Tarea 4: Elección de λ: El riesgo de predicción

Angie Rodríguez Duque & César Saavedra Vanegas

Diciembre 04 de 2020

Actividad 1

En los métodos de regresión no paramétrica los estimadores en general no son insesgados, por lo que la varianza del estimador no será suficiente para evaluar la incertidumbre inherente a estos métodos.

De acuerdo a lo anterior, el presente documento tiene como objetivo responder a la pregunta: ¿Cuál valor de λ sería una "buena elección"?, para ello se hará uso del estimador rice y del estimador UBRE.

1. Base de datos

El conjunto de datos empleados en el presente documento proviene del repositorio de la base de datos de aprendizaje automático de UCI. Los datos originales consisten en variables del vino portugués "Vinho Verde" y cuenta con 1599 observaciones de vino rojo y 4898 observaciones de vino blanco. Para cada uno se evalúa la calidad del vino (Calificación entre 0 y 10) y 11 variables químicas (cuantitativas), que son las siguientes: Acidez fija, Acidez volátil, Ácido cítrico, Azúcar residual, Cloruros, Dióxido de azufre libre, Dióxido de azufre total, Densidad , PH, sulfatos y alcohol. Específicamente se hará uso de las observaciones procedentes del vino rojo.

2. Muestra aleatoria

Se procede a seleccionar una muestra aleatoria de 60 vinos de la base de datos y se escoge las variables "Acidez fija" como respuesta y "pH" como predictora, cuyas descripciones son las siguientes:

- **Acidez fija:** La mayoría de los ácidos involucrados con el vino o fijos o no volátiles (no se evaporan fácilmente)
- **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.

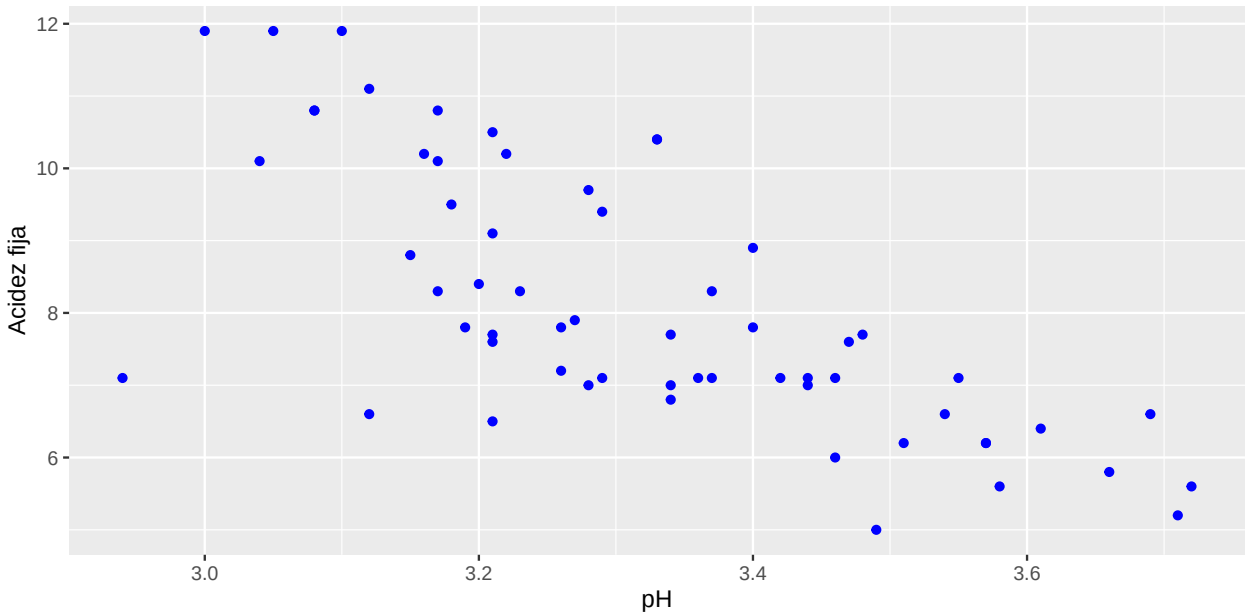
```
# Tamaño de la muestra
n <- 60
# Seleccion de la muestra
set.seed(12345)
muestra <- Datos %>% sample_n(size=n,replace=FALSE)
muestra <- muestra %>% arrange(pH)
```

