Regresión: Modelos Estadísticos

Conjunto de Datos: Cheddar Faraway

true true true true

25/03/2022

Abstract

Hemos analizado con las herramientas proporcionadas en el curso de Modelos Estadísticos el conjunto de datos, *Cheddar*, distribuido en la librería Faraway de R. Para ello hemos utilizado diversas técnicas de regresión lineal y no lineal.

1 Introducción

En un estudio de queso Cheddar realizado en el Valle de Latrobe (Victoria, Australia), se estudiaron muestras de queso en las que se analizó su composición química y fueron dadas a probar a distintos sujetos para que valoraran su sabor. Los valores asignados a cada queso son el resultado de combinar las distintas valoraciones.

El DataFrame **cheddar** de la librería **faraway** consiste de 30 muestras de queso Cheddar en las que se ha medido el sabor (*taste*) y las concentraciones de ácido acético (*Acetic*), ácido sulfhídrico (*H2S*) y lactosa (*Lactic*).

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

A lo largo del documento hacemos uso de las siguientes librerias de R:

Vamos a utilizar el dataset *Cheddar*. Cargamos los datos y enseñamos las primeras observaciones.

Si en nuestro dataset tuviesemos entradas vacías (NA), tenemos varias posibilidades para lidiar con este problema:

- No utilizar/Eliminar las observación que contienen valores.
- No utilizar/Eliminar las variables que contienen las entradas vacías.
- Intentar completar los valores. Existen métodos menos y más sofisitcados:
  - Remplezar con la **media, media o moda**.
  - Crear una **nueva categoría** para valores vacíos.
  - Utilizar algún modelo de **regresión**.
  - Usar un modelo de **K-Nearest Neighbors (KNN)**.

A continuación, comprobamos que no hay entradas vacías,

```
## [1] FALSE
```

Toas las variables son numéricas (**cuantitativas**). No hay que transformar las variables no cuantitativas (**cualitativas**), conviertiendolas en variables binarias. Para ello, podríamos deberíamos hacer encoding a variables binarias, el lenguage de programación R nos permite utilizar *as.numeric*.

Con estas variables vamos a intentar **explicar** cómo los valores observados de una variable Y (taste) dependen de los valores de otras variables (Acetic, H2S, Latcic), a través de una relación funcional del tipo Y = f(X) También vamos a intentar **predecir** el valor de la variable Y para valores no observados de las variables X.

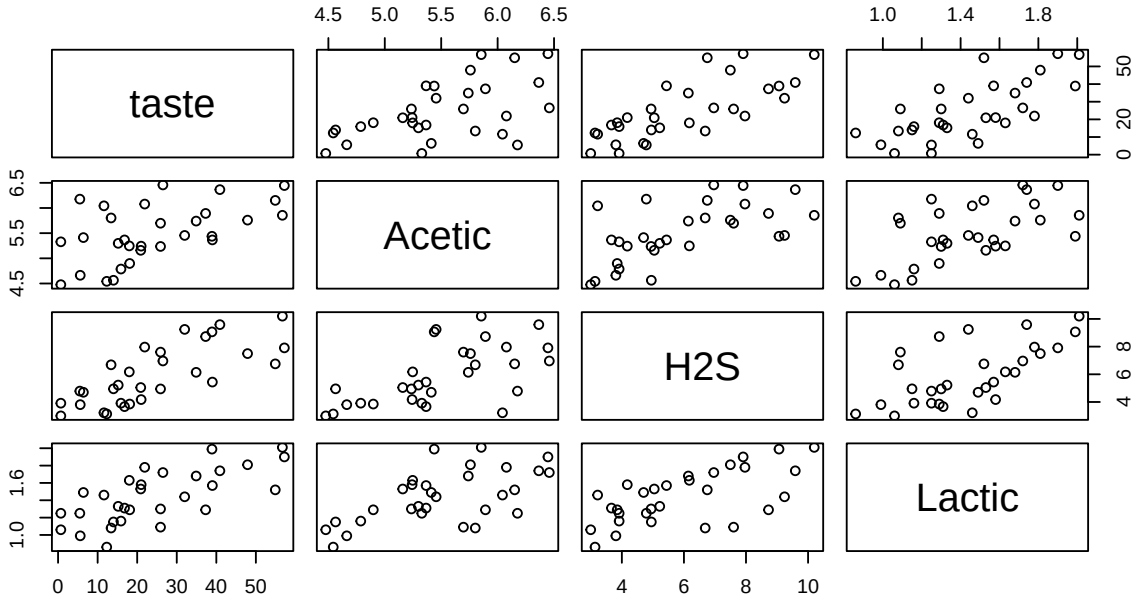
Usamos el número de observaciones para determinar los conjuntos de train y test.

