Cheddar Faraway

1. Introducción

Tenenemos un conjunto de datos en el que se recogen observaciones de una cata de quesos, nuestras variables son:

- **Taste:** una valoración subjetiva de los jueces.
- **Acetic:** la concentración de ácido acético en un queso de terminado en esca la logarítmica
- **H2S:** la concentración de sulfito de hidrógeno en escala logarítmica.
- **Lactic:** Concentración de ácido láctico

```
library(faraway)
data(cheddar)
head(cheddar)
```

```
##   taste Acetic   H2S Lactic
## 1  12.3  4.543 3.135   0.86
## 2  20.9  5.159 5.043   1.53
## 3  39.0  5.366 5.438   1.57
## 4  47.9  5.759 7.496   1.81
## 5   5.6  4.663 3.807   0.99
## 6  25.9  5.697 7.601   1.09
```

```
sapply(cheddar, class)
```

```
##     taste    Acetic       H2S    Lactic
## "numeric" "numeric" "numeric" "numeric"
```

Al ser todas numéricas no hay que hacer encoding a variables binarias

Usamos el número de observaciones para determinar los conjuntos de train y test

```
numObs <- dim(cheddar)[1]
numObs
```

```
## [1] 30
```

Tenemos 30 observaciones, ¿Tenemos alguna en la que falten datos? ¿Tenemos outliers?
Esto lo trataremos en la sección 4

```
# Para asegurar que sea reproducible

set.seed(101)
sample <- sample.int(n = nrow(cheddar),
                     size = floor(.7 * nrow(cheddar)),
                     replace = F)
train <- cheddar[sample, ]
test <- cheddar[-sample, ]
```

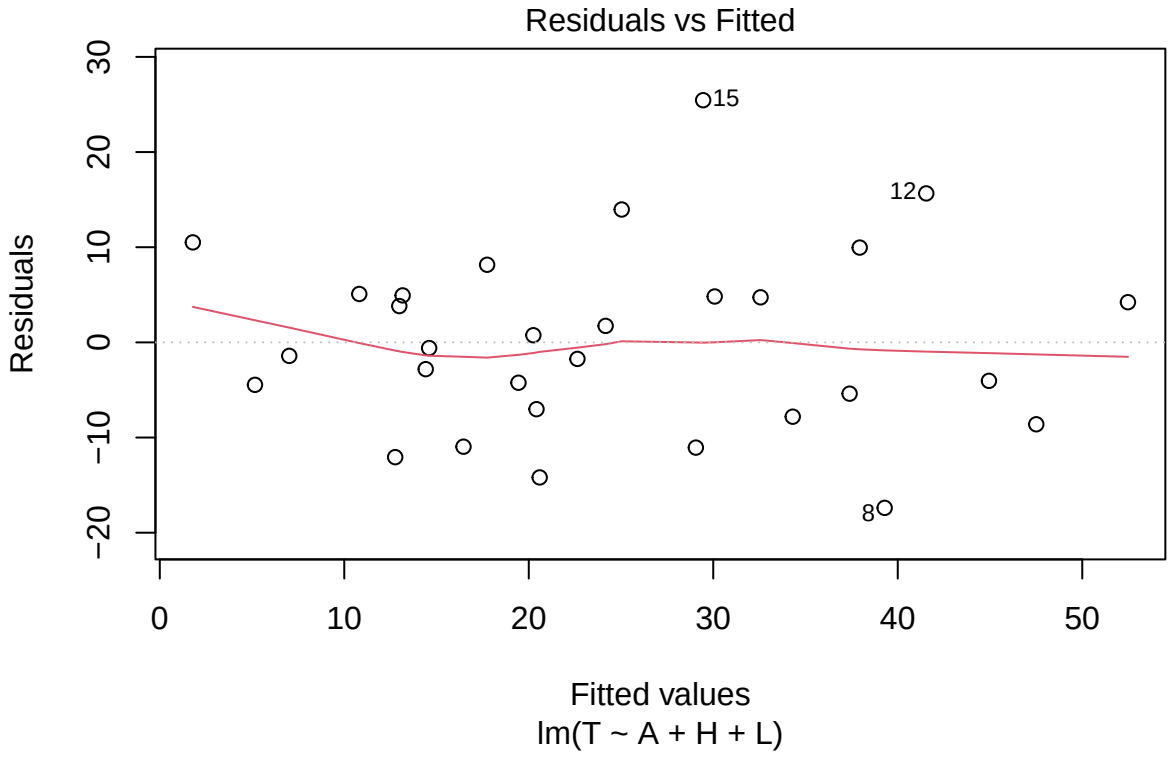
2.Estudio y evaluación del modelo completo.

