Tarea 4 Inferencia Estadística

Code

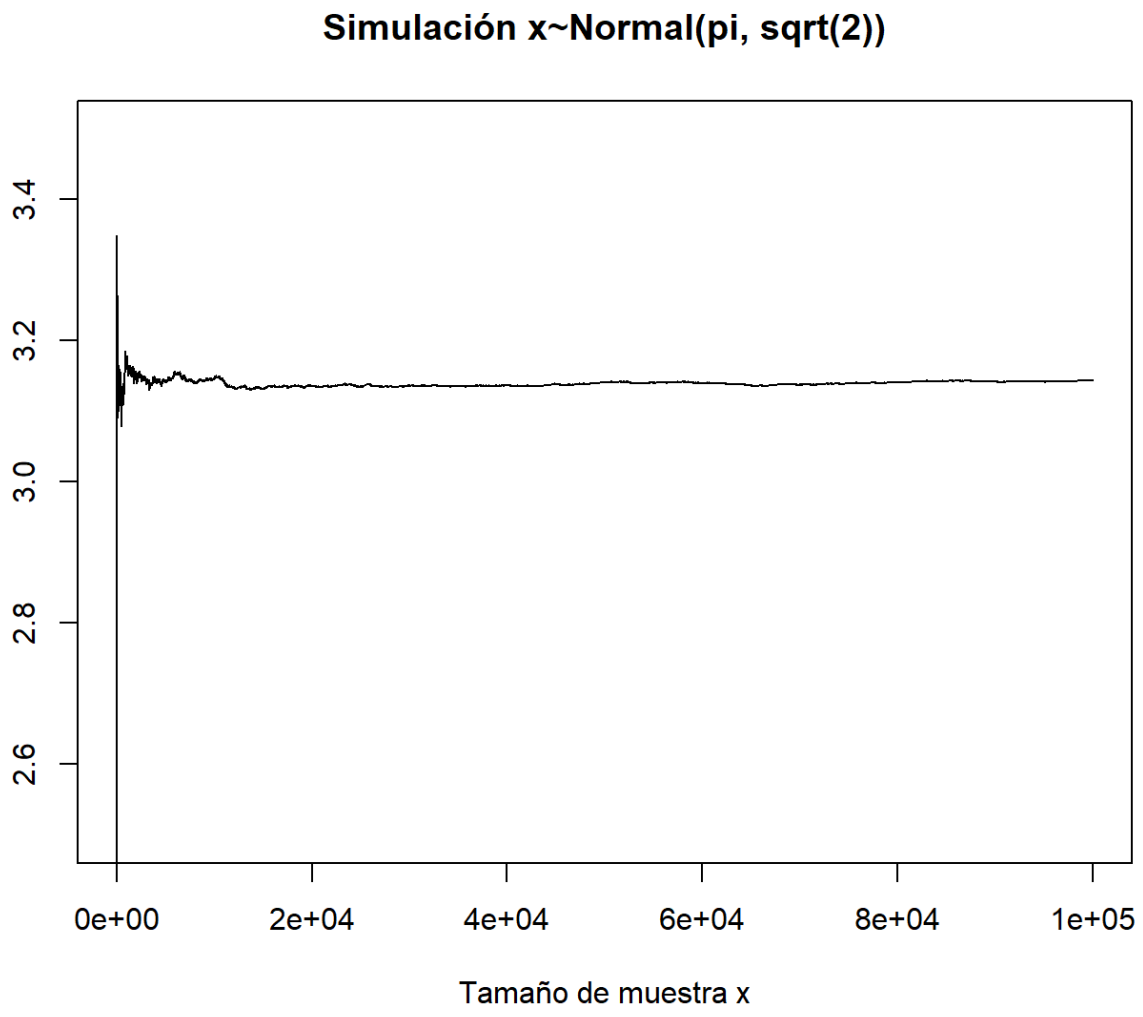
Hairo Ulises Miranda Belmonte

2 de Octubre de 2018

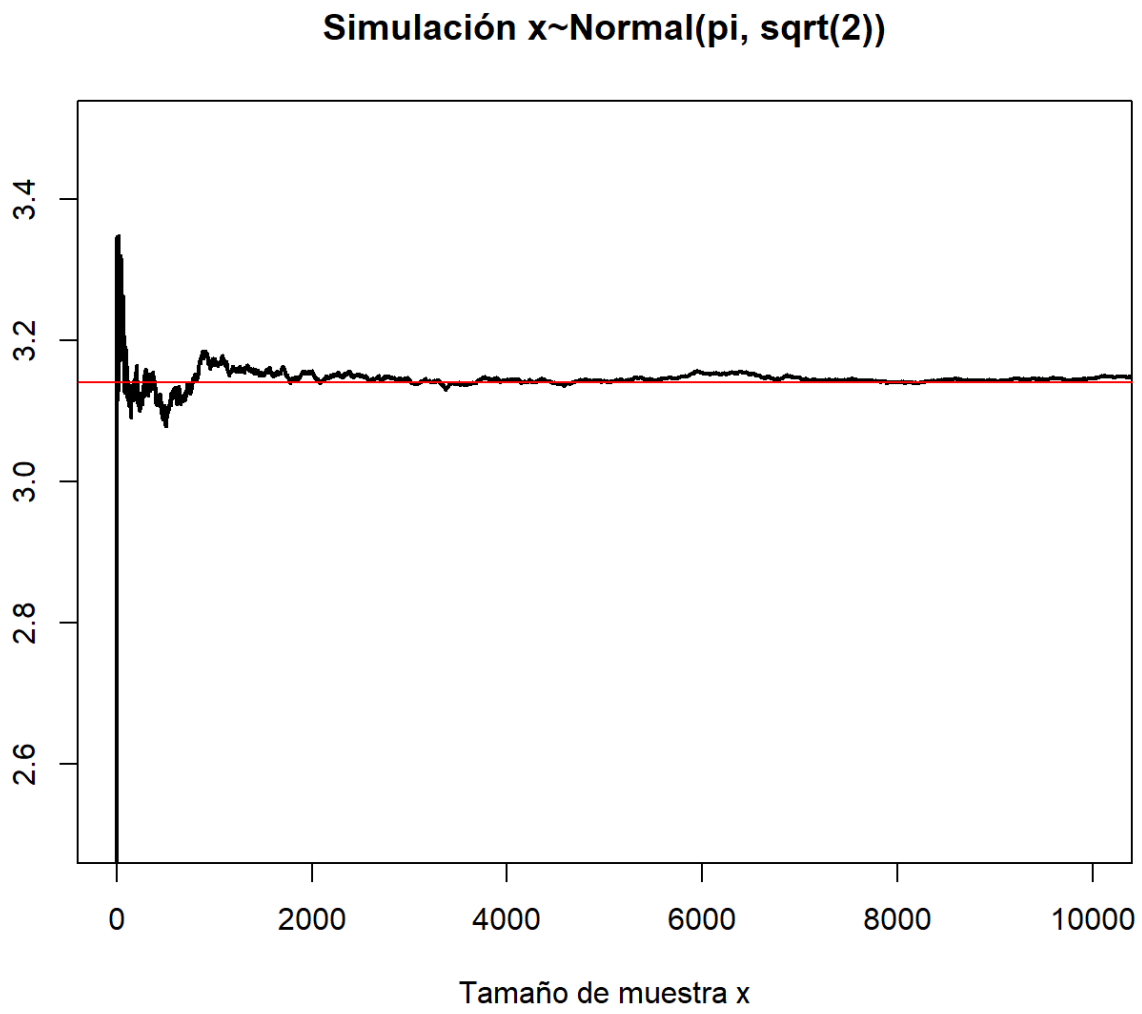
EJERCICIO 1

b. Simule una muestra $x_1, \ldots, x_n$ de una v.a. Normal de tamaño $n = 10^5$. Defina $y_m = \Sigma_{i=1}^{n} x_i / n$ y grafique esta cantidad. ¿Que observa? ¿Como está esto relacionado con la LGN?

Code


Simulación x~Normal(pi, sqrt(2))
3.4
3.2
3.0
2.8
2.6
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Tamaño de muestra x