Tenemos 30 observaciones en nuestro dataset. Ahora procedemos a divirlo en el *conjunto de train y test*. El primero lo utilizaremos para entrenar nuestros modelos y el segundo lo usamos para cuantificar el error de los modelos.

Ahora nos hacemos las siguientes preguntas: ¿Podemos suponer que la distribución de las variables es normal?,¿Tenemos alguna en la que falten datos?, ¿Tenemos *outliers*?, etc. En las siguientes secciones trataremos de responder a estas preguntas y muchas otras acercar de nuestro conjunto de datos.

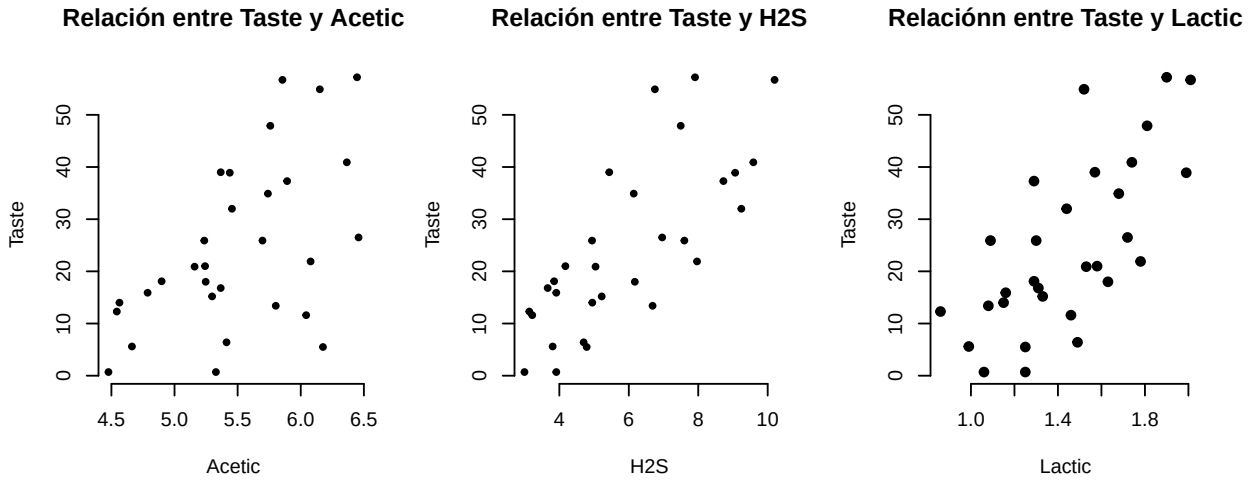
Para asegurar que sea reproducible utilizamos una semilla, que permite fijar los valores pseudoaleatorios obtenidos en muchas de las funciones utilizadas.

Hacemos un pequeño estudio preliminar de nuestras variables. Mostramos un scatter plot de las cada variable contrastada con el resto. Esto permite ver *a ojo* si algún par de variables tiene correlación.