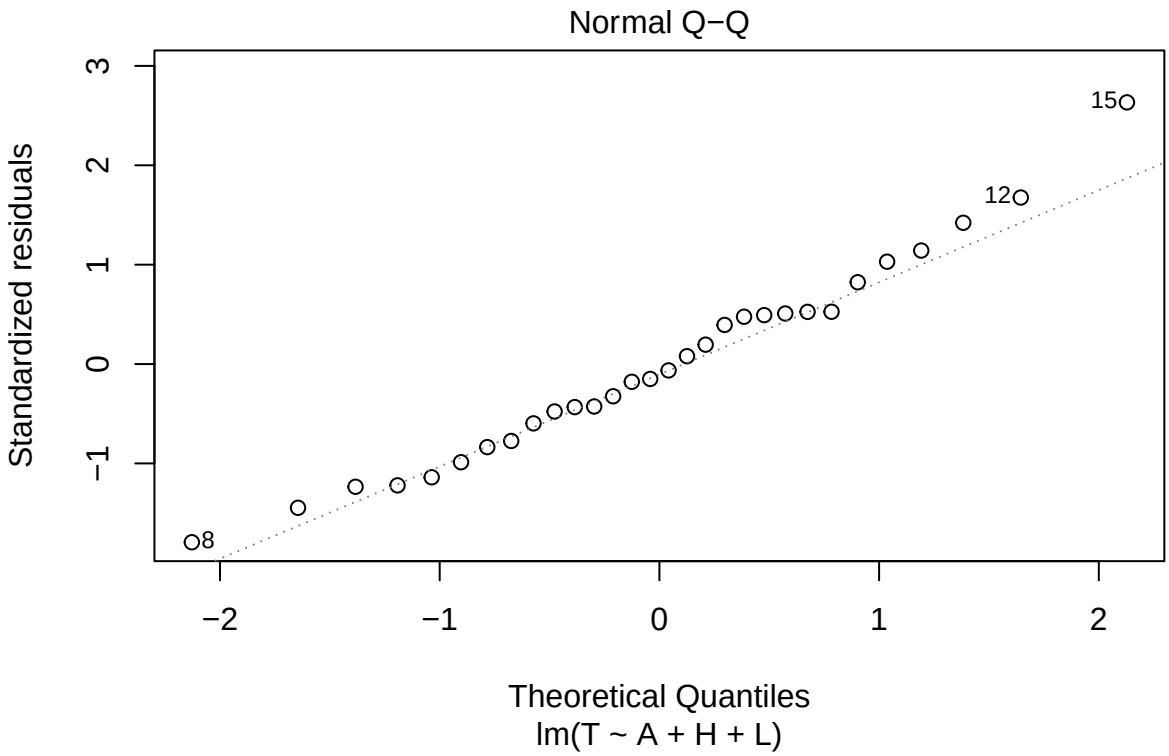
Simplificamos nuestra notación para las variables, la que intentaremos predecir es taste.