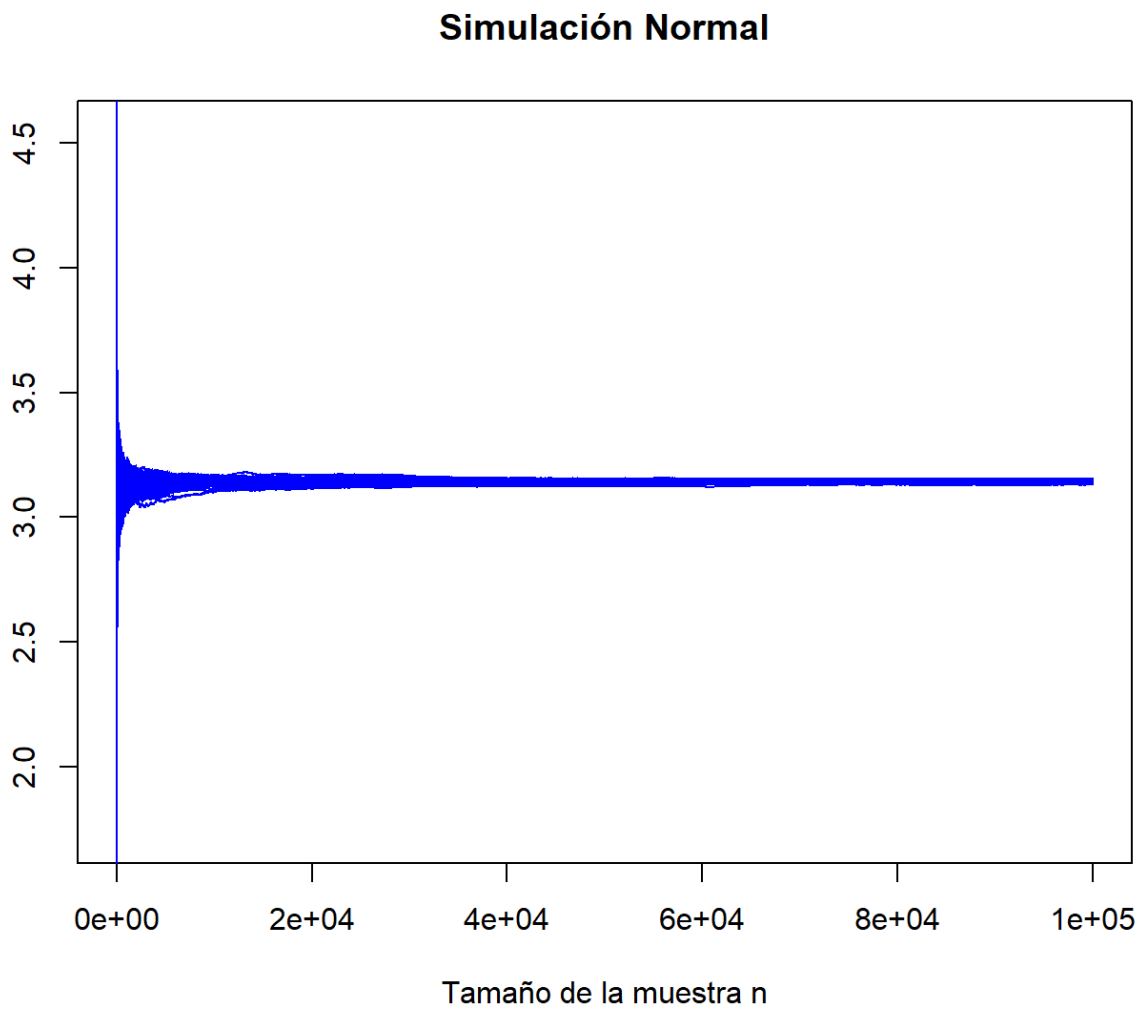
Code



Se observa que cuando n -el numero de simulaciones- es grande, la media mustral se aproxima a la media poblacional (π). A su vez, lo anterior se puede relacionar con la convergencia en probabilidad,dado a que una secuencia de variables aleatorias se aproxima a una variable aleatoria -(i.e, la media muestral es una variable aleatoria enerada por una combinación lineal de una secuencia de variables aleatorias)- cuando el tamaño de muestra tiende a infinito.

c. Repita el procesoanterior 100 veces y grafique y de cada iteración sobre una misma gráfica