Ahora, utilizamos la funcón *summary* de R, la cual nos permite estimar algunos de las características de la distribución del dataset. La siguiente tabla nos muestra los estadísticos más comunes: el mínimo, máximo, mediana, media y el 1er y 3er cuartil.

Ploteamos las gráficas de dispersion entre la variable respuesta *taste* y las variables predictoras *Acetic, H2S, Lactic*.



Podemos observar que la que aperentemente guarda una menor relación lineal con taste es la variable Acetic, esto será comprobado con distintos tests

2 Estudio y evaluación del modelo completo.

Simplificamos nuestra notación para las variables. Intentaremos predecir la variable *taste* usando el resto de variables. Para empezar, definimos el modelo completo, el cual se usan todas las variables para nuestro modelo lineal múltiple.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i(p-1)} + \epsilon_i, \quad i = 1, \ldots, n$$

donde Y_i es el valor de la variable respuesta para el individuo i-ésimo,
β_0 y los β_j son los par´ametros $j = 1, ..., p-1$,
x_{ij} son los elementos de la matriz de las variables explicativas
ϵ_i es el término del error aleatorio que suponemos que se distribuye como una $\mathcal{N}(0, \sigma^2)$, donde σ^2 es la varianza que suele ser desconocida.

2.1 Resolución mediante matrices

Utilizamos el método de mínimos cuadrados que estima los valores $\hat{\beta}$ intentando minimizar los errores ϵ. Como hemos visto en clase, la fórmula que se deduce es:

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

donde X es una columna de 1's concatenada con las variables que usamos para predecir. Es importante clarificar que en este proceso solo usamos el training set.

Por tanto, aproximamos las β modelo lineal completo con los valores de $\hat{\beta}_0, \ldots, \hat{\beta}_3$ con los siguientes valores:

$$\hat{\beta}_0 = -28.8767696, \quad \hat{\beta}_1 = 0.3277413, \quad \hat{\beta}_2 = 3.911841, \quad \hat{\beta}_3 = 19.6705434$$

2.2 Resolución usando librerias de R

Podemos utilizar la función *lm*, ya programada en R. Definimos el modelo completo:

$$taste \sim Acetic + H2S + Lactic,\ data = cheddar$$

```
## (Intercept)      Acetic         H2S      Lactic
## -28.8767696   0.3277413   3.9118411  19.6705434
```

Evidentemente los resultados son los mismos.

Estudiemos preliminarmente si es un modelo lineal adecuado, para ello comprobaremos las hipótesis estándar del modelo lineal de regresión usando el **test de normalidad Shapiro-Wilk**. La función *shapiro.test* le pasamos por parámetro el residuo/error de cada una de las muestras y nos devuelve un p-valor.

Observamos que estamos en la hipótesis de que el error nuestro modelo se distribuye de manera normal, ya que el p-valor es $0.8865 > 0.05$.

Ahora nos preguntamos si hay variables que tienen un mayor impacto en el modelo. Estas variables podrían ser *outliers* y podríamos deshecharlas del training set ya que podrían estar perjudicando la predicción del modelo negativamente.

