Tarea 2: Modelos lineales generalizados y paramétricos

Angie Rodriguez Duque & Cesar Saavedra Vanegas

Octubre 22 de 2020

Actividad 1

Se dispone de los tiempos de vida (tiempos hasta que fallan, en horas) de 49 recipientes de presión sometidos a un nivel de carga del 70%

Distribución Weibull

Para estudiar este tipo de variable se acostumbra utilizar la distribución de Weibull, cuya función de densidad es:

$$f(y;\lambda,\theta) = \frac{\lambda y^{\lambda-1}}{\theta^{\lambda}} exp\left[-\left(\frac{y}{\theta}\right)^{\lambda}\right]$$

El método de Newton-Raphson

Para analizar este problema usaremos el método de Newton (conocido también como el método de Newton Raphson o el método de Newton Fourier), el cual es un algoritmo basado en la derivada que permite encontrar aproximaciones de los ceros o raíces de una función real derivable. En este caso particular se hara uso de la función U de scoring para la Weibul y se asumira λ conocido y U será el estimador $\hat{\theta}$ del parametro de escala θ.

1. Se elige x_0 en el eje de las x, asumiendo que está cerca de la solución de $f(x) = 0$ (raíz buscada)
2. Calculamos la ecuación punto pendiente de la recta tangente de la función en $(x_0, f(x_0))$, a saber $y - f(x_0) = f'(x_0)(x - x_0)$ (1).
3. Esta recta debe intersecar al eje de las x, en un punto más cercano a la raíz buscada; en el punto $(x_1, 0)$.
4. El punto $(x_1, 0)$ satisface la ecuación (1) y sustituyendo queda:

$$0 - f(x_0) = f'(x_0)(x - x_0) \qquad (2)$$

5. Si $f'(x_0) \neq 0$, entonces despejando x_i en (2) se obtiene:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

6. Repetimos el procedimiento anterior para x_0, pero ahora comenzando en x_1, en cuyo caso se obtiene:

$$x_2 = x_1 - \frac{f(x_0)}{f'(x_0)}$$

De forma que x_2 está más cerca de la raíz buscada que x_1.

7. Iterando cada vez con el número obtenido, se construye una secuencia: $x_0, x_1, x_2, ..., x_n$ de números cada vez más próximos a la raíz tales que:

$$x_{n+1} = x_n - \frac{f(x_0)}{f'(x_0)} \qquad (3)$$

Como asumimos que conocemos λ, la solución de $U = 0$ será el estimador $\hat{\theta}$ del parámetro de escala θ.

Como en este caso estamos asumiendo que $\lambda = 2$, entonces la expresión anterior será aplicable en este caso.

Nos proponemos encontrar θ tal que $U = 0$

Interacción	$\lambda = 1$	$\lambda = 1.5$	$\lambda = 2$	$\lambda = 2.5$	$\lambda = 3$
1	388.5787	388.5787	388.5787	388.5787	388.5787
2		408.2362	420.6631	428.5779	433.5102
3		410.7335	428.6631	443.3022	454.9522
4		410.7684	429.1123	444.8338	458.7434
5		410.7684	429.1133	444.8485	458.7434
6			429.1133	444.8485	458.7434

Actividad 2

Base de datos

```
dim(Datos)
```

```
## [1] 1599   12
```

Este conjunto de datos de vino tinto consta de 1599 observaciones y 12 variables, 11 de las cuales son sustancias químicas. Las variables son:

1. **Acidez fija:** La mayoría de los ácidos implicados en el vino son fijos o no volátiles (no se evaporan fácilmente).

2. **Acidez volátil:** La cantidad de ácido acético en el vino, que en niveles demasiado altos puede provocar un sabor desagradable a vinagre.

3. **Ácido cítrico:** Encontrado en pequeñas cantidades, el ácido cítrico puede agregar "frescura" y sabor a los vinos.

4. **Azúcar residual:** Es la cantidad de azúcar que queda después de que se detiene la fermentación, es raro encontrar vinos con menos de 1 gramo / litro y los vinos con más de 45 gramos / litro se consideran dulces.

5. **Cloruros:** Es la cantidad de sal del vino.

6. **Dióxido de azufre libre:** La forma libre de SO_2 existe en equilibrio entre el SO_2 molecular (como gas disuelto) y el ion bisulfito; Previene el crecimiento microbiano y la oxidación del vino.

