Tarea 3: Elección de λ óptimo

Angie Rodríguez Duque & César Saavedra Vanegas

Octubre 14 de 2020

Introducción

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

La base de datos empleada se denomina "Vino Rojo". Este conjunto de datos de vino tinto consta de 1599 observaciones y 12 variables, 11 de las cuales son sustancias químicas.

2. Muestra aleatoria

Se procede a seleccionar una muestra de 60 vinos de la base de datos y se escoge las variables "acidez fija" como respuesta y "ph" como predictora

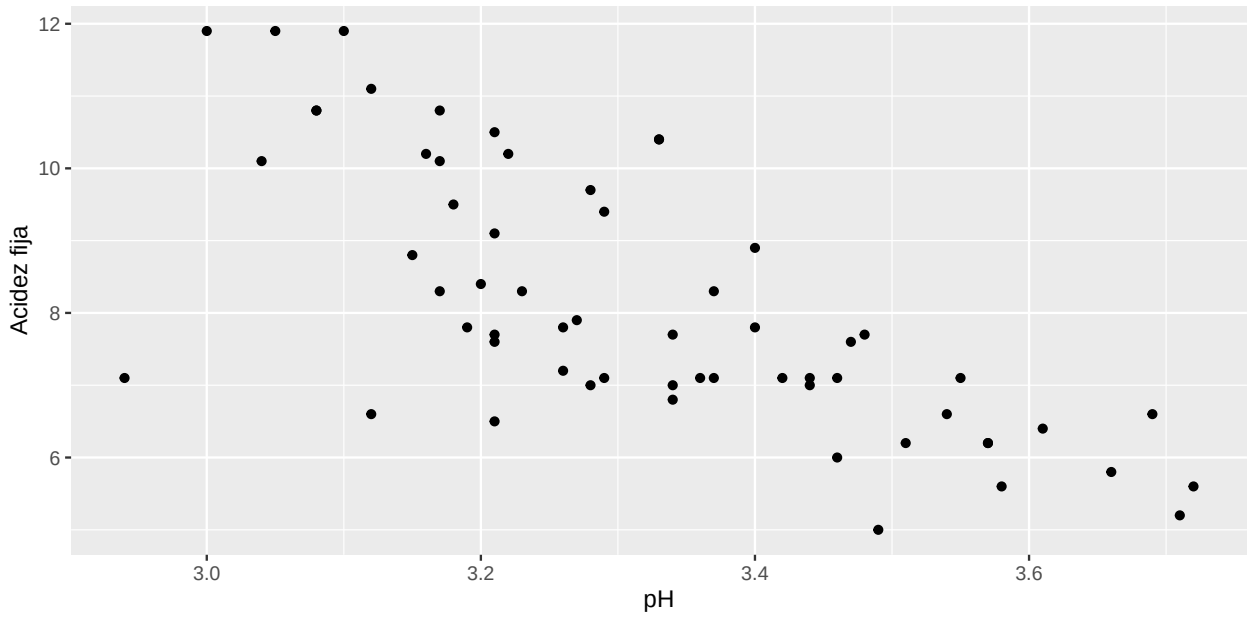
```
# Tamaño de la muestra
n <- 60
# Seleccion de la muestra
set.seed(12345)
muestra <- Datos %>% sample_n(size=n,replace=FALSE)
muestra <- muestra %>% arrange(pH)
```

Representación gráfica

A continuación se procede a graficar el comportamiento de diagrama de dispersion

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

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



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.

```
lambda <- 28
all.R <- all.R(x, lambda)
all.R
```