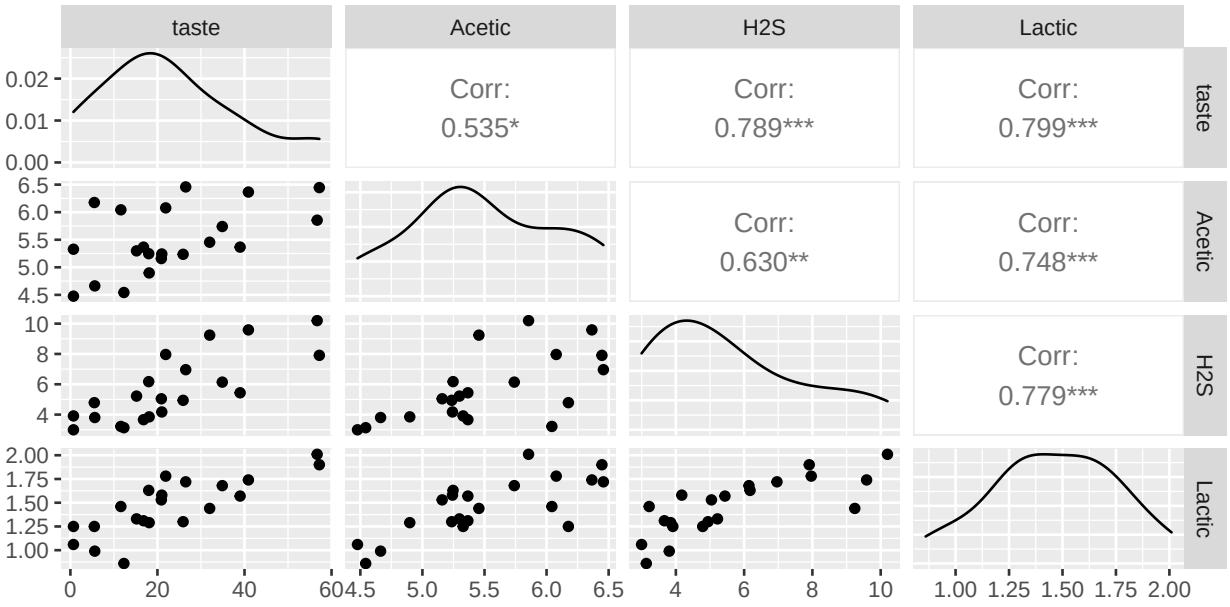
Obervamos el p-valor de Acetic en el resumen del modelo nos indica que con casi toda seguridad Acetic no tiene impacto real en el modelo (> 0.94)

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

Una vez hemos concluido que aunque estamos en las hipótesis de regresión lineal el modelo completo a pesar de ser el más complejo probablemente da resultados similares a otro más simple.

2.2.0.1 Correlaciones y tabla de resultados con el estudio de sus p-valores

Usamos el paquete *GGplot* de R, el cual nos permite visualizar la correlación y dispersión entre las distintas variables.



Ahora utilizamos el anális **anova** (Analysis of Variance), para justificar si podemos eliminar alguna variable del modelo.

	Acetic	H2S	Lactic
p-valor	6.53e-05	0.0001035	0.0310795

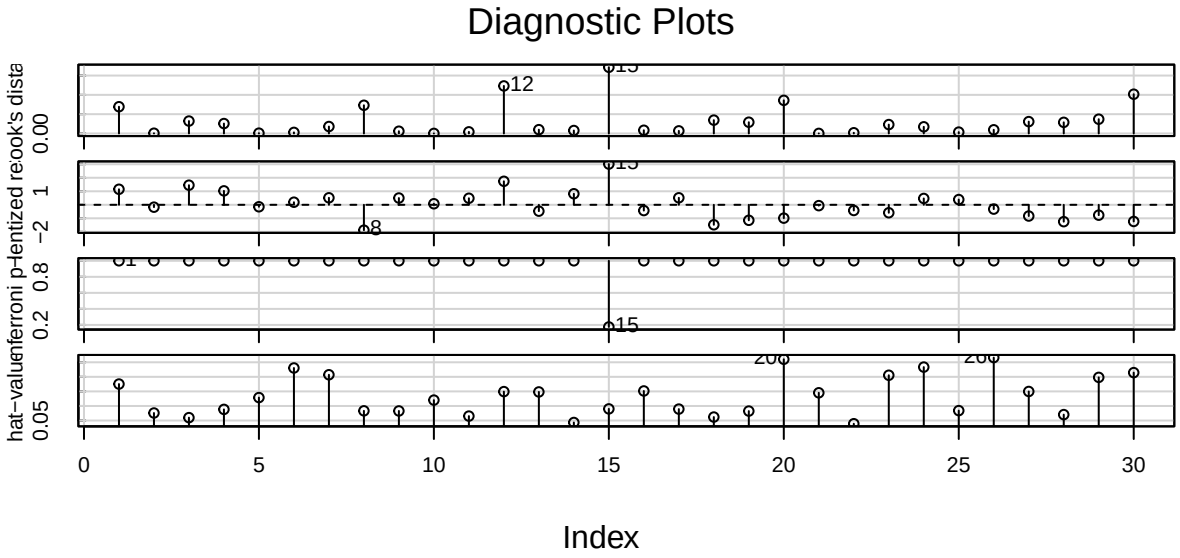
Todos nuestros p-valores son adecuados a un nivel $\alpha = 0.05$. En nuestro caso nos vale, sin embargo, la variable *Lactic* no lo cumpliría si disminuimos el nivel de α a una cota inferior como 0.01.