7. **Dióxido de azufre total:** Es la cantidad de formas libres y unidas de SO_2; en concentraciones bajas, el SO_2 es mayormente indetectable en el vino, pero en concentraciones de SO_2 libre superiores a 50 ppm, el SO_2 se hace evidente en la nariz y el sabor del vino.

8. **Densidad:** La densidad es cercana a la del agua dependiendo del porcentaje de alcohol y contenido de azúcar.

9. **pH:** Describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); la mayoría de los vinos están entre 3-4 en la escala de pH.

10. **Sulfatos:** Aditivo del vino que puede contribuir a los niveles de dióxido de azufre (SO_2), que actúa como antimicrobiano y antioxidante.
11. **Alcohol:** El porcentaje de contenido de alcohol del vino.
12. **Calidad:** Variable de respuesta (basada en datos sensoriales, puntuación entre 0 y 10).

Estadísticas descriptivas

```
summary(Datos)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates          alcohol          quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

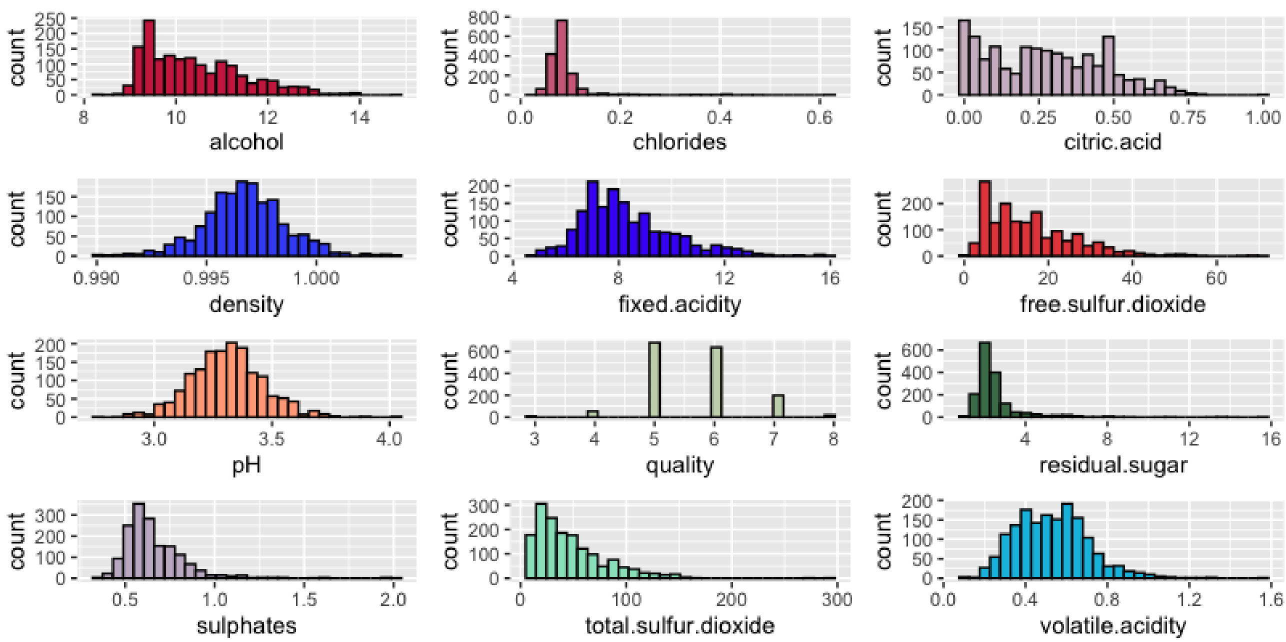
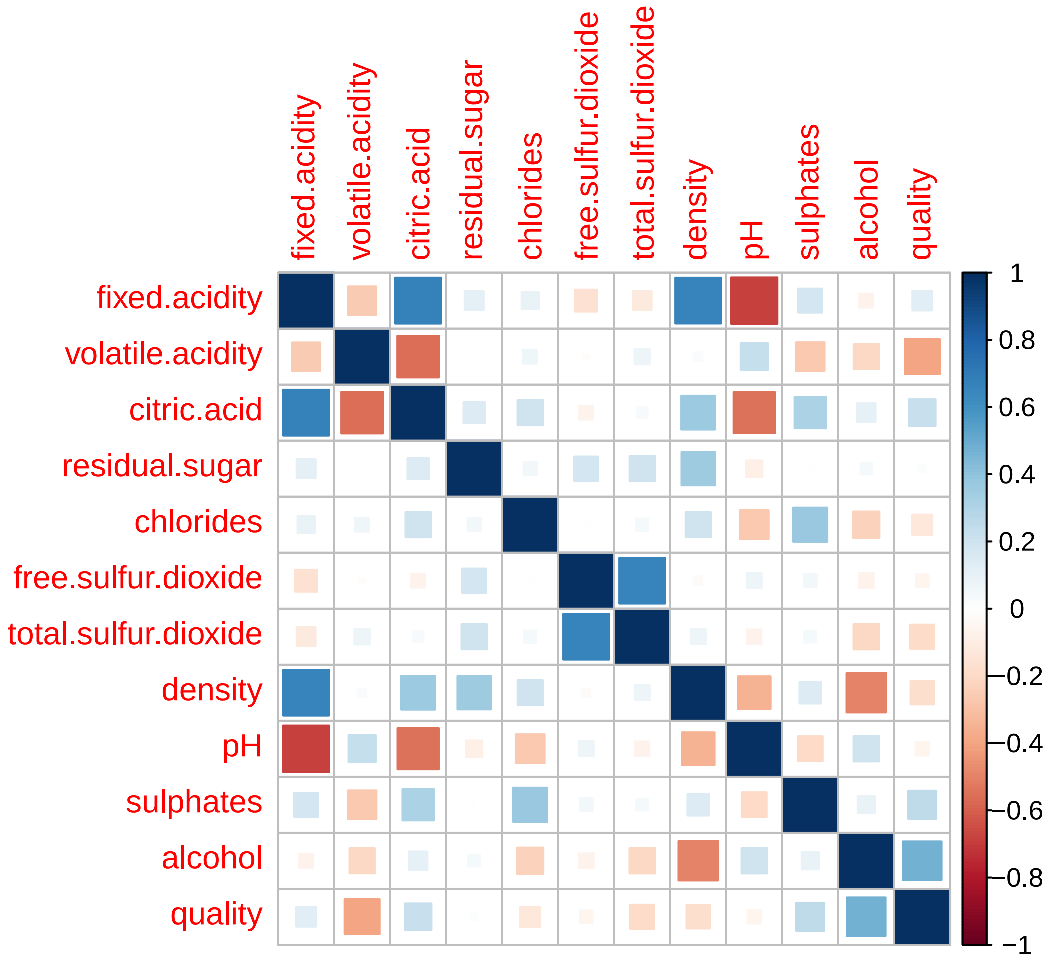


Figure 1: Distribución de las variables

Observaciones:

- Algunas de las variables tienen distribuciones normales (densidad, acidez fija, pH, acidez volátil).
- Algunas variables están un poco sesgadas hacia el extremo inferior de los valores (cloruros, ácido cítrico, azúcar residual, dióxido de azufre total).
- La variable calidad tiene solo 6 valores discretos.