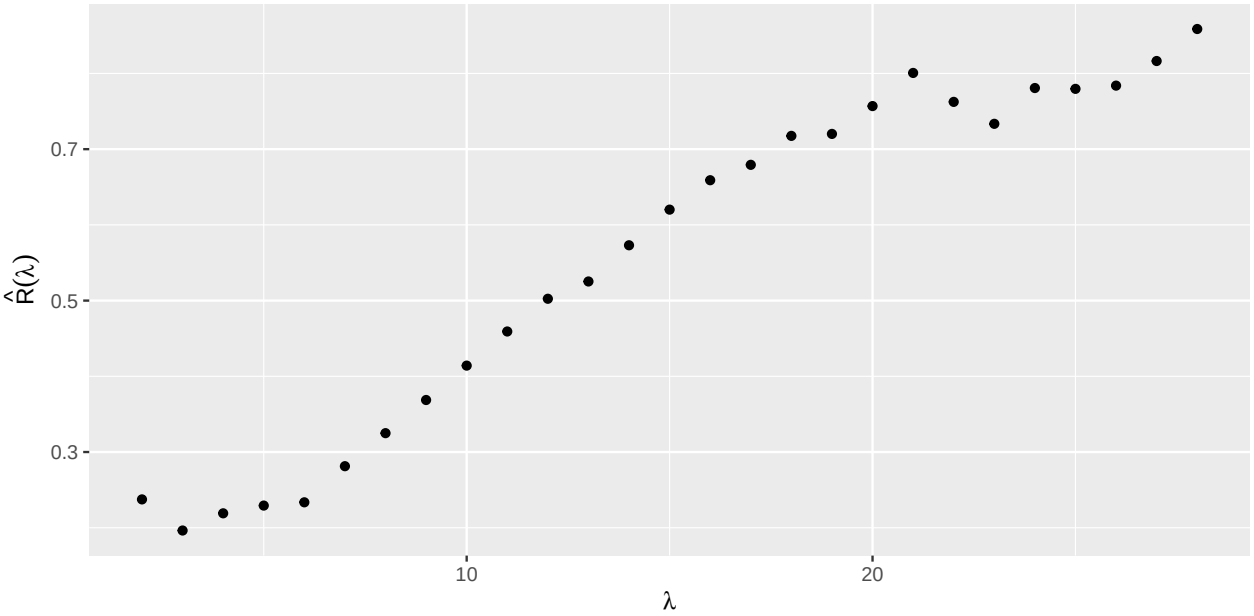
```
##         UBRE           CV      GCV LAMBDA
## 1  0.2374374 1.679390e+00 1.688573      2
## 2  0.1963344 1.629180e+00 1.649753      3
## 3  0.2190494 1.651224e+00 1.680319      4
## 4  0.2292779 1.656378e+00 1.697177      5
## 5  0.2336055 1.644459e+00 1.706858      6
## 6  0.2812813 1.686110e+00 1.771622      7
## 7  0.3249054 1.736705e+00 1.834760      8
## 8  0.3687135 1.804683e+00 1.901791      9
```

```
## 9  0.4141164 1.895187e+00 1.975067     10
## 10 0.4592090 1.966977e+00 2.052335     11
## 11 0.5024213 2.192840e+00 2.131458     12
## 12 0.5252461 2.096499e+00 2.182303     13
## 13 0.5730721 2.305738e+00 2.278137     14
## 14 0.6201230 5.137728e+00 2.379051     15
## 15 0.6589800 1.366162e+01 2.471653     16
## 16 0.6793066 3.206695e+00 2.534318     17
## 17 0.7175330 1.794698e+03 2.636750     18
## 18 0.7201338 2.061520e+03 2.669987     19
## 19 0.7568830 1.160571e+03 2.780127     20
## 20 0.8006563 4.934097e+05 2.914822     21
## 21 0.7624407 8.954444e+06 2.855627     22
## 22 0.7334757 4.081682e+05 2.810014     23
## 23 0.7807420 3.117660e+08 2.966609     24
## 24 0.7796438 3.033008e+09 2.994636     25
## 25 0.7838992 2.231881e+10 3.037549     26
## 26 0.8164224 1.320724e+10 3.173689     27
## 27 0.8587045 7.919373e+13 3.355489     28
```

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 del método UBRE. A partir de lo anterior, se puede decir que $\hat{\mu}_3$ es una buena aproximación a μ.

```
ggplot()+
  geom_point(data = all.R, aes(x = LAMBDA, y = UBRE)) +
  labs(x =  expression(lambda), y = expression(hat(R)(lambda)))
```



Finalmente se observa mediante los graficos que el valor de λ que minimiza las estimaciones segun el criterio de UBRE es un valor de $\lambda = 3$.