Representación gráfica

A continuación se procede a graficar el comportamiento de ambas variables a partir del diagrama de dispersión:

```
x <- muestra  %>% dplyr::select(fixed.acidity, pH)
```

```
ggplot() + geom_point(data = x, aes(x = pH, y = fixed.acidity), col="blue") +
  ylab("Acidez fija") + xlab("pH")
```



Mediante la gráfica de dispersión se puede interpretar un tipo de relación lineal negativa entre ambas variables de estudio, esto es, mientras mayor es el pH menor es la acidez fija, y mientras menor sea el pH del vino mayor será su acidez. Es por esta razón que en la enología, es decir, en la ciencia, técnica y arte de la producción del vino, los vinos tintos no se caracterizan por tener una acidez tan fuerte en comparación con los vinos blancos, pues el gusto amargo de algunos de sus taninos, se acentúa demasiado.

De acuerdo a lo anterior, el pH influye significativamente en la sensación de astringencia de los vinos tintos. Se observa fácilmente que el incremento del pH reduce la sensación de astringencia de los vinos o de los zumos de frutas tánicas. Este fenómeno se explica, al menos parcialmente, por la interacción de la acidez con la precipitación o la desnaturalización de las proteínas encargadas de la lubricación de la cavidad bucal en presencia de polifenoles.

3. Estimación de la varianza $(\hat{\sigma}^2)$

En esta sección se estimará la varianza del modelo haciendo uso del estimador de Rice denotado como σ_R^2 y propuesto por John Rice en 1984. Su expresión es la siguiente:

$$\sigma_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2$$

4. Elección de λ

La elección del λ más apropiado para la estimación de μ en el ejemplo de vino rojo se lleva a cabo mediante el estimador insesgado del riesgo, también conocido como **UBRE** (UnBiased Risk Estimator) el cual hace uso de series de cosenos.

$$\hat{R}(\lambda) = \frac{1}{n} RSS(\lambda) + \frac{2}{n} \hat{\sigma}^2 tr\left[S_\lambda\right] - \hat{\sigma}^2$$

Donde: $\lambda \in (1, 2, ..., 60)$ es el número de funciones f_i

Deseamos entonces construir un dataframe tomando como variable respuesta "acidez fija" y como variable predictora "pH" donde f es la base de cosenos (CONS) que elegimos previamente.

Se imprimen los 10 primeros valores UBRE, los cuales generan el λ optimo que minimiza el riesgo y que resulta se el mas adecuado para la estimación de μ.