```
corrplot(cor(Datos), method = "square")
```



- La densidad tiene una correlación muy fuerte con la acidez fija.
- Las variables más fuertemente correlacionadas con la calidad son la acidez volátil y el alcohol.
- El alcohol tiene una correlación negativa con la densidad. Esto es evidente por el hecho de que la densidad del agua es mayor que la densidad del alcohol.

Variable indicadora: pHi

Se convierte la variable "pH" en una variable indicadora con tres niveles: "alto", "medio" y "bajo", esta nueva variable se denomina: "pHi". Para dicha transformación se realiza el siguiente procedimiento:

$$Rango = \frac{Máx(pH) - Mín(pH)}{3}$$

$$Rango = \frac{4.01 - 2.74}{3} = 0.4233333$$

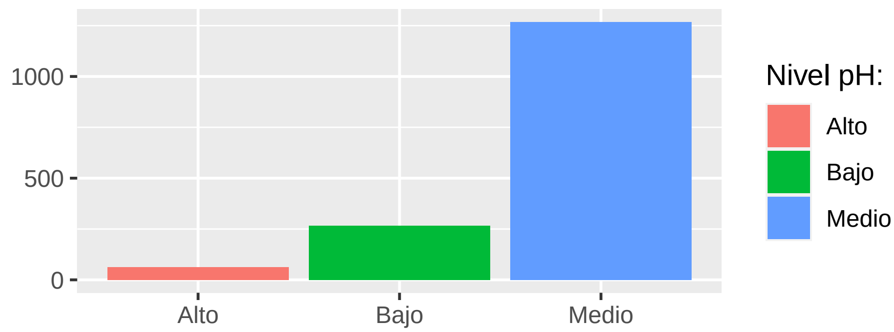
De esta manera, los límites de cada intervalo son:

- $a = mín(pH) = 2.74$
- $b = a + Rango = 3.163333$
- $c = b + Rango = 3.586667$
- $d = c + Rango = 4.01$

Nivel	Criterio	Intervalo	Conteo
Bajo	pH<b	$[2.74; 3.163333)$	267
Medio	pH $\geq$ b & pH<c	$[3.163333; 3.586667)$	1269
Alto	pH $\geq$ c	$[3.586667; 4.01)$	63

A partir de la siguiente figura es posible observar como el nivel de *pH* con mayor frecuencia es aquel que se denomina como "medio" con 1269 observaciones, mientras que los niveles "bajo" y "alto", presentan frecuencias muy bajas, esto es, 267 y 63 respectivamente.

