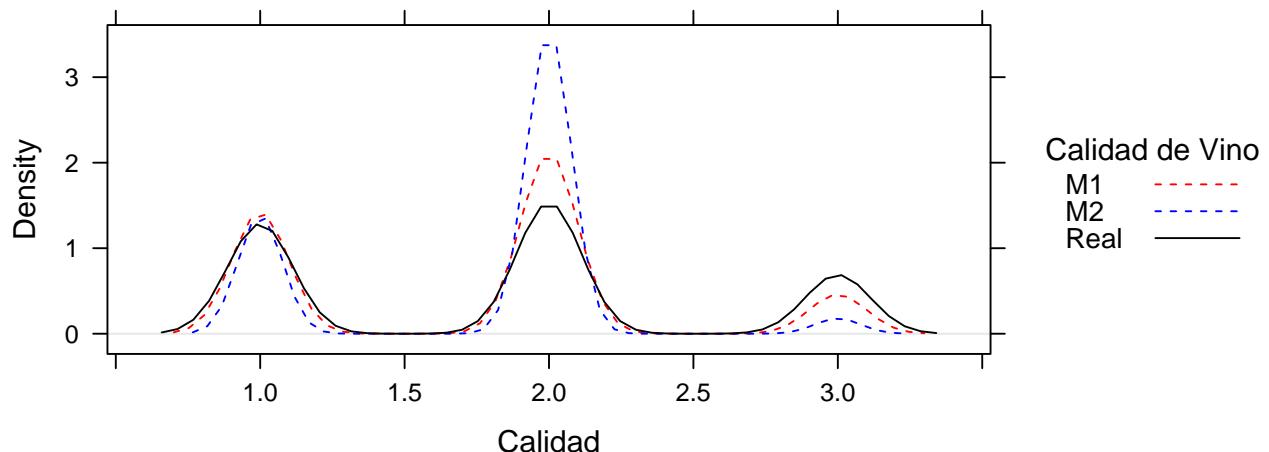
En la figura ??, puede encontrarse la densidad real y las dos estimaciones de los modelos. Puede notarse que las categorías con menor cantidad de muestras tienen una mala estimación, mientras las categorias con mayor cantidad de muestras tienen una sobreestimación, esto podria corregirse balanceando la base de datos, sin embargo serian necesarias una mayor cantidad de muestras.



Sin embargo si se realiza una transformación de la variable Calidad a 3 categorias:

1. 3, 4, 5
2. 6
3. 7, 8, 9

Que pueden representarse como Bueno, Regular y Malo, se obtiene el siguiente resultado:



Es posible evaluar los resultados obtenidos mediante el uso de una matriz de confusión:

```
##           Reference
## Prediction   1    2    3
##           1 1502  709  94
##           2  849 1841 756
##           3   33  286 427
```

Es notorio que las categorías 1 y 2 tienen una buena estimación sin embargo la categoría 3 al estar desbalanceada tiene una estimación menor. En la figura 5, se encuentran los residuales obtenidos del modelo simplificado, cada color representa una categoría estimada, los residuales de color rojo representan la categoría 3, los de color verde la categoría 2 y los de color negro la categoría 1. Como era de esperarse, los residuales de la categoría 3 tienen una magnitud mayor, dado que es la categoría con la menor cantidad de muestras, y es la que presenta mayor cantidad de errores de Tipo I y Tipo II. Sin embargo, analizando la matriz de confusión, la estimación de las densidades, y el orden de los residuales, puede inferirse que el modelo obtenido tiene un buen nivel de ajuste.