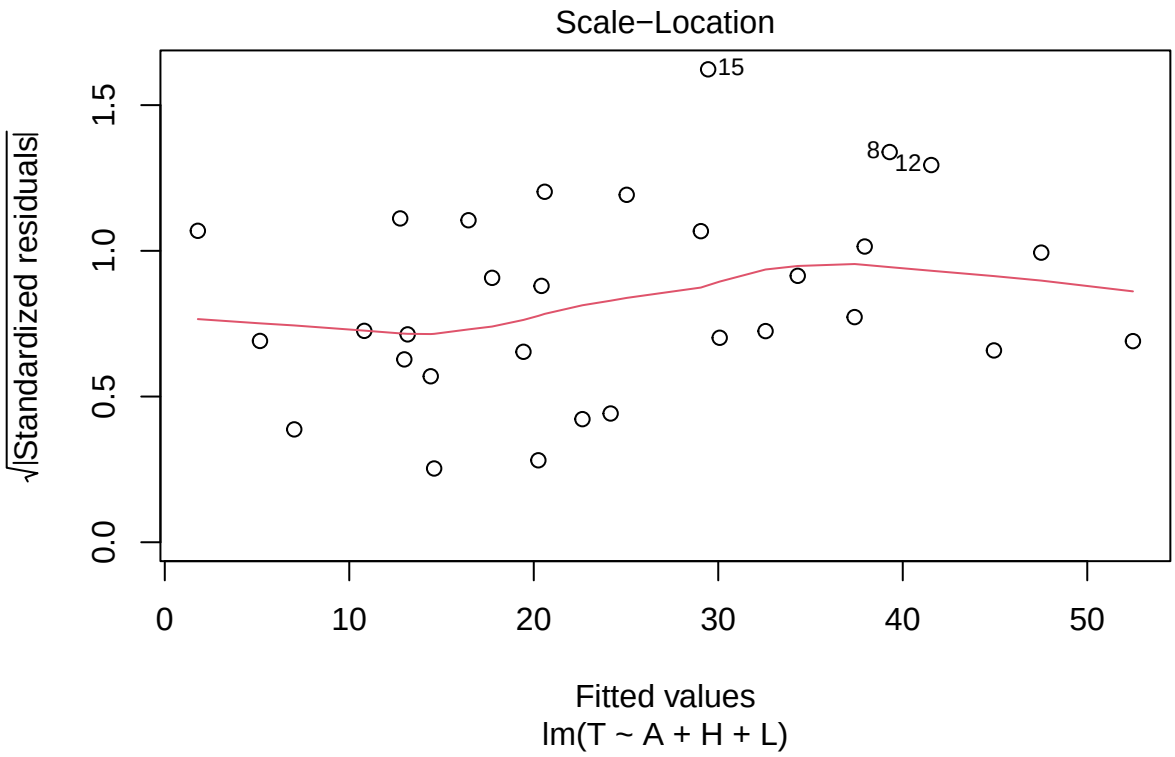
```
T <- cheddar$taste
A <- cheddar$Acetic
H <- cheddar$H2S
L <- cheddar$Lactic
```