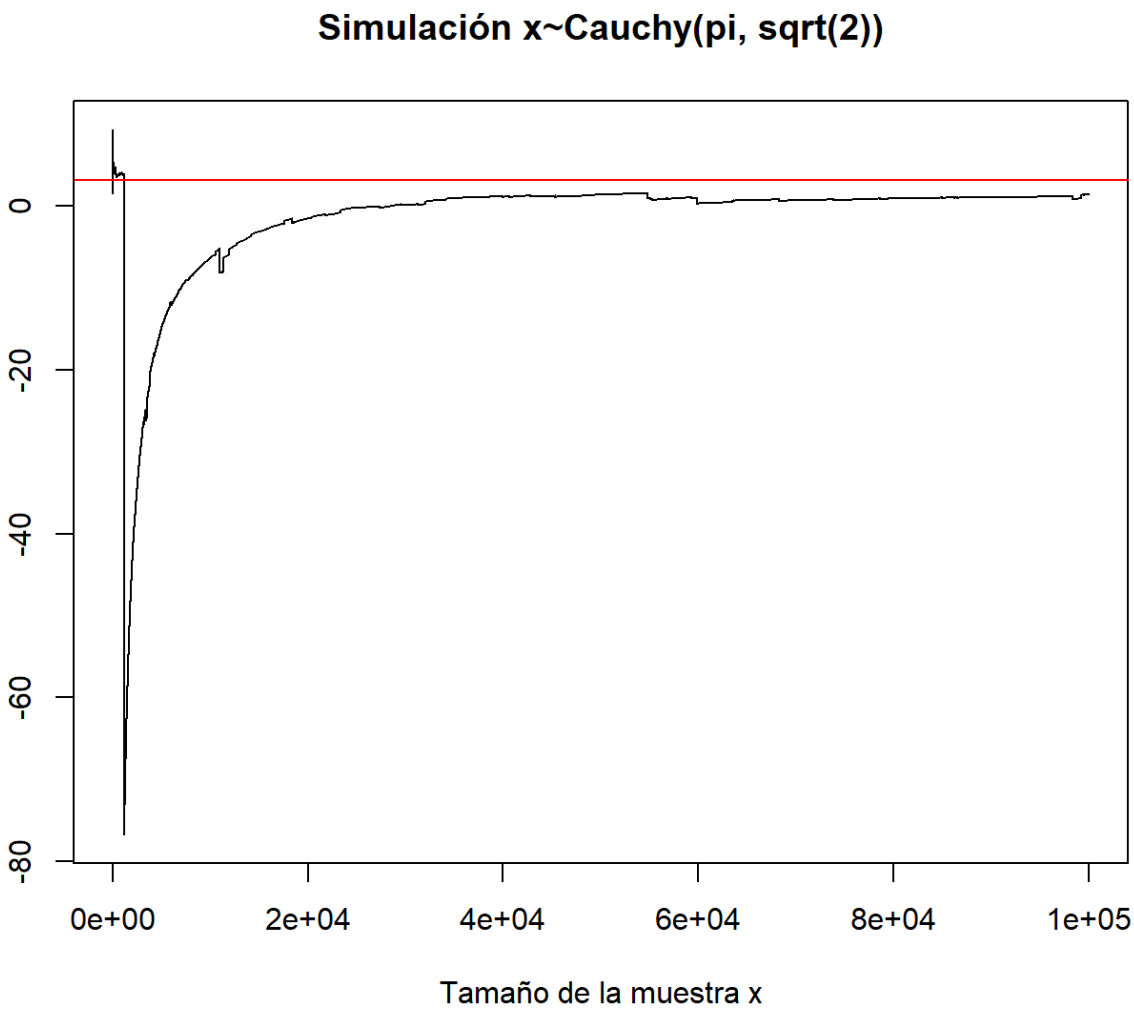
Code



Code

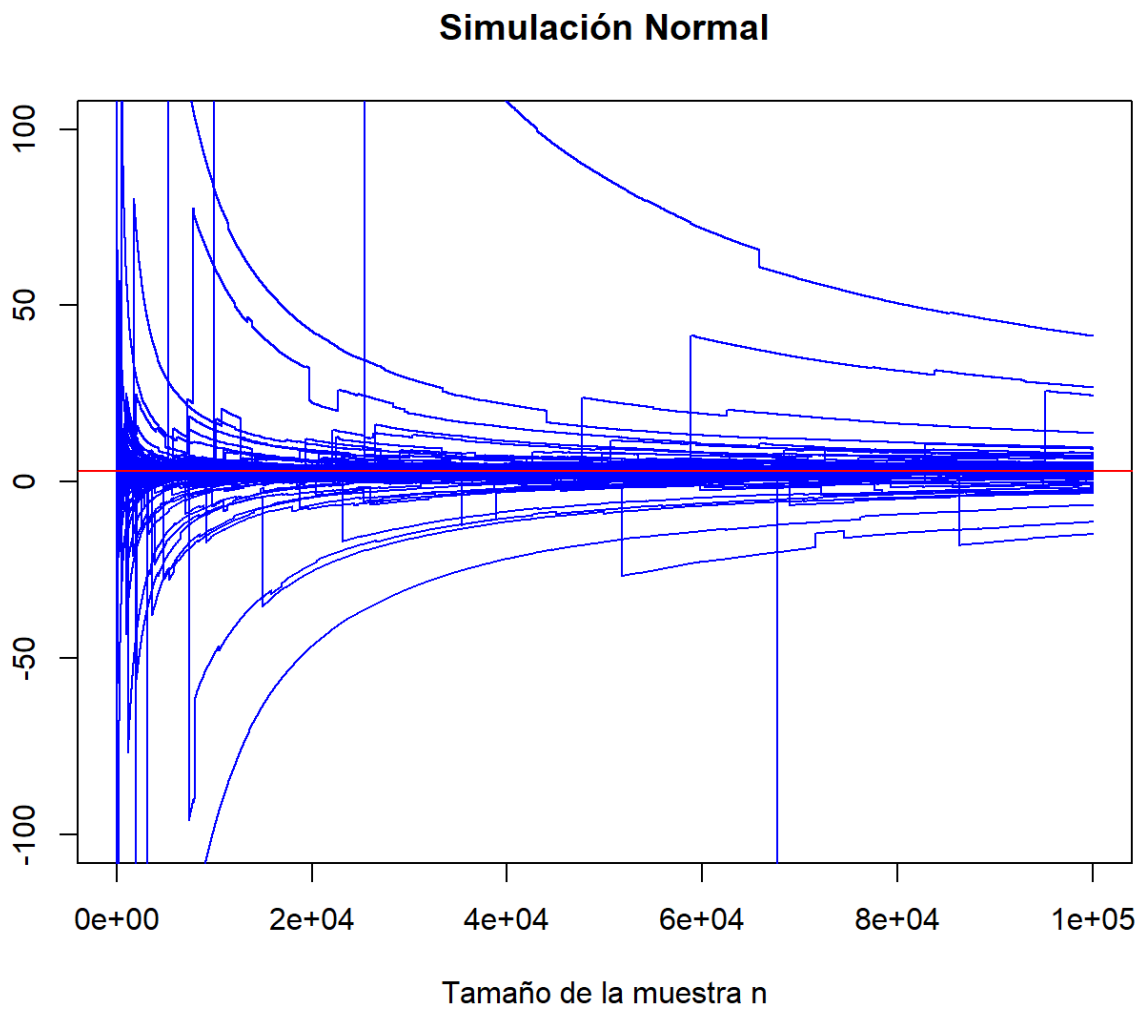
d. Repita los dos incisos anteriores para una distribución cauchi ¿Qué observa?