2.3 ¿Tiene *outliers* nuestra muestra?

Para comprobarlo basta realizar el **test de Bonferroni** sobre nuestro modelo completo:

Concluimos con un nivel $\alpha = 0.05$ que no tenemos ningún outlier en nuestra. Lo más cercano a un *outlier* que tenemos es la observación número 15, que tiene un valor **Bonferroni p** de 0.17453 (no se acerca a 0.05). Por tanto, no tenemos razones por las que eliminar alguna observación inusual de nuestro conjunto de datos.

Esto se puede comprobar graficamente a través del siguiente gráfico, el cual mide la influencia de cada observación sobre cada una de las betas de nuestro modelo.



Vemos que la que más influye es la antes mencionada obsercación 15 y por tanto es posible que en el resto de modelos que estudiemos con más detalles salga de la muestra como observación influyente, si está en el conjunto *train*

2.4 ¿Cuál es el mejor modelo?

Como dice el **Principio de la Navaja de Ockham**, a menudo la explicación más simple es la correcta. Queremos seleccionar predictores que explican los datos de la manera más simple posible, sin disminuir la calidad de las predicciones mucho.

2.4.1 Separacion del dataset en conjuntos de entrenamiento y test (70-30%)

Hemos escogido distintas semillas para estar en condiciones de realizar un estudio más amplio, en la elección de las mismas se ha intentado evitar aquellas que generaban muestras demasiado similares. Las semillas usadas son 1, 1100 y 5 posteriormente se introducirán dos más.

Consideremos los conjuntos de entrenamiento resultantes de las semillas: *train.1*(semilla 1), *train.2*(semilla 1100), *train.3*(semilla 5)

2.4.2 Método Backward

Partimos del modelo completo estudiado en la sección anterior y aplicamos con $\alpha = 0.05$, el metodo de Backward, que consiste en eliminar la variable que *menos influya* a la predicción. Primero realizamos una iteración explícita del método, posteriormente se construyen a través de la libreria *mixlm* de R.

	Acetic	H2S	Lactic
p-valor	6.53e-05	0.0001035	0.0310795

Eliminamos Acetic del modelo debido que su p-valor es > 0.05.

	H2S	Lactic
p-valor	0.0017429	0.0188499

Repetimos el proceso con la variable H2S, ya que tiene un p-valor mayor que $\alpha = 0.05$

	Lactic
p-valor	1.4e-05

El p-valor es menor que $\alpha = 0.05$, por lo que hemos concluido, ya que no tenemos suficiente certeza para poder eliminar otra variable.Por tanto, tenemos como resultado que la variable que mejor explica el *taste* es *Lactic*. Modelo resultante: **taste ~ Lactic, data = cheddar[train.1,]**.

Por otro lado los modelos backward resultantes por *mixlm* son:

taste ~ H2S + Lactic, data = cheddar[train.2,] y por otro lado **taste ~ H2S + Lactic, data = cheddar[train.3,]**.

2.4.3 Método Forward

El método Forward consiste en empezar con un modelo de una variable y vamos añadiendo las que más influyan, desarrollaremos el primer modelo de forma explícita y el resto los generaremos con *mixlm*. De esta manera tenemos:

	Acetic	H2S	Lactic
p-valor	0.0125179	2.12e-05	1.39e-05

Actualizamos añadiendo Lactic por tener el menor p-valor.