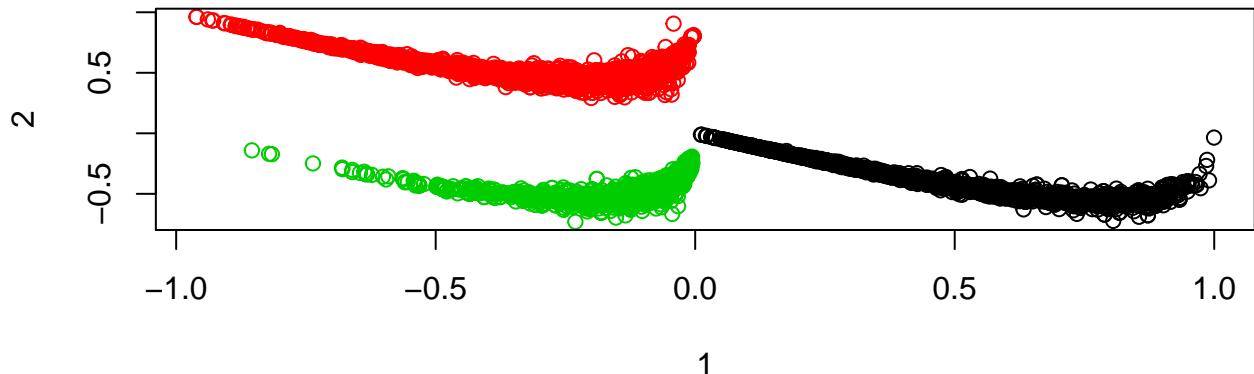


Figure 5: Residuales

### 3.2 Regresión Poisson

Se crea un modelo con la función:

```
Freq ~ Calidad+Tipo+Tipo_Azucar+pHB
```

```
## Freq ~ Calidad + Tipo + Tipo_Azucar + pHB
```

Donde se obtiene:

```
summary(m3)
```

```
##
## Call:
## glm(formula = Freq ~ Calidad + Tipo + Tipo_Azucar + pHB, family = poisson(link = "log"),
##      data = X2)
##
## Deviance Residuals:
##       Min        1Q     Median        3Q       Max
## -16.6181  -1.9496  -0.4785  -0.0479  25.3138
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.42829   1.01656  -7.307 2.73e-13 ***
##
```

```

## Calidad4          1.97408   0.19484  10.132 < 2e-16 ***
## Calidad5          4.26643   0.18385  23.206 < 2e-16 ***
## Calidad6          4.54895   0.18354  24.785 < 2e-16 ***
## Calidad7          3.58259   0.18509  19.355 < 2e-16 ***
## Calidad8          1.86149   0.19625   9.485 < 2e-16 ***
## Calidad9          -1.79176   0.48305  -3.709 0.000208 ***
## TipoTinto         -1.11945   0.02880  -38.867 < 2e-16 ***
## Tipo_AzucarExtra_seco 8.17216   1.00008   8.172 3.04e-16 ***
## Tipo_AzucarSeco      7.65255   1.00017   7.651 1.99e-14 ***
## Tipo_AzucarSemidulce 4.40672   1.00601   4.380 1.18e-05 ***
## Tipo_AzucarSemiseco  6.64249   1.00059   6.639 3.17e-11 ***
## pHBSi              0.65711   0.02616  25.115 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 24863.5 on 279 degrees of freedom
## Residual deviance: 5850.6 on 267 degrees of freedom
## AIC: 6390.7
##
## Number of Fisher Scoring iterations: 7

```

Puede notarse que cada categoría en general es significativa a excepción de valor de Calidad 5. Es por esto que se decide dejar el modelo con todas las variables explicativas. Los coeficientes son consistentes con lo encontrado en la base de datos, donde se tienen una mayor cantidad de vinos con una Calidad de 5, 6 y 7, en comparación con los vinos con Calidad 3. Puede notarse el mismo comportamiento para las variables del Tipo de Azúcar. Se calculan algunos coeficientes adicionales:

```

## [1] "Deviance: 19012.847693"
## [1] "R Deviance: 0.764690"

```

Es notorio el alto resultado R Deviance, en los datos que se presentan a continuación, pueden identificarse las categorías con un peso mayor para la determinación del resultado del modelo propuesto, esto puede deberse a que estas categorías tienen una mayor probabilidad de cumplir con la Acidez Voltatil deseada.

```
exp(cbind(RR=coef(m3), confint(m3)))
```

