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:

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

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.

Los pasos para desarrollar el algoritmo de Newton son los siguientes:

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
```

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).

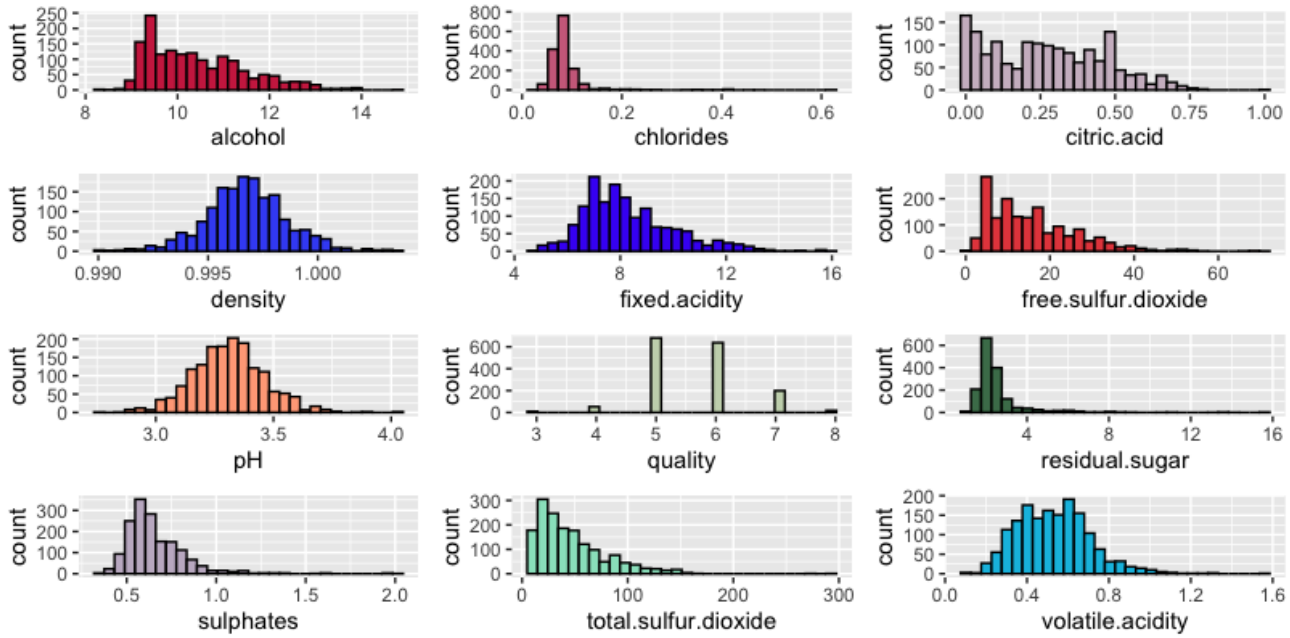
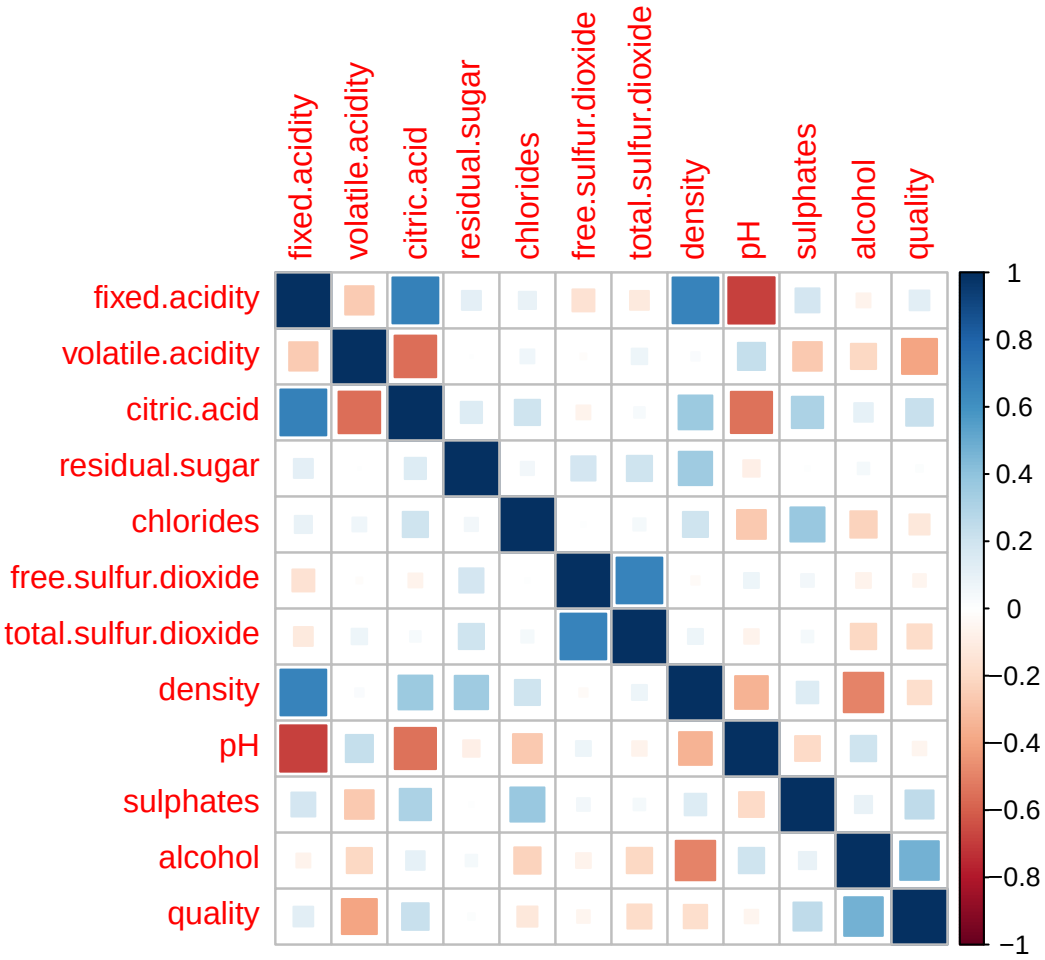


Figure 1: Distribución de las variables

- 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:

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

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

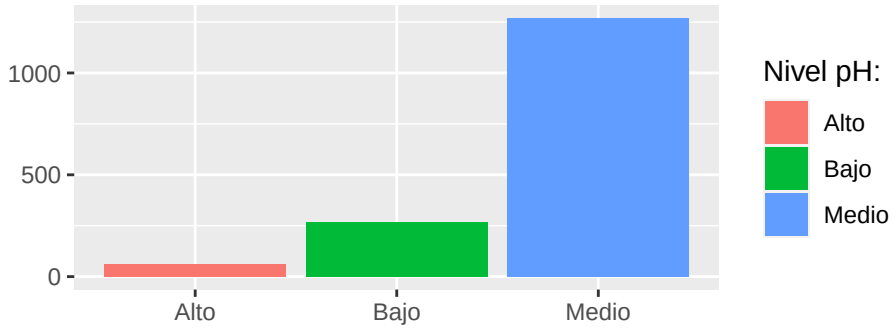
De esta manera, los límites de cada intervalo son:

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

Nivel	Criterio	Intervalo	Conteo
Bajo	pH<b	[2.74; 3.163333)	267
Medio	pH ≥ b & pH<c	[3.163333; 3.586667)	1269
Alto	pH ≥ 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 lineal generalizado (GLM)

Los modelo lineal generalizado (GLM) es una generalización flexible de la regresión lineal ordinaria que permite variables de respuesta que tienen modelos de distribución de errores distintos de una distribución normal, fueron desarrollados por Nelder - Wedderburn (1972), permiten ampliar la gama de distribuciones de la variable respuesta a todas aquellas que pertenezcan a la familia exponencial de densidades.

Al igual que la regresión lineal múltiple, permite cumplir con dos objetivos, determinar si existe relación lineal entre las xj y y y cuál es su magnitud.

En esta sección, se ajustan dos modelos lineales generalizados, en primer lugar un modelo lineal generalizado usando como respuesta la variable "calidad" (quality) y como variables de predicción la variable "acidez fija" (fixed acidity) y en segundo lugar teniendo como variables predictoras las variables "acidez fija" (fixed acidity) y "pHi", tal como sigue:

Modelo sin variable indicadora pHi

