Inferencia Estadística (Fecha: 11/10/2023)

Tarea #4

Estudiante: Juan Javier Monsivais Borjón *ID:*

Problema 1

a) Sea $X \sim$ Exponencial(β). Encuentre $P(|X - \mu_X| \geq k\sigma_X)$ para $k > 1$. Compare esta probabilidad con la que obtiene de la desigualdad de Chebyshev.

b) Sean $X_1, \ldots, X_n \sim$ Bernoulli(p) y $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Usando las desigualdades de Chebyshev y Hoeffding, acote $P(|\bar{X}_n - p| > \varepsilon)$. Demuestre que para n grande la cota de Hoeffding es más pequeña que la cota de Chebyshev. ¿En qué beneficia esto?

(Solución)

(a)

Dada nuestra función de distribución:

$$f_x(x) = \frac{1}{\beta} e^{\frac{x}{\beta}} \quad x \geq 0$$

Sabemos para esta distribución la media y la desviación son iguales:

$$\mu_x = \beta$$
$$\sigma_x = \beta$$

Podríamos subdividir en casos la siguiente expresión:

$$\begin{aligned} P(|X - \beta| \geq k\beta) = \\ = P(X - \beta > k\beta) + P(-X + \beta > k\beta) = \\ = P(X > k\beta + \beta) + P(X < \beta - k\beta) \end{aligned}$$

Lo anterior por ser eventos disjuntos, y además el último término es igual a cero dado que $x > 0$, tendremos pues:

$$\begin{aligned} P(X > \beta(k+1)) = e^{-\beta(k+1)\cdot\frac{1}{\beta}} = \\ = e^{-(k+1)} \end{aligned}$$

La desigualdad de Chevyshev nos da la cota:

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}$$

Es claro que:

$$e^{-(k+1)} < \frac{1}{k^2}$$

Es lo que se espera dado que el lado izquierdo es propio de la distribució, veamos para algunos k:

k	Exponential Distribution	Chebyshev's Inequality
2	0.0498	0.2500
3	0.0183	0.1111
4	0.0067	0.0625

Cuadro 1

Para cuatro sigma ya hay una diferencia de un orden de magnitud completo.

(b)

Para comparar las desigualdades, empecemos con la desigualdad de Chevyshev:

$$Var(X_i) = p(1-p)$$
$$Var(\bar{X_n}) = \frac{p(1-p)}{n}$$

La desigualdad de Chevyshev, se define como:

$$P(|X - \mu_X| \geq t) \leq \frac{\sigma^2}{t^2} \tag{1}$$

Ahora sistituyendo para obtener la cota de Chevyshev de la suma de variables Bernoulli:

$$P(|X - \mu_X| \geq \epsilon) \leq \frac{Var(\bar{X})}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}$$

A lo más la multiplicación de dos probabilidades es la simétrica $(0.5 \cdot 0.5)$, es decir: $p \cdot q = p(1-p) \leq \frac{1}{4}$, por lo que podemos acotar aún más:

$$P(|X - \mu_X| \geq \epsilon) \leq \frac{1}{4n\epsilon^2}$$

Por otro lado para Hoeffding, tenemos:

$$P(|\bar{X} - p| \geq \epsilon) \leq 2e^{(-2n\epsilon^2)} \tag{2}$$

Podemos ahora comparar ambas funciones obtenidas de las desigualdades y compararlas, mediante el límite de la razón de ambas:

$$\lim_{n\to\infty} \frac{f(n)}{g(n)}$$

Con $f(n) = \frac{1}{4n\epsilon^2}$ y $g(n) = 2e^{-2n\epsilon^2}$

$$\lim_{n\to\infty} \frac{\frac{1}{4n\epsilon^2}}{2e^{-2n\epsilon^2}} = \infty \tag{3}$$

Lo que nos dice que $g(n)$ decrese mucho más rápido que $f(n)$.

La ventaja de esto es que para n grande tenemos una cota mucho más ajustada con la desigualdad de Hoefdding. Que se acerca más a la función verdadera.