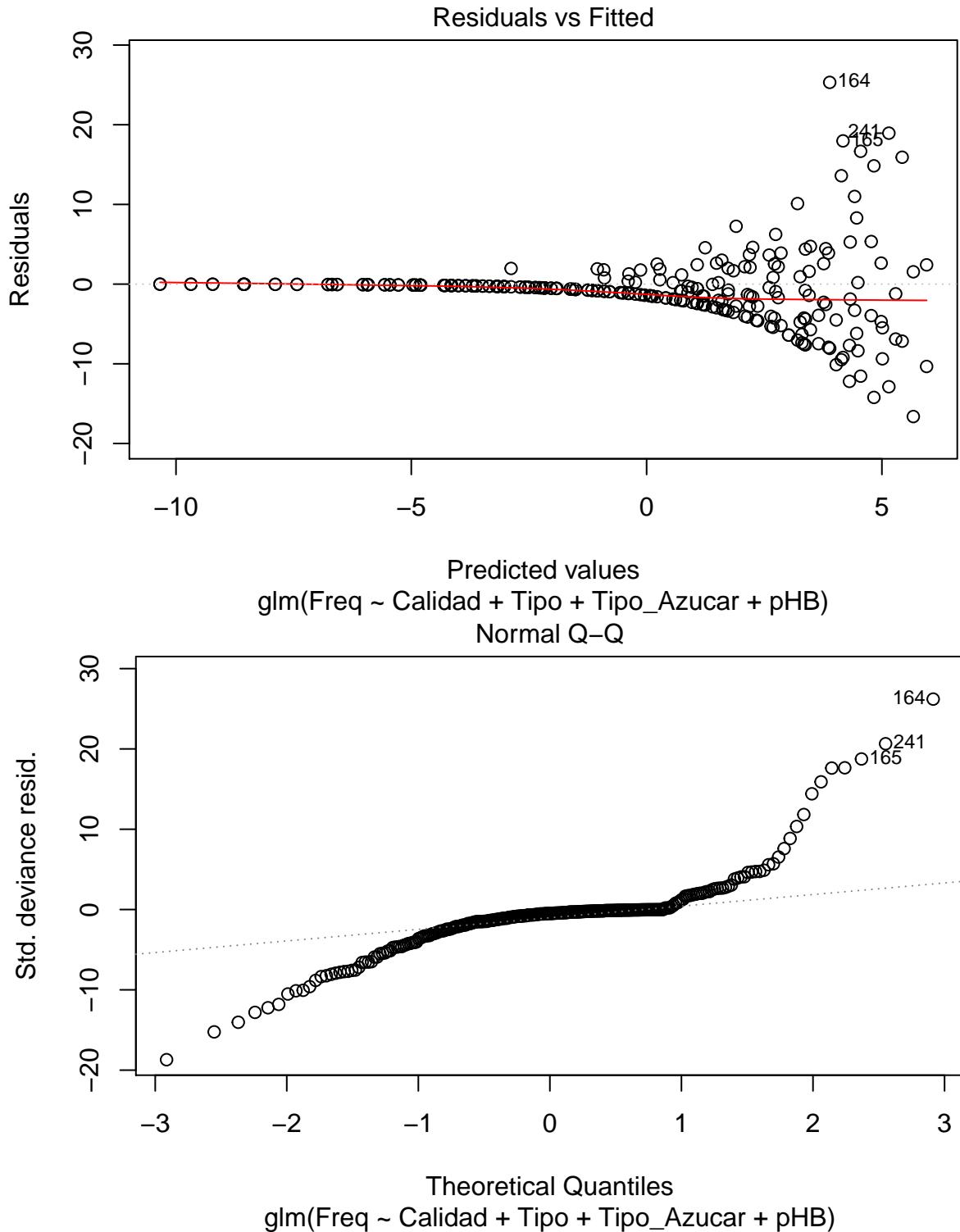
```

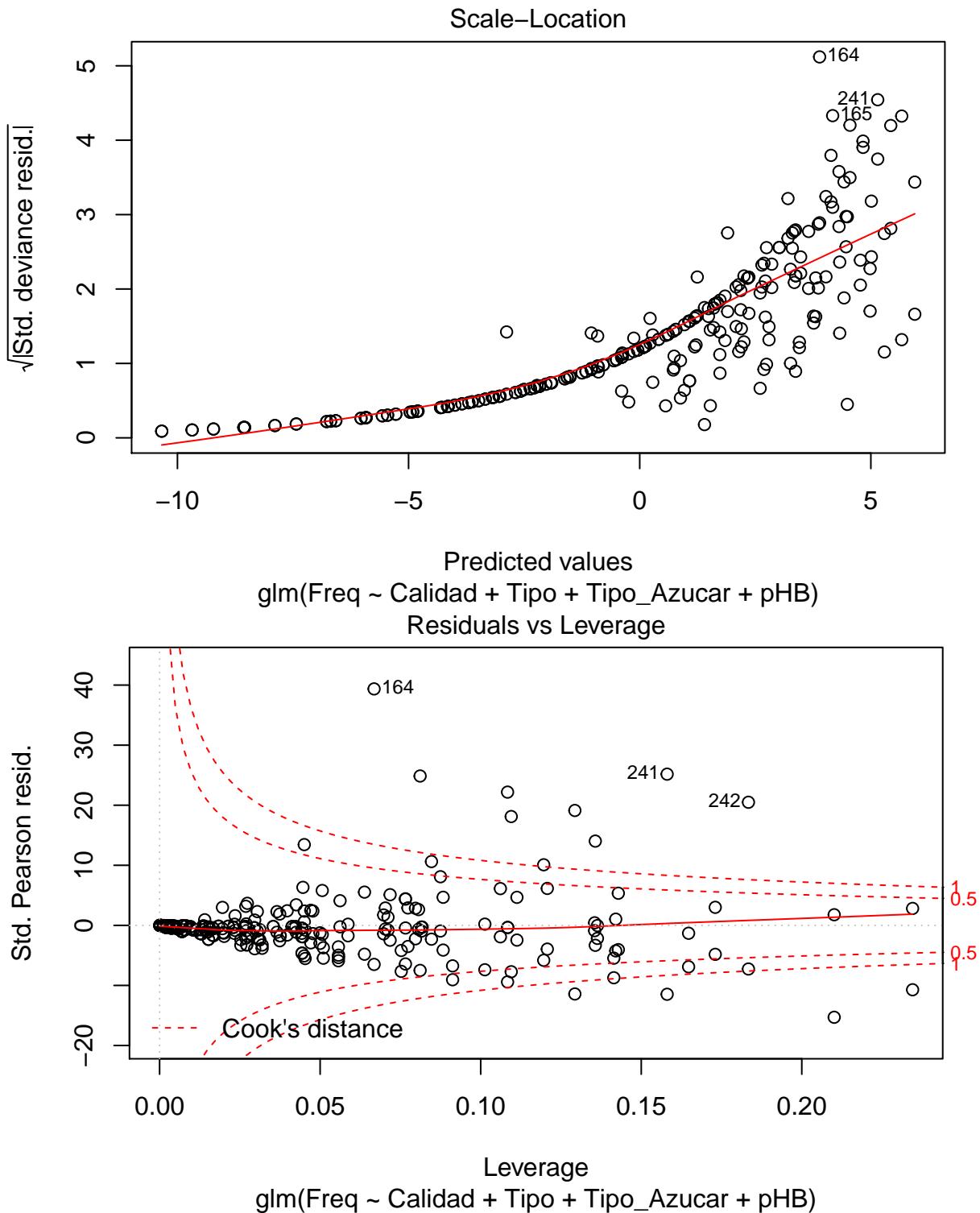
## Waiting for profiling to be done...
##                               RR      2.5 %     97.5 %
## (Intercept)      5.942005e-04 3.334773e-05 2.758711e-03
## Calidad4        7.200000e+00 4.999106e+00 1.075951e+01
## Calidad5        7.126667e+01 5.069979e+01 1.044986e+02
## Calidad6        9.453333e+01 6.729837e+01 1.385394e+02
## Calidad7        3.596667e+01 2.551693e+01 5.285173e+01
## Calidad8        6.433333e+00 4.452825e+00 9.637163e+00
## Calidad9        1.666667e-01 5.679426e-02 3.932694e-01
## TipoTinto       3.264598e-01 3.084567e-01 3.453269e-01
## Tipo_AzucarExtra_seco 3.541000e+03 8.046127e+02 6.209726e+04
## Tipo_AzucarSeco  2.106000e+03 4.783826e+02 3.693556e+04
## Tipo_AzucarSemidulce 8.200000e+01 1.825389e+01 1.446154e+03
## Tipo_AzucarSemiseco 7.670000e+02 1.739772e+02 1.345716e+04
## pHBSi           1.929216e+00 1.833028e+00 2.031016e+00