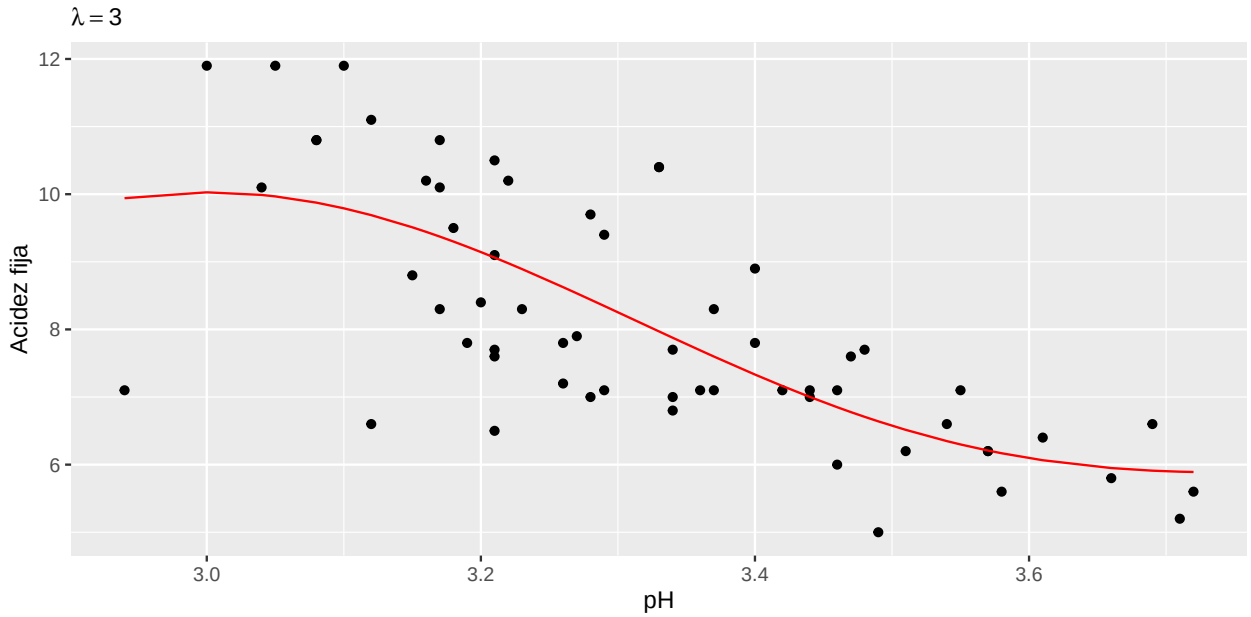
5. Estimación del modelo de regresión no paramétrica

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)) +
  geom_line(data = x, aes(x =pH, y = fitted), col="red") +
  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, podríamos decir que μ_3 es una buena aproximación a μ.

6. Interpretaciones

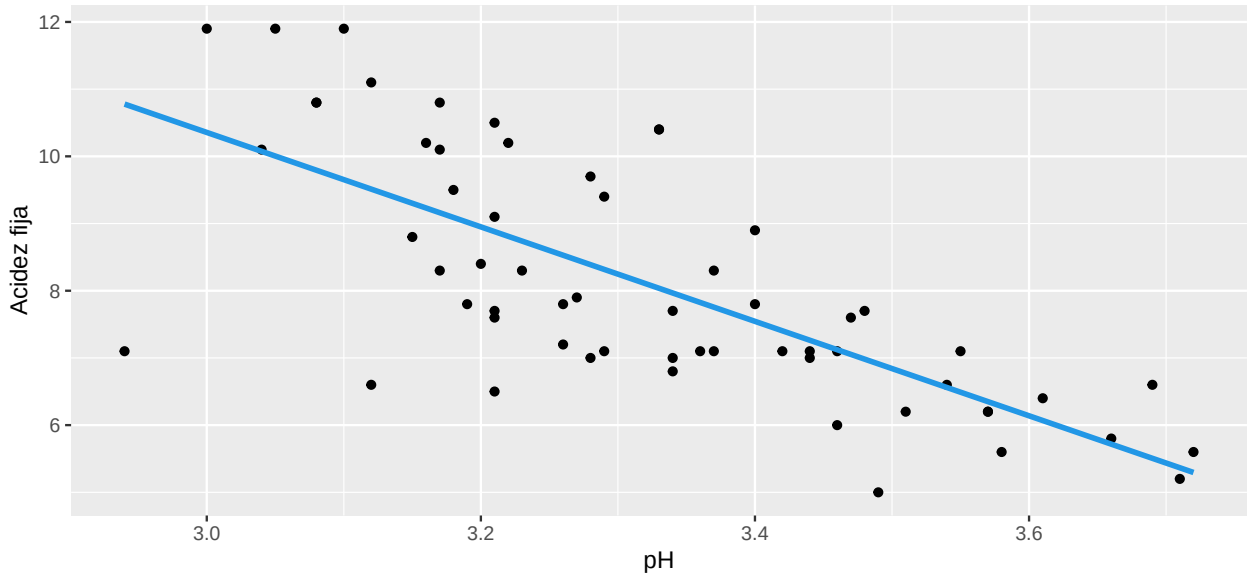
A partir del modelo anterior se puede decir que:

- 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

```
qplot(x = pH, y = fixed.acidity,data = muestra,
      main = "", ylab = "Acidez fija",
      xlab = "pH", geom = c("point"),
      method = "lm") + geom_line(aes(y=pHp), lwd = 1.2, color = 4)
```



Bibliografía

- 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/
- Eubank (1999), Nonparametric Regression and Spline Smoothing, second edn, Marcel Dekker, New York, NY