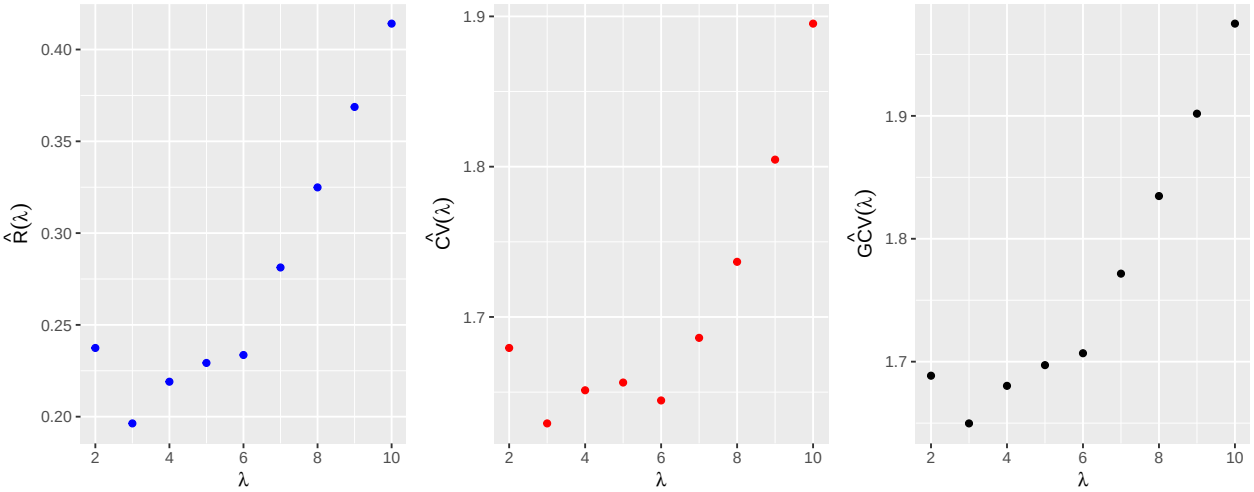
```
# Estimador UBRE, CV y GCV
all.R
```

```
##        UBRE       CV      GCV LAMBDA
## 1 0.2374374 1.679390 1.688573      2
## 2 0.1963344 1.629180 1.649753      3
## 3 0.2190494 1.651224 1.680319      4
## 4 0.2292779 1.656378 1.697177      5
## 5 0.2336055 1.644459 1.706858      6
## 6 0.2812813 1.686110 1.771622      7
## 7 0.3249054 1.736705 1.834760      8
## 8 0.3687135 1.804683 1.901791      9
## 9 0.4141164 1.895187 1.975067     10
```

Selección de λ

Ahora, tenemos la estimación del comportamiento de la acidez fija de acuerdo al pH de los vinos usando series de Fourier con base de cosenos y con un $\lambda = 3$, el cual fue seleccionado por medio de los métodos UBRE, CV, GCV. A partir de lo anterior, se puede decir que $\hat{\mu}_3$ es una buena aproximación a μ.

```
grid.arrange(plot1, plot2, plot3, nrow=1)
```



Finalmente se observa mediante los graficos que el valor de λ que minimiza las estimaciones segun los criterios son:

Estimador	Mejor Lambda
UBRE	3
CV	3
GCV	3

Para los cuales los resultados son los siguientes:

UBRE	CV	GCV	Lambda
0.1963344	1.629180	1.649753	3

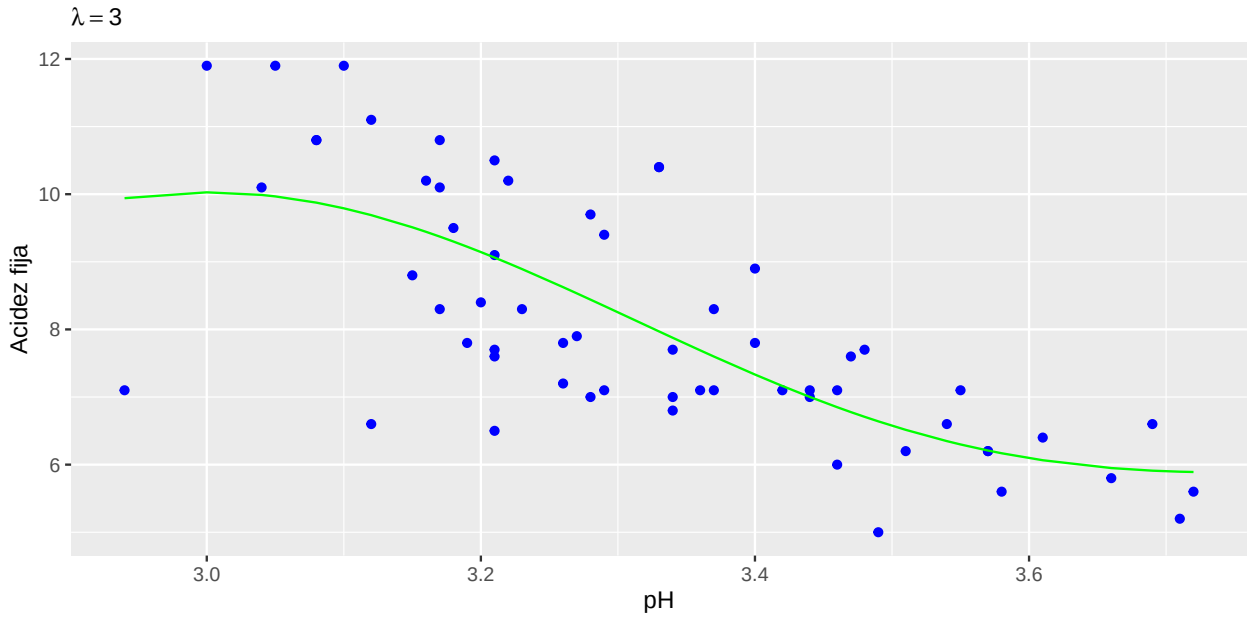
5. Estimación del modelo de regresión no paramétrica