```

Por último se realiza el análisis de residuales del modelo propuesto, pueden encontrarse algunas muestras “outliers”, las cuales pueden corresponder a datos atípicos, que pueden ser resultado de variables cualitativas

como la Calidad del vino, sin embargo en términos generales se encuentra que el modelo tiene un buen ajuste del modelo, como lo comprueba el R Deviance obtenido.





## 4 Conclusiones

El presente trabajo nos permite concluir, que es fundamental el análisis exploratorio de los datos, dado que estas labores exploratorias ayudan el argumento de la eliminación de una variable explicativa en un modelo,

como fue el caso del primer modelo, o el análisis del modelo resultante como fue el caso del segundo modelo, dado que el análisis para cada caso de las categorías, permite identificar que categorías no contienen suficiente información para la construcción de un modelo, o puedan apoyar decisiones para la retención de variables categóricas. Esta exploración permite la toma de decisiones para la transformación de algunas variables como se realizó en el primer modelo al reducir el número de categorías para la variable respuesta, permitiendo balancear las categorías.

## Bibliografía

- [1] R. Ferrarini, C. Carbognin, E. M. Casarotti, E. Nicolis, A. Nencini, and A. M. Meneghini, “The emotional response to wine consumption,” *Food Quality and Preference*, vol. 21, no. 7, pp. 720–725, 2010 [Online]. Available: <http://dx.doi.org/10.1016/j.foodqual.2010.06.004>
- [2] N. Wariishi, B. Flanagan, T. Suzuki, and S. Hirokawa, “Sentiment Analysis of Wine Aroma,” *Proceedings - 2015 IIAI 4th International Congress on Advanced Applied Informatics, IIAI-AAI 2015*, pp. 207–212, 2016.
- [3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009 [Online]. Available: <http://dx.doi.org/10.1016/j.dss.2009.05.016>
- [4] X. Wang and Z. Guan, “Evaluation Model of Grape Wine Quality Based on BP Neural Network,” no. 71273123, 2016.
- [5] Z. Song, T. Liu, and S. Bai, “Modeling based on the effects of grapes for wine,” *Proceedings - 2014 IEEE Workshop on Electronics, Computer and Applications, IWECA 2014*, pp. 948–951, 2014.
- [6] L. Dooley, R. T. Threlfall, and J. F. Meullenet, “Optimization of blended wine quality through maximization of consumer liking,” *Food Quality and Preference*, vol. 24, no. 1, pp. 40–47, 2012 [Online]. Available: <http://dx.doi.org/10.1016/j.foodqual.2011.08.010>
- [7] S. Lee, J. Park, and K. Kang, “Assessing wine quality using a decision tree,” pp. 10–12, 2015.
- [8] P. Cortez and J. Teixeira, “Using Data Mining for Wine Quality Assessment,” pp. 66–79, 2009.
- [9] A. M. Johansen, A. L. Mumma, and H. C. Pinkart, “Cost efficient prediction of Cabernet Sauvignon wine quality,” 2016.
- [10] M.-y. Wu, J.-h. Lee, and S.-w. Kuo, “A Hierarchical Feature Search Method for Wine Label Image Recognition,” pp. 568–572, 2015.