	Acetic	H2S
p-valor	0.5039877	0.0512171

Con nivel de significación $\alpha = 0.05$ este sería nuestro modelo final. Modelo resultante = **taste ~ Lactic, data = cheddar[train.1,]**
Por otro lado los modelos forward resultantes por *mixlm* son:

taste ~ H2S + Lactic, data = cheddar[train.2,] y por otro lado **taste ~ H2S + Lactic, data = cheddar[train.3,]**.

2.5 Construcción por criterios

En esta subsección trataremos de encontrar un candidato a mejor modelo, construyendo nuestros modelos usando distintos enfoques. Tras aplicar los siguientes criterios a la hora del desarrollo de modelos: R^2 ajustado, Cp de Mallows, Criterio de Informacion de Bayes (BIC), Criterio de Informacion de Akaike (AIC) (los desarrolos se pueden encontrar en el script), llegamos a las siguientes conclusiones:
solo aparece un modelo nuevo usando el criterio del estadístico R^2, **taste ~ H2S + Lactic, data = cheddar[train.1,]**.
Notamos que la combinación de H2S + L aparece en todos nuestros conjuntos de entrenamiento en algón momento, es candidata a ser nuestra mejor elección.

Comparamos los modelos obtenidos hasta ahora en su respectiva muestra de entrenamiento con el modelo completo en ese conjunto de entrenamiento.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
19	1853.543	NA	NA	NA	NA
17	1410.701	2	442.8415	2.668285	0.0982145

	train1: L vs Completo	train1: H2S + L vs Completo	train2: H2S + L vs Completo	train1: H2S+ L vs Completo
p-valor	0.0982145	0.3362689	0.9551123	0.5551292

Con un estos p-valores podemos decir que con un nivel de significación α ningún modelo es notablemente diferente de su contraparte salvo en el caso de los modelos resultantes en *train2* esto puede ser por la cantidad de observaciones influyentes presentes en la muestra, lo trataremos en la siuigente sección.
Diagnostico: Comprobaciones de hipotesis, outliers y observaciones influyentes

En esta sección estudiaremos si nuestros modelos cumplen las condiciones necesarias de un modelo de regresión lineal.

Nuestro enfoque consistirá en un analisis gráfico, acompañado de tests estadísticos en los casos en los que se aprecie una discrepancia notable.

2.6 ¿Son nuestros *modelos*, modelos de regresión lineal?: Comprobación de hipótesis.

En la sección 3 se toma un enfoque *naïve* a la hora de construir los modelos, ya que no hemos estudiado si hay observaciones influyentes, podríamos tener una muestra que no es la adecuada para el estudio de nuestros datos.

Un modelo de regresión lineal debe satisfacer las siguientes hipótesis con nivel de siginificación α adecuado:

1. Los errores ϵ_i tienen distribución normal.

2. Los errores ϵ_i tienen media cero.

3. Los errores ϵ_i tienen varianza constante.

4. Los errores ϵ_i no están correlacionados.

	train1: L	train1: H2S + L	train2: H2S + L	train3: H2S + L	Nivel de siginificación	Test utilizado
Linealidad	0.14	0.609	0.812	0.816	0.05	Resettest
Normalidad	0.255	0.933	0.982	0.452	0.05	Test Shapiro Wilk
Media = 0	1	1	1	1	0.05	t-test
Varianza constante	0.63	0.438	0.392	0.483	0.05	Test ncv
Correlación	0.608	0.508	0.924	0.958	0.05	Test de Durbin-Watson

Podemos observar que se verifican a nivel de signifcación $\alpha = 0.05$ se verifican todas las hopótesis de modelo de refresión lineal.

```
#PENDIENTE: qqplot qqnorm en grid de 2x2
```