Problema 2

Sean $X_1, \ldots, X_n \sim$ Bernoulli(p). a) Sea $\alpha > 0$ fijo y defina

$$\epsilon_n = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}.$$

Sea $\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Defina $C_n = (\hat{p}_n - \epsilon_n, \hat{p}_n + \epsilon_n)$. Use la desigualdad de Hoeffding para demostrar que

$$P(C_n \text{ contiene a } p) \geq 1 - \alpha.$$

Diremos que C_n es un $(1-\alpha)$-intervalo de confianza para p. En la práctica, se trunca el intervalo de tal manera de que no vaya debajo del 0 o arriba del 1.

b)Sea $\alpha = 0.05$ y $p = 0.4$. Mediante simulaciones, realice un estudio para ver qué tan a menudo el intervalo de confianza contiene a p (la cobertura). Haga esto para $(n = 10, 50, 100, 250, 500, 1000, 2500, 5000, 10000)$. Grafique la cobertura contra (n).

c) Grafique la longitud del intervalo contra n. Suponga que deseamos que la longitud del intervalo sea menor que 0.05. ¿Qué tan grande debe ser n?

(Solución)

(a)

Para que p se encuentre dentro fuera de C_n debe ocurrir que $P(|\hat{p} - p| \geq \epsilon_n)$

Ya sabemos por el ejercicio anterior que se cumple lo siguiente:

$$P(|\hat{p} - p| \geq \epsilon_n) \leq 2e^{-2n\epsilon_n^2} \tag{4}$$

De aquí podemos sustituir el valor del error ϵ_n:

$$\begin{aligned} P(|\hat{p} - p| \geq \epsilon_n) \leq 2e^{-2n\left(\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\right)_n^2} = \\ = 2\exp\left(-\frac{2n}{2n} log\left(\frac{1}{\alpha}\right)\right) = \\ = 2\exp\left(\log\left(\frac{\alpha}{2}\right)\right) = \\ 2 \cdot \frac{\alpha}{2} = \alpha \end{aligned}$$

Y de esta manera, para p se encuentre dentro de C_n, lo único que debemos hacer es obtener el complemento:

$$P(|\hat{p} - p| < \epsilon_n) \geq 1 - \alpha$$

(b)

Ahora veamos cómo se comporta este estadístico en las simulaciones y cuántos puntos se logran cubrir dependiendo del "n". A Continuación se muestran los gráficos respectivos de las simulaciones para cada n.

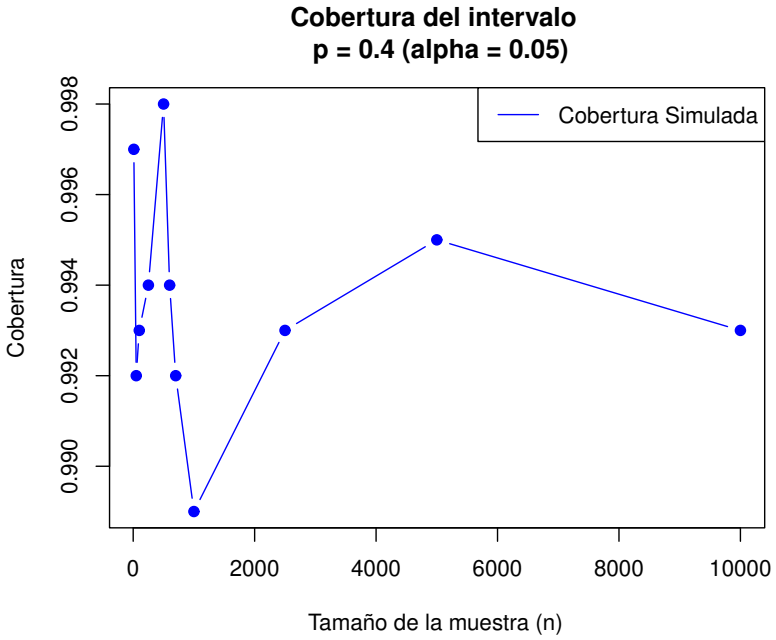

Figura 1

En general vemos que se dispersan de manera prácticamente uniforme, a lo largo del eje x, muy cercanos a 1, de hecho esperamos justo que no haya demasiados datos fuera de nuestro intervalo de confianza.