Code



Repita el procesoanterior 100 veces y grafique y de cada iteración sobre una misma gráfica.

Code



Caso contrario a la media muestral generada con variables aleatorias normales, al tener un conjunto de variables aleatorias con distribución cauchy, el estadistico -media muestral-, no coverge en ´probabilidad (i.e, conjunto de variables aleatorias tienden a una variable aleatoria). Lo que sucede, al contrario de las normales, es que al incrementar el número de muestras, la media muestral no se aproxima a su media poblacional. Esto se puede explicar, dado que la distribución cauchy se encuentra ausente de momentos. Otro determinante, es el no cumplimiento de la ley de los grandes números (debil), dado a que no sólo se es necesario un conjunto de datos iid, también se require que las variables aleatoria presenten media y varianza finitas, (que existan primer y segundo momento).

EJERCICIO 3

b. Sea $\alpha = 0.05$ y $p = 0.4$. Mediante simulaciones, realice un estudio para ver que 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.

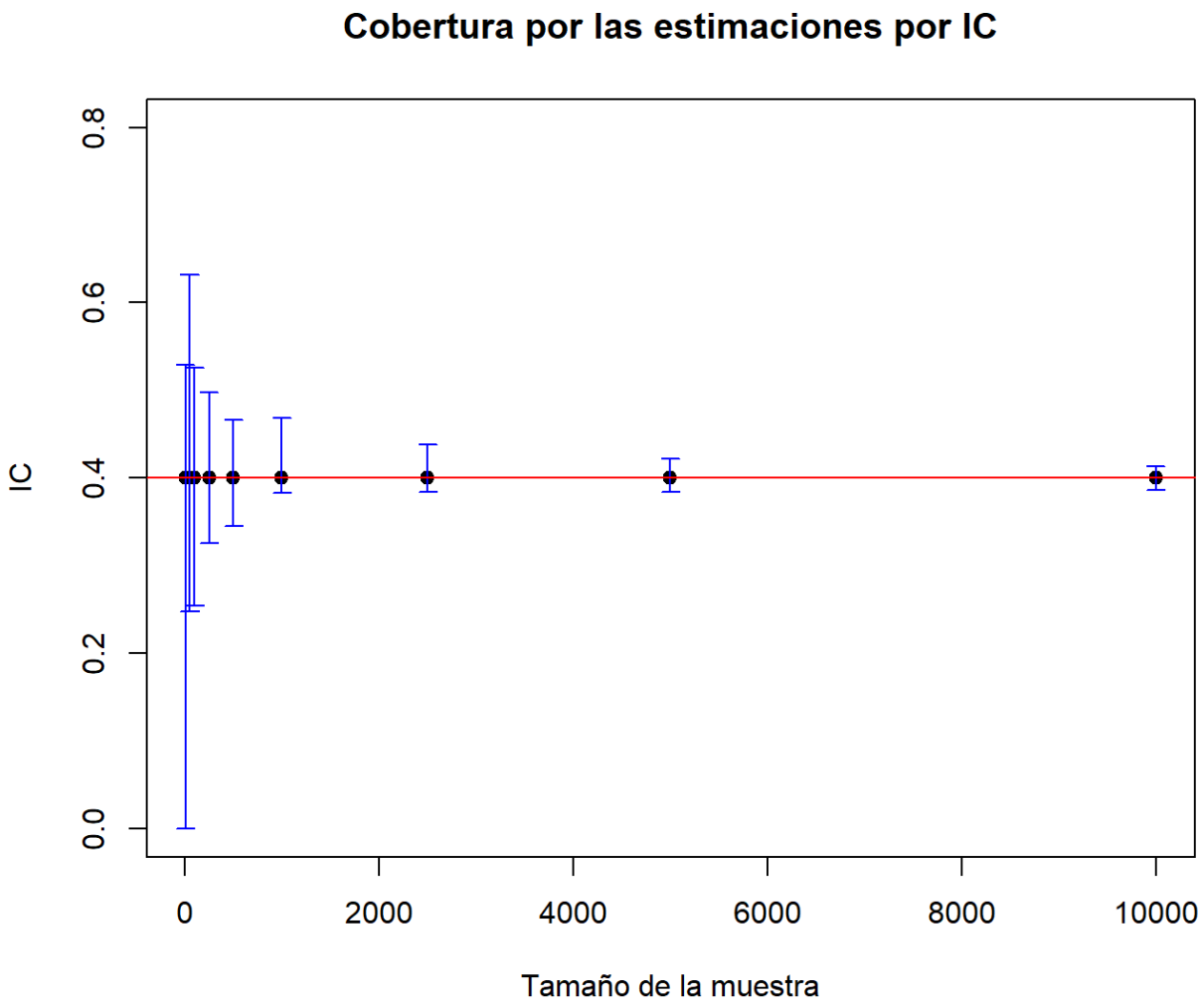
Code

```
##           Fuera IC Dentro IC
## n = 10         172     99828
## n = 50         559     99441
## n = 10         553     99447
## n = 250        552     99448
## n = 500        544     99456
## n = 1000       593     99407
## n = 2500       595     99405
## n = 5000       511     99489
## n = 10000      584     99416
```