Tras haber elegido el valor optimo de λ se prosigue a estimar el modelo de regresión no paramétrica. Los resultados obtenidos se presentan a continuación en la tabla que reune el valor del $\hat{R}(\lambda)$ para cada λ de acuerdo al método UBRE.

Representación de $\mu_3(X)$

Se observa el ajuste con $\lambda = 3$ con los datos reales (puntos) y los datos ajustados por el modelo (línea) de la variable "Acidez fija" vs "pH".

```
ggplot()+ geom_point(data = x, aes(x = pH, y = fixed.acidity),col="blue") +
  geom_line(data = x, aes(x =pH, y = fitted), col="green") +
  labs(subtitle = expression(lambda==3)) +
  ylab("Acidez fija") + xlab("pH")
```



Tenemos entonces la estimación del comportamiento de la acidez fija para el pH usando series de Fourier con base de cosenos y con un $\lambda = 3$, que seleccionamos por medio del método UBRE, CV, GCV, podríamos decir que μ_3 es una buena aproximación a μ.

6. Interpretaciones

A partir del modelo anterior se puede decir que:

- Se evidencia que tanto la varianza como el sesgo tienden a 0 cuando n crece, esto es, cuando $n = 60$ se obtiene una varianza de .
- De acuerdo con los resultados de la tabla y de la figura, el valor óptimo de λ, basado en el estimador UBRE, es $\lambda = 3$.
- En otras palabras, basados en este indicador, elegiremos a μ_3 como el mejor estimador de μ en el problema de vino tinto usando el estimador de cosenos.

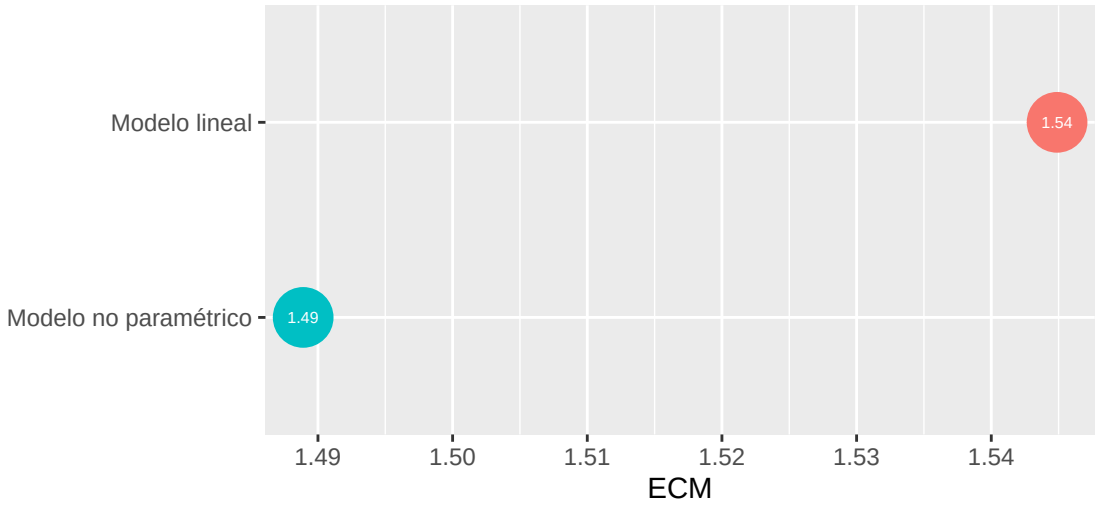
7. Ajuste de modelo lineal y comparación