```
Modelo1 <- glm(Datos$quality ~ Datos$fixed.acidity, data=Datos)
```

$$Calidad = 5.15732 + 0.05754 * Acidezfija$$

Coeficientes	Estimación	Error estándar	Valor t	$Pr(>\|t\|)$	Significancia
Intercepto	5.15732	0.09789	52.684	< 2e-16	***
Acidez fija	0.05754	0.01152	4.996	6.5e-07	***

Modelo con variable indicadora pHi

```
Modelo2 <- glm(Datos$quality ~ Datos$fixed.acidity + pHi, data=Datos)
```

$$Calidad = 5.03884 + 0.07194 * Acidezfija - 0.11957 * pHiBajo + 0.02348 * pHiMedio$$

Coeficientes	Estimación	Error estándar	Valor t	$Pr(>\|t\|)$	Significancia
Intercepto	5.03884	0.13061	38.579	< 2e-16	***
Acidez fija	0.07194	0.01365	5.272	1.54e-07	***
pHiBajo	-0.11957	0.12566	-0.952	0.341	
pHiMedio	0.02348	0.10672	0.220	0.826	

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 en ambos modelos. Además, se encuentra positivamente relacionada con la variable de respuesta "Calidad", así la puntuación de la calidad del vino aumentaría 0.05754 y 0.07194 respectivamente por cada unidad que aumenta la acidez fija.

Finalmente, respecto al criterio de información empleado, esto es, el criterio de información de Akaike (AIC), se tiene que mientras más pequeño sea el AIC, mejor se ajustará el modelo a los datos. De esta manera, el modelo que mejor se ajusta a los datos es el modelo con la variable indicadora pHi.

Sin pHi	Con pHi
3834.5	3833

Conlusiones

Biblíografia