Y a continuación mostramos el gráfico de los n y su tamaño de intervalo respectivo calculado a partir del estadístico ϵ, siendo el tamaño del intervalo el doble del estadístico.

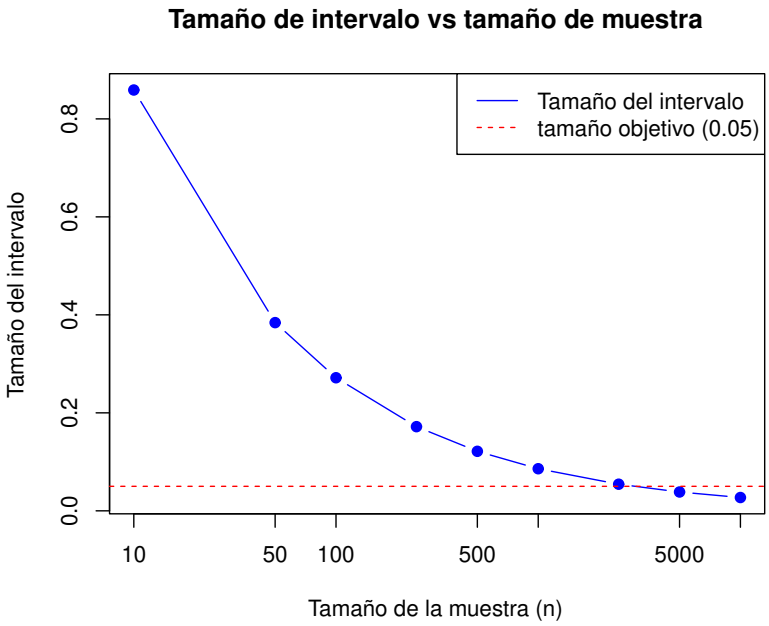

Figura 2

Como podemos ver para lograr un tamaño de intervalo de 0.05 es necesario aproximadamente $n = 5000$.

Problema 3

Considera el problema 5 de la Tarea 3. Utilizando la desigualdad de Dvoretzky-Kiefer-Wolfowitz, escriba una función en R que calcule y grafique una región de confianza para la función de distribución empírica. La función debe tomar como parámetros el conjunto de datos que se utiliza para construir la función de distribución empírica.

(Solución)

Para el problema ya teníamos contruida la distribución empírica, por lo que lo único que necesitamos es definir el estadístico de Dvoretzky-Kiefer-Wolfowitz, como sigue:

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)} \tag{5}$$

Con esto, simplemente construimos las bandas de confianza sumando a la ECDF teórica:

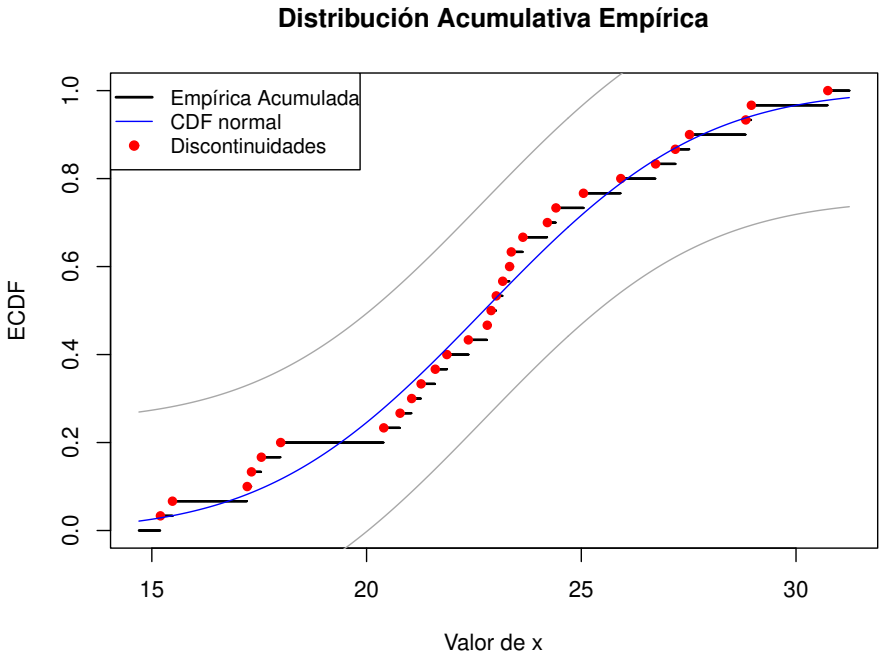

Figura 3: Bandas de confianza con cotas de Dvoretzky Kiefer Wolfowitz, con $\alpha = 0.05$

Problema 4

a) Escriba una función en R que estime una densidad por el método de kerneles. La función deberá recibir al punto x donde se evalúa al estimador, al parámetro de suavidad h, al kernel que se utilizará en la estimación y al conjunto de datos.

b) Cargue en R el archivo "Tratamiento.csv", el cual contiene la duración de los períodos de tratamiento (en días) de los pacientes de control en un estudio de suicidio. Utilice la función del inciso anterior para estimar la densidad del conjunto de datos para $h = 20, 30, 60$. Grafique las densidades estimadas. ¿Cuál es el mejor valor para h? Argumente.

c) En el contexto de la estimación de densidades, escriba una función en R que determine el ancho de banda que optimiza el ISE. Grafique la densidad con ancho de banda óptimo para el conjunto de datos de "Tratamiento.csv".

(Solución)

Para la resolución del problema, hemos de identificar que es un método no paramétrico, es decir, no se propone de antemano una distribución, sino que es un método de localidad, y ajuste del ancho de las celdas que se usará, ello nos permite particularizar en problemas mucho más complejos, adecuando estos parámetros anteriormente mencionados.

Nuestro objetivo al final de cuentas es obtener una función de densidad estimada a partir de los datos, que tienda a ser la función real para n muy grandes (lo que sea que muy grande signifique).

Por simplificación usaremos el kernel gaussiano, que es el kernel más usado:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

Donde u representa la lejanía del punto evaluado a los otros, estandarizados en h:

$$u = \frac{x - x_i}{h} \quad h > 0$$

Y con ello contruimos nuestra función con el método de Kernels definido como:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(u) \quad x \in R, \quad h > 0$$

A continuación se muestran las gráficas obtenidas, en las que se comparan las estimaciones mediante Kernel y los histogramas:

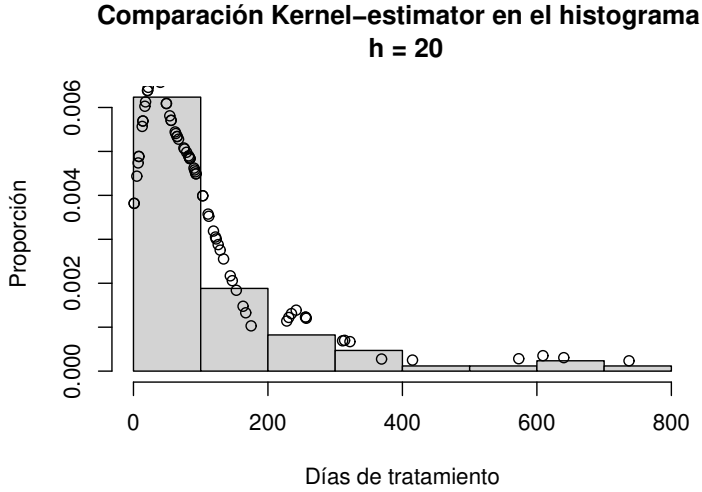

Figura 4

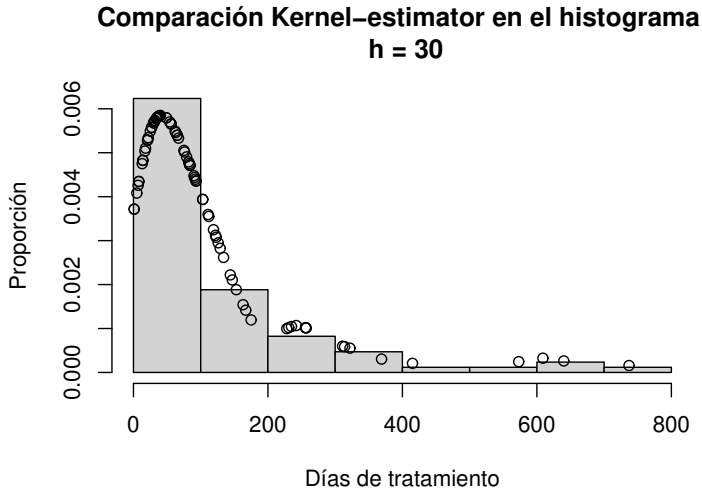

Figura 5

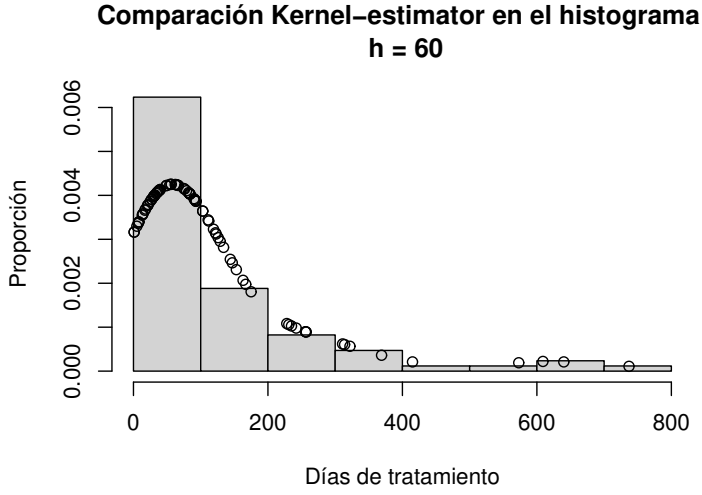

Figura 6

Existen dos maneras de validar nuestro modelo, una es simplemente viendo el ajuste con los datos, ("rule of thumb"). Que no resulta muy formal, la otra es la la validación cruzada, que puede subdividirse en múltiples. Esta última la veremos más adelante en la que usaremos el ISE.

Por lo pronto no podemos hacer más que equiparar la suavidad de nuestra función de densidad con su ajuste, para nuestro caso, parece ser que un $h = 30$ sería una buena elección del ancho, ya que no parece estar sobrealimentado, se percibe buena suavidad sobretodo en la media de los datos y no se aleja del ajuste de los mismos. En cambio para los valores $h = 20$ y $h = 60$ tenemos justo esos casos no deseados; para 20, tenemos que un sobreajuste debido a la alta localidad, y por otro lado, para 60, tenemos un exceso de suavidad, cuestión que evitaría poder realizar una buena estimación.

Una vez realizadas estas observaciones podemos pasar al siguiente punto, que es justamente minimizar el **ISE**, recordemos la expresión:

$$\text{ISE } = \int_{-\infty}^{\infty} [\hat{f}(y) - f(y)]^2 dy. \tag{6}$$

Básicamente lo que nos permitiría esta expresión es medir qué tan alejados nuestra función de densidad estimada está de la teórica, en este punto debemos notar un gran problema y es que no tenemos la función de densidad "real" o teórica, por lo que nos es imposible usar esta expresión. Procederemos a usar un concepto conocido como validación cruzada, en particular la validación cruzada insesgada:

$$\text{UCV}(h) = R(\hat{f}) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}\,(x_i)$$

En donde R definido como:

$$R(\psi) \equiv \|\psi\|_2^2 = \int_{-\infty}^{\infty} \psi(x)^2 dx$$

Notemos que en nuestra expresión para UCV, aparece un $\hat{f}_{-i}$ lo que nos indica que debemos dejar una de las observaciones fuera, esto se hace normalmente para evitar sesgo, y un overfitting. Imaginemos que tenemos una serie de datos identico, esto nos llevaría a inflar este valor y la capacidad predictiva del modelo se va perdiendo.