Como se observa en la tabla anterios, a un 95% de confianza, la mayoría de de veces el intervalo de confinza contiene a p.

Se gráfica la cobertura contra el numero de muestra " n "

Code

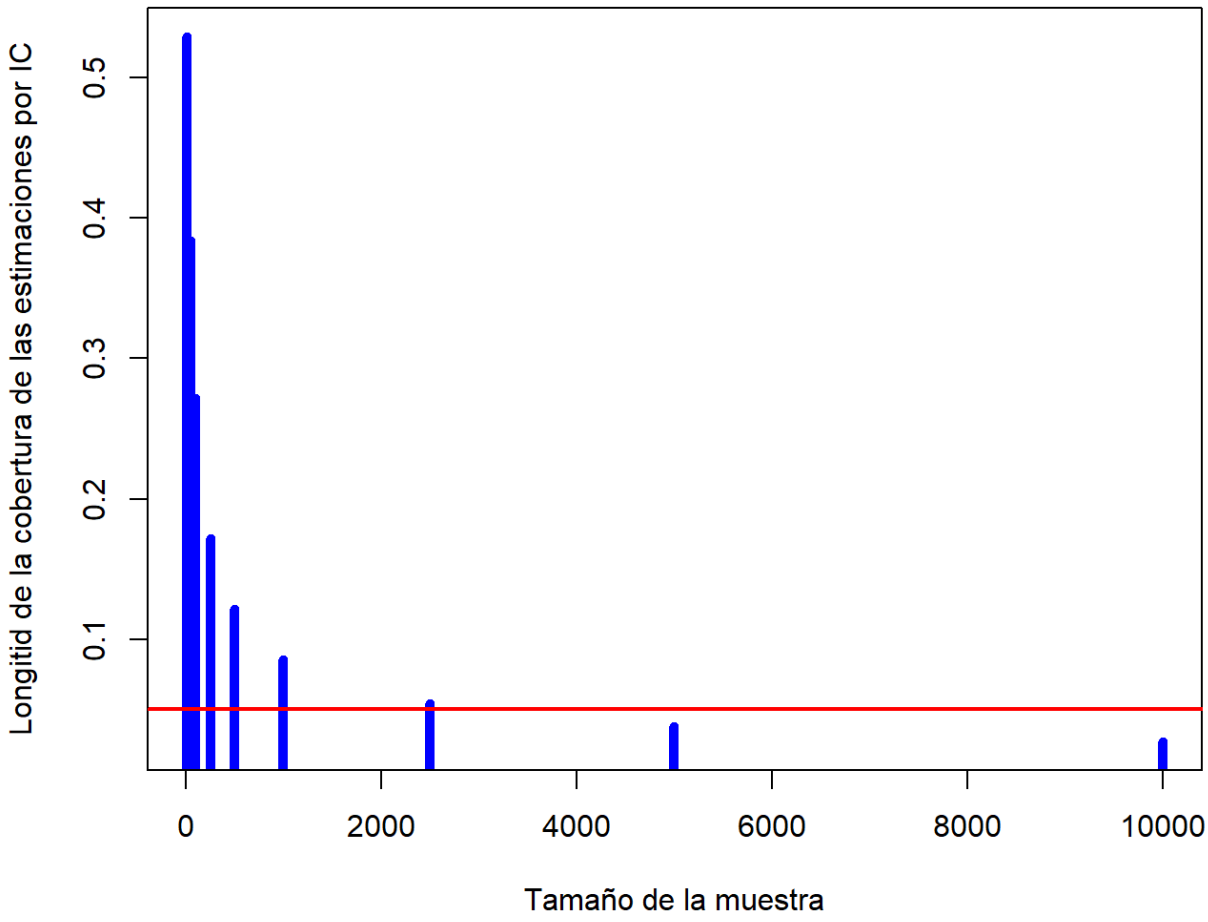


c) Grafique la longitud del intervalo contra n. Suponga que deseamos que la longitud del intervalo sea menor que 0.05. ¿Qué tan grande debe ser n?