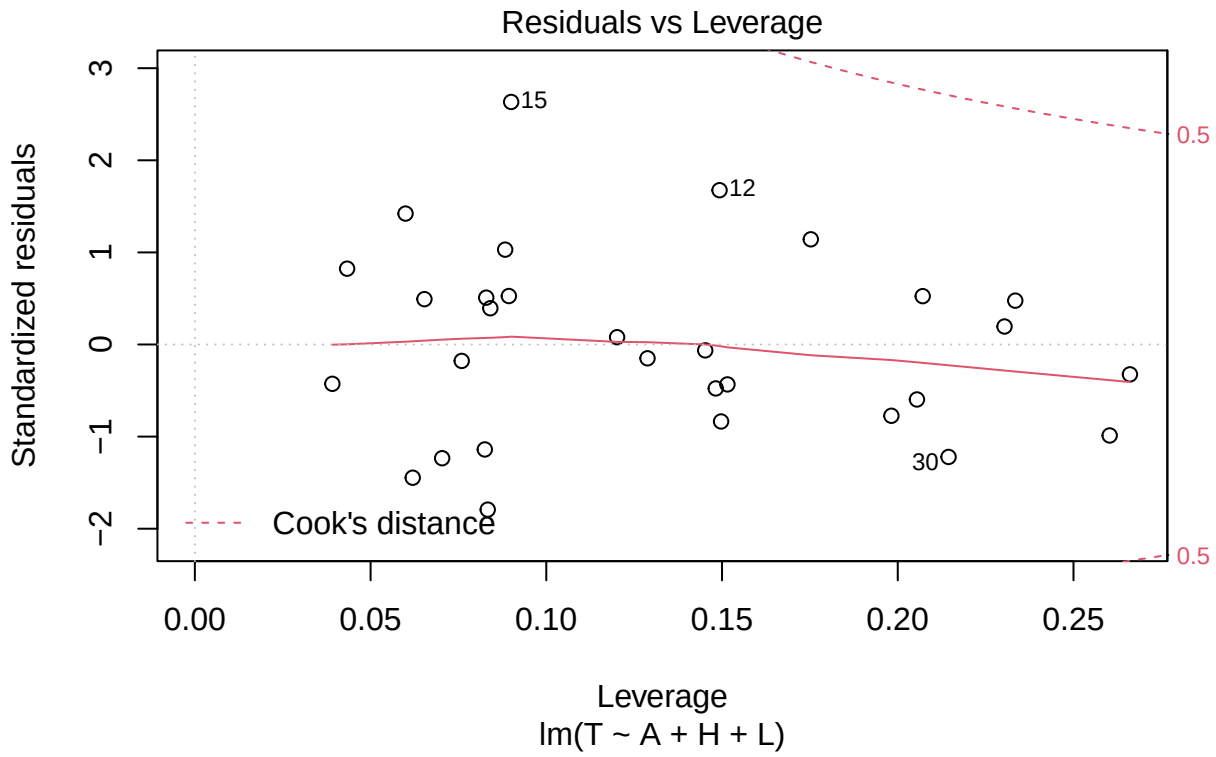
Definimos nuestro modelo completo:

```
model.completo.lm <- lm(T~A+H+L, data=train)
plot(model.completo.lm)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
15
12
8
−2
−1
0
1
2
Theoretical Quantiles
lm(T ~ A + H + L)


Scale−Location
15
8
12
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
10
20
30
40
50
Fitted values
lm(T ~ A + H + L)




```
cor(train)
```

```
##            taste    Acetic       H2S    Lactic
## taste  1.0000000 0.4944292 0.7769666 0.7598731
## Acetic 0.4944292 1.0000000 0.5844270 0.7243627
## H2S    0.7769666 0.5844270 1.0000000 0.6776785
## Lactic 0.7598731 0.7243627 0.6776785 1.0000000
```

```
shapiro.test(resid(model.completo.lm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model.completo.lm)
## W = 0.98021, p-value = 0.8312
```

Observamos que estamos en la hipótesis de que el error nuestro modelo se distribuye de manera normal.

¿Tienen todas las variables un impacto relevante en el modelo?

```
summary(model.completo.lm)
```

```
##
## Call:
```

```
## lm(formula = T ~ A + H + L, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## A             0.3277     4.4598   0.073  0.94198
## H             3.9118     1.2484   3.133  0.00425 **
## L            19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

```
betahat<- matrix(coef(model.completo.lm),ncol=1)
```

Obervamos el p-valor de A nos indica que con casi toda seguridad A no tiene impacto real en el modelo ($>$ 0.94)

```
anova(model.completo.lm)
```

```
## Analysis of Variance Table
##
## Response: T
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## A          1 2314.14 2314.14 22.5481 6.528e-05 ***
## H          1 2147.02 2147.02 20.9197 0.0001035 ***
## L          1  533.32  533.32  5.1964 0.0310795 *
## Residuals 26 2668.41  102.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los p-valores son lo suficientemente bajos como para rechazar varianza constante