Usando dicha validación cruzada obtenemos que el h que optimiza a la función es:

$$h = 14.7194$$

Ahora observemos finalmente la estimación mediante kernel con el h óptimo:

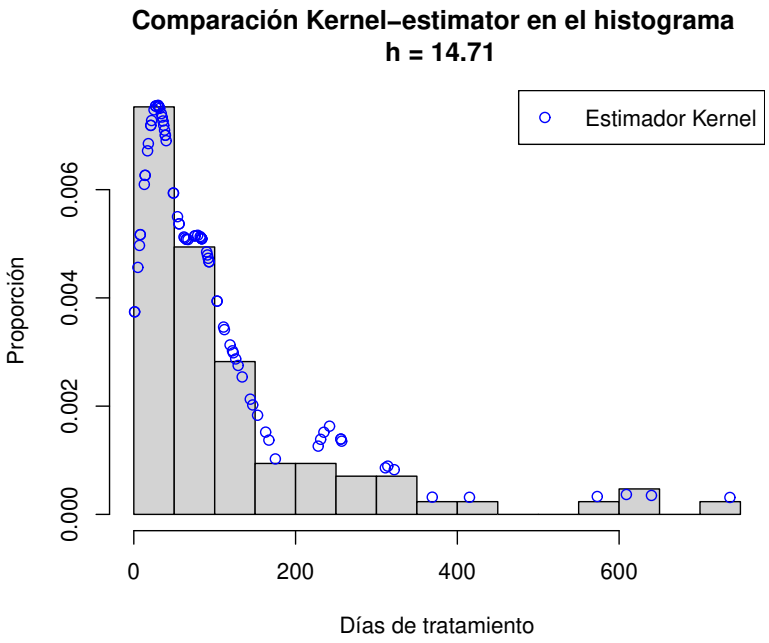

Figura 7: Caption

Problema 5

Considera el siguiente experimento en dos etapas: primero se elige un punto X con distribución uniforme en (0, 1); luego se selecciona un punto Y con distribución uniforme en (-X, X). El vector aleatorio (X, Y) representa el resultado del experimento. ¿Cuál es su densidad conjunta de (X, Y)? ¿Cuál es la densidad marginal de Y? ¿Cuál es la densidad condicional de X dado Y?

(Solución)

Primero obtengamos la función de densidad del evento X, que es simplemente la uniforme:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{si } a < x < b \\ 0 & \text{fuera del intervalo} \end{cases}$$

Por lo tanto como tenemos $b = 1$, $a = 0$ tendremos:

$$f_X(x) = \begin{cases} 1 & \text{si } 0 < x < 1 \\ 0 & \text{fuera del intervalo} \end{cases}$$

Una vez el evento ha ocurrido, se da el evento Y, que depende directamente del evento anterior en los límites y en la distribución, dado que es otra uniforme:

Ya que la uniforme de Y está definida en $(-X, X)$, tendremos entonces la condicional de Y dado X:

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{2x} & \text{si } -x < y < x \\ 0 & \text{otro} \end{cases}$$

Con lo anterior, al multiplicar la condicional y nuestra marginal de X:

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{2x} & \text{si y } -x < y < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Para obtener la marginal de Y, simplemente integramos nuestra función sobre las X:

$$f_Y(y) \int_y^1 \frac{1}{2x} dx = \frac{1}{2} ln(x) \Big|_y^1 = ln(\frac{1}{\sqrt{y}})$$

Y finalmente usamos teorema de Bayes para el último inciso:

$$f_{Y|X}(x, y) = \frac{f_{Y|X}(x, y)}{f_X(x)}$$

Tenemos pues:

$$f_{Y|X} = \frac{1}{2 \cdot x \cdot ln(\frac{1}{\sqrt{y}})}$$

Problema 6

Cargue en R el conjunto de datos `Maíz.csv`, que contiene el precio mensual de una tonelada de maíz y el precio de una tonelada de tortillas en USD. En este ejercicio, se le pedirá estimar los coeficientes de una regresión lineal simple.