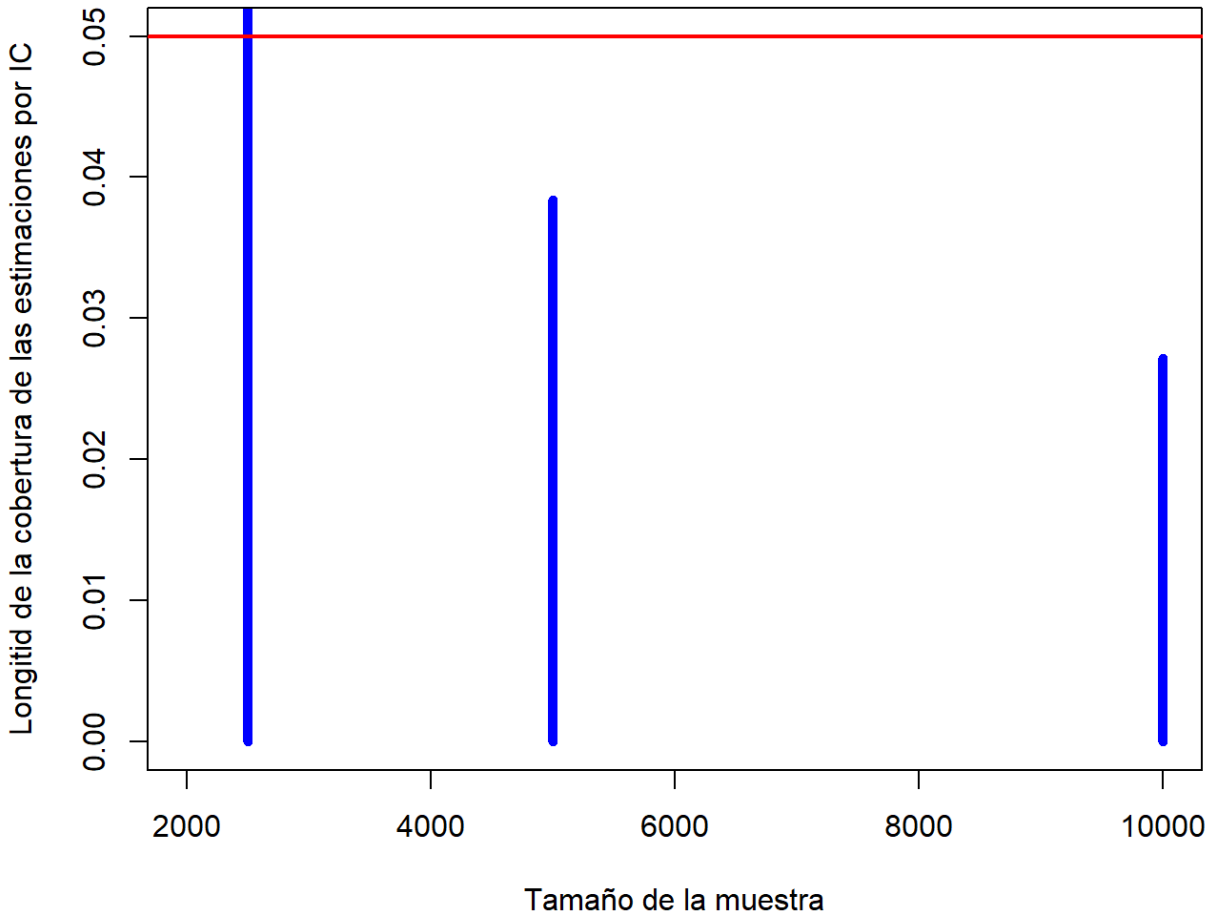
Code

Para observar que tan grande debe ser n se realiza por medio de un análisis visual. El siguiente plot se le delimitó el eje de las abcisas para observal el tamaño de muestra necesaria para que la longitud del intervalo sea menor a 0.05.