```
G3
```



Modelo con variable indicadora pHi

En esta sección, se procede a generar un modelo logístico con variable de respuesta ordinal, ya que la variable de respuesta "calidad" tiene una jerarquía, esto es, una puntuación entre 0 y 10, donde 0 representa una mala calidad y 10 una calidad de vino excelente.

```
fit = vglm(quality ~ fixed.acidity + pHi, data = Datos, family = cumulative(parallel = TRUE))
summary(fit)
```

```
##
## Call:
## vglm(formula = quality ~ fixed.acidity + pHi, family = cumulative(parallel = TRUE),
##     data = Datos)
##
## Pearson residuals:
##                          Min       1Q   Median       3Q     Max
## logitlink(P[Y<=1])  -0.9198 -0.06304 -0.05795 -0.05153 17.7241
## logitlink(P[Y<=2])  -0.3286 -0.26343 -0.15649 -0.13068  7.5465
## logitlink(P[Y<=3])  -1.4357 -1.04615 -0.35390  1.04408  1.8178
## logitlink(P[Y<=4])  -3.4466  0.21827  0.24634  0.53341  0.7579
## logitlink(P[Y<=5]) -12.3625  0.07290  0.08043  0.09285  0.2943
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -3.68405    0.43567  -8.456  < 2e-16 ***
```

```
## (Intercept):2 -1.80815    0.32560  -5.553 2.80e-08 ***
## (Intercept):3  1.27227    0.30934   4.113 3.91e-05 ***
## (Intercept):4  3.29577    0.31992  10.302  < 2e-16 ***
## (Intercept):5  5.93415    0.39326  15.090  < 2e-16 ***
## fixed.acidity -0.18017    0.03220  -5.595 2.21e-08 ***
## pHiBajo        0.43078    0.29710   1.450    0.147
## pHiMedio       0.01117    0.25256   0.044    0.965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  5
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3]), logitlink(P[Y<=4]), logitlink(P[Y<=5])
##
## Residual deviance: 3755.387 on 7987 degrees of freedom
##
## Log-likelihood: -1877.694 on 7987 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):5'
##
##
## Exponentiated coefficients:
## fixed.acidity       pHiBajo      pHiMedio
##     0.8351289     1.5384581     1.0112315
```

Coeficientes	Estimación	Error Estándar	Valor - t	Pr(>|t|)	Significancia
Intercepto 1	-3.68405	0.43567	-8.456	< 2e-16	***
Intercepto 2	-1.80815	0.32560	-5.553	2.80e-08	***
Intercepto 3	1.27227	0.30934	4.113	3.91e-05	***
Intercepto 4	3.29577	0.31992	10.302	< 2e-16	***
Intercepto 5	5.93415	0.39326	15.090	< 2e-16	***
Acidez fija	-0.18017	0.03220	-5.595	2.21e-08	***
pHiBajo	0.43078	0.29710	1.450	0.147	
pHiMedio	0.01117	0.25256	0.044	0.965	

Conclusiones

- Se obtienen $G - 1$ interceptos, esto es $6 - 1 = 5$. Dado que, la variable de respuesta "Calidad" presenta 6 categorías.
- De acuerdo a los resultados obtenidos y teniendo en cuenta que la interpretación de los valores p es similar a la del modelo lineal. Es posible evidenciar que la variable denominada "Acidez fija" es significativa. Además, se encuentra negativamente relacionada con la variable de respuesta "Calidad", así la puntuación de la calidad del vino disminuiría 0.18017 por cada unidad que aumenta la acidez fija.