a) Calcule de forma explícita la estimación de los coeficientes utilizando el método de mínimos cuadrados y ajuste la correspondiente regresión. Luego, proporcione sus conclusiones.

b) Calcule de forma explícita la estimación de los coeficientes utilizando el método de regresión no paramétrica tipo kernel (consulte Nadaraya, E. A. (1964). 'On Estimating Regression'. Theory of Probability and its Applications. 9 (1): 141–2. doi:10.1137/1109020) y ajuste la correspondiente regresión. Luego, proporcione sus conclusiones.

c) Compare ambos resultados. ¿Qué diferencias observa?

(Solución)

Para proceder con el análisis del problema, debemos observar dar por hecho que el precio de las tortillas debe ser proporcional al precio del maíz, de no ser así, no tendríamos una relación lineal entre estas variables y una regresión lineal **directa** no sería viable. Podríamos hablar de una linealización del modelo "Estirar los ejes" dependiendo de la relación que exista entre ambas variables. Pero para nuestro caso haremos esa suposición (relación lineal).

Si nosotros deseamos obtener una función lineal $g(x)$, que mejor ajuste a nuestros datos usaremos la minimizacion de la diferencia entre nuestros puntos y la línea que deseamos, de esta manera podremos obtener los coeficientes:

$$g(x) = \alpha x + \beta$$

$$\alpha = \frac{\Sigma_i x_i y_i - \Sigma_i x_i \Sigma_i y_i}{\Sigma_i {x_i}^2 - \left(\Sigma_i x_i\right)^2} \tag{7}$$

$$\beta = \frac{1}{n}\Sigma_i y_i - \alpha \frac{1}{n}\Sigma_i x_i \tag{8}$$

Esto para una regresión tipo mínimos cuadrados, en cambio si deseamos la aproximación mediante Kernel, usaremos lo siguiente:

$$\bar{y}_n(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h(n)}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h(n)}\right)}, \tag{9}$$

De este método no podemos obtener directamente los coeficientes, de hecho veremos que la ventaja de este método es que, nosotros podemos observar y cambiar la suavidad del ajuste (que

también puede resular en algo negativo), ya que nosotros podemos elegir el tamañp de nuestra celda $h(n)$.

En particular se ha elegido $h(n) = 2$ con un kernel gaussiano, a continuación se muestran los resultados:

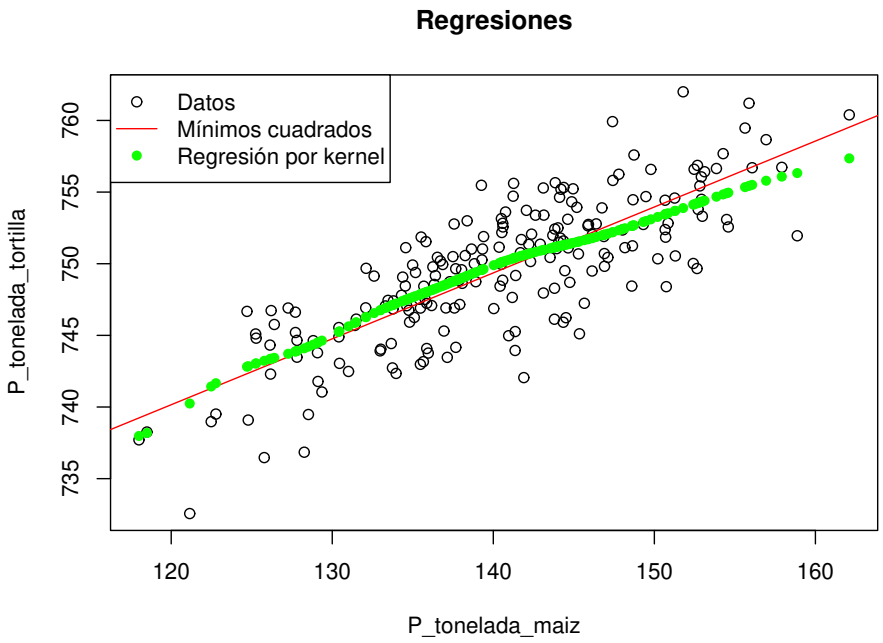

Figura 8: Comparativa entre regresión por mínimos cuadrados y mediante Kernel Gaussiano.