A continuación se realiza el ajuste del modelo lineal general

8. Comparación de modelos

Finalmente buscamos realizar la comparación entre modelos para escoger el que seria el de mejor ajuste, esto mediante el error cuadrático medio (ECM) y esto da como resultado la figura que se presenta a continuación.

```
ECM
```



Después de graficar el modelo lineal, se procede a realizar la comparación entre el modelo de regresión no paramétrica y el modelo lineal:

- Respecto a la fuerza de la relación, se evalúa qué tan cerca se ajustan los datos a cada uno de los modelos para estimar la fuerza de la relación entre la variable predictora pH (X) y la variable de respuesta Acidez fija (Y). Se evidencia una relación fuerte en ambos casos, sin embargo la relación es notoriamente más significativa en el modelo de regresión no paramétrica, donde los puntos correspondientes a los vinos se adhieren mucho más a la línea de regresión ajustada por lo cual el método no paramétrico modelará con mayor precisión nuestros datos.

Esto se verifica mediante el calculo del error cuadrático medio, como se observa en la figura anterior, donde el modelo no paramétrico obtiene un $MSE = 1.49$ y en contraste al $MSE = 1.54$ obtenido para el modelo lineal, lo cual permite concluir que este seria el modelo que se ajusta mejor a nuestros datos.

Actividad 2

1. Base de datos

El conjunto de datos empleados en el presente documento proviene del repositorio de la base de datos de aprendizaje automático de UCI. Los datos originales consisten en variables del vino portugués “Vinho Verde” y cuenta con 1599 observaciones de vino rojo y 4898 observaciones de vino blanco. Para cada uno se evalúa la calidad del vino (Calificación entre 0 y 10) y 11 variables químicas (cuantitativas), que son las siguientes: Acidez fija, Acidez volátil, Ácido cítrico, Azúcar residual, Cloruros, Dióxido de azufre libre, Dióxido de azufre total, Densidad , PH, sulfatos y alcohol. Específicamente se hará uso de las observaciones procedentes del vino rojo.

2. Curva de regresión

Se seleccionan cinco días consecutivos (de lunes a viernes) y se construye la curva de regresión para cada día, usando seis funciones de la base de cosenos.

3. Datos funcionales

Asuma que cada curva es la observación del día correspondiente. Así que ahora usted tiene cinco datos, uno para cada día. A este tipo de datos se les llama Datos Funcionales (Ramsay y Silverman 2005, Ferraty y Vieu 2006, Ramsay, Spencer y Hooker 2010, Ramsay, Wickham, Graves y Hooker 2011)

4. Representación gráfica

Represente sus cinco datos funcionales en un solo gráfico. Luego, calcule e interprete la media y la desviación estándar funcionales de estos cinco datos. Represente esta curvas en el mismo gráfico

Bibliografía

- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.
- Eubank (1999), Nonparametric Regression and Spline Smoothing, second edn, Marcel Dekker, New York, NY
- Olaya, J. (2012). Métodos de Regresión No Paramétrica. Universidad del Valle.
- R Core Team. (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.r-project.org/