Los parámetros obtenidos por mínimos cuadrados son los siguientes:

$$\alpha = 0.46$$
$$\beta = 684.95$$

Lo que nos representaría el valor de alpha es nuestra constante de proporcionalidad, es decir, la velocidad a la que aumenta el precio de la tortilla dado el del maíz.

Problema Ejercicio 1

Se desea verificar la programación del tiempo que debe durar la luz verde en un semáforo ubicado en una intersección específica que permite giros a la izquierda. Dado que no se dispone de información al respecto, se ha enviado a un estudiante graduado para realizar observaciones sobre la cantidad de automóviles (X) y camiones (Y) que llegan en un ciclo completo (desde que la luz se pone en verde hasta que vuelve a ponerse en verde). A partir de estas observaciones, se construye una tabla de distribución conjunta. Se plantean una serie de preguntas con el objetivo de evaluar la eficiencia del ciclo planificado.

Cuadro 2: Tabla de Distribución Conjunta de Automóviles (X) y Camiones (Y) que llegan por Ciclo

$p(x,y)$	$x=0$	$x=1$	$x=2$
$y=0$	0.025	0.015	0.010
$y=1$	0.050	0.030	0.020
$y=2$	0.125	0.075	0.050
$y=3$	0.150	0.090	0.060
$y=4$	0.100	0.060	0.040
$y=5$	0.050	0.030	0.020

(Solución)

(1)

Para el primer ejercicio podemos sumar columnas y filas y verificar que se cumple:

$p(x,y)$	$x=0$	$x=1$	$x=2$	Suma por fila
$y=0$	0.025	0.015	0.010	0.050
$y=1$	0.050	0.030	0.020	0.100
$y=2$	0.125	0.075	0.050	0.250
$y=3$	0.150	0.090	0.060	0.300
$y=4$	0.100	0.060	0.040	0.200
$y=5$	0.050	0.030	0.020	0.100
	0.500	0.300	0.200	1.000

Con lo cual cumple para ser una tabla de probabilidades conjuntas

(2)

Lo anterior contesta también al cálculo de las marginales de Y y X, la última columna representa la marginal de Y y la fila de debajo representa la marginal de X

(3)

Para calcular que haya 1 carro y un camión en un ciclo, basta con encontrar la intersección de ambas variables, en nuestro caso es $P(X = 1, Y = 1) = 0.03$

(4)

Si el límite de aceptación es 5, basta con sumar las probabilidades que cumplen con llenar el carril dadas las condiciones, pero será más sencillo usar el complemento:

$$P(X + 3Y \geq 5) = 1 - P(X + 3Y < 5) = 1 - 0.13 = 0.87$$

(5)

Podemos ver este último punto como si estuviésemos en un mini-universo en el que ya se dio cierto evento, es decir, ya nos encontramos en una fila, con ello en mente solo tenemos que observar la proporción que representa nuestro evento en ese universo, es decir:

$$P(X = 1|Y = y) = \frac{P(X = 1 \cap Y = y)}{P(Y = y)}$$

Para $Y = 1$:

$$\begin{aligned}
P(X = 1|Y = 0) &= \frac{0.015}{0.05} = 0.3 \\
P(X = 1|Y = 1) &= \frac{0.03}{0.1} = 0.3 \\
P(X = 1|Y = 2) &= \frac{0.075}{0.250} = 0.3 \\
P(X = 1|Y = 3) &= \frac{0.090}{0.3} = 0.3 \\
P(X = 1|Y = 4) &= \frac{0.06}{0.2} = 0.3 \\
P(X = 1|Y = 5) &= \frac{0.03}{0.1} = 0.3
\end{aligned}$$

Referencias