Code



Code



Code

```
## [1] "n"    "Dist"
```

Code

```
##           n      Dist
##  [1,]    10 0.52946941
##  [2,]    50 0.38412912
##  [3,]   100 0.27162030
##  [4,]   250 0.17178776
##  [5,]   500 0.12147229
##  [6,]  1000 0.08589388
##  [7,]  2500 0.05432406
##  [8,]  5000 0.03841291
##  [9,] 10000 0.02716203
```

Code

Como se observa en la tabla, un n mayor a los 2500 mil, genera una longitud de los intervalos de confianza de un 0.05. Pero siendo exactos, al 2951.104

$$2\sqrt{(\frac{1}{2n})\log(\frac{2}{.05})} < .05$$

$$\sqrt{(\frac{1}{2n})\log(\frac{2}{.05})} < .05$$

$$(\frac{1}{2n})\log(\frac{2}{.05}) < (.025)^2$$

$$\frac{\log(\frac{2}{.05})}{(2 * .025^2)} > n$$

EJERCICIO 5

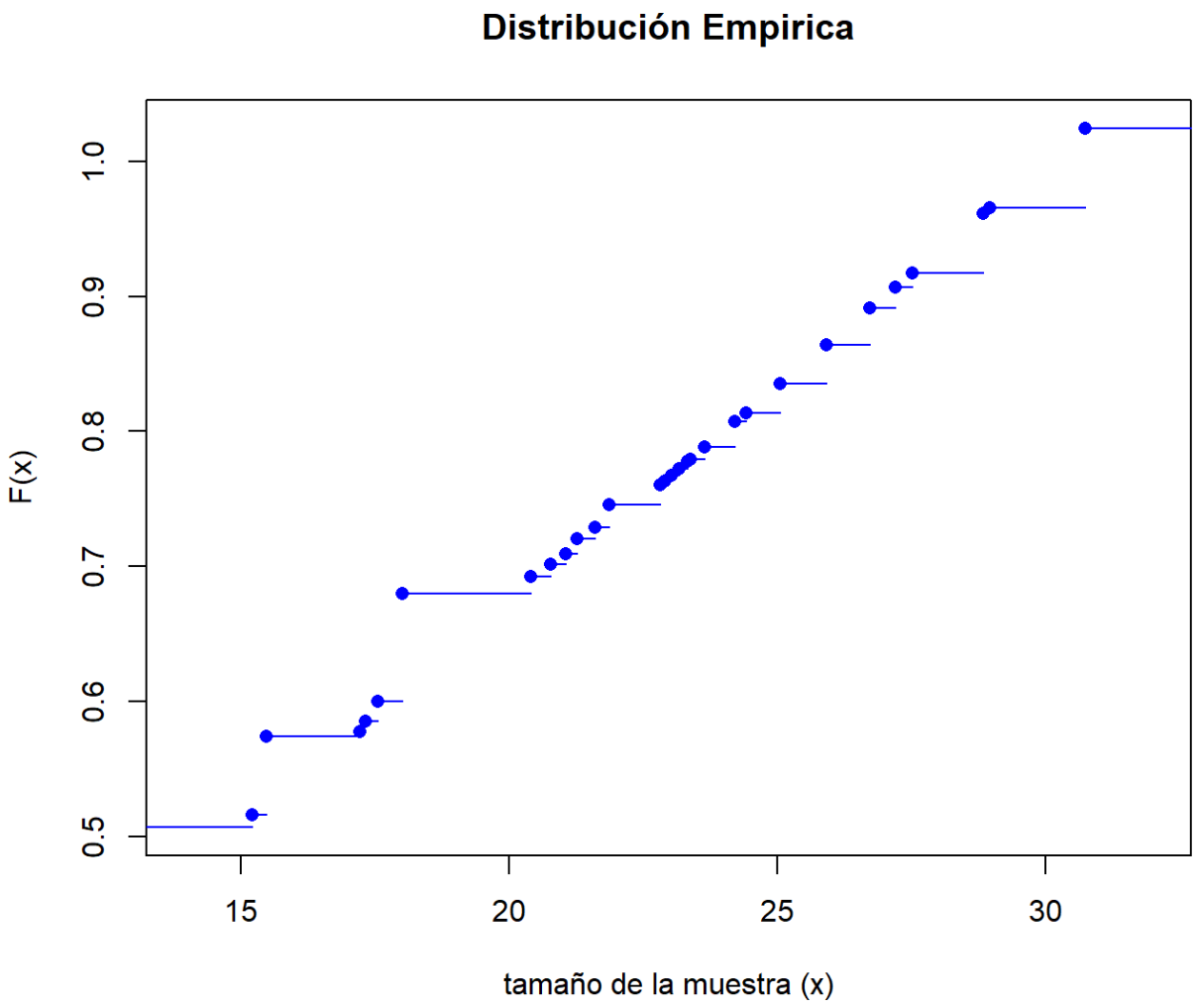
El siguiente conjuntos de datos contiene mediciones del diametro de un agave, medido en decmetros, en distintas localizaciones no cercanas.

a. Escriba una funci□on en R que calcule la función de distribución empirica para un conjunto #de datos dado. La funci□on debe tomar como parámetros al punto x donde se evalúa y alconjunto de datos D. Utilizando esta función grafique la función de distribuci□on emp□□rica asociada al conjunto de datos de lluvias. Ponga atenci□on a los puntos de discontinuidad. ¿Qu□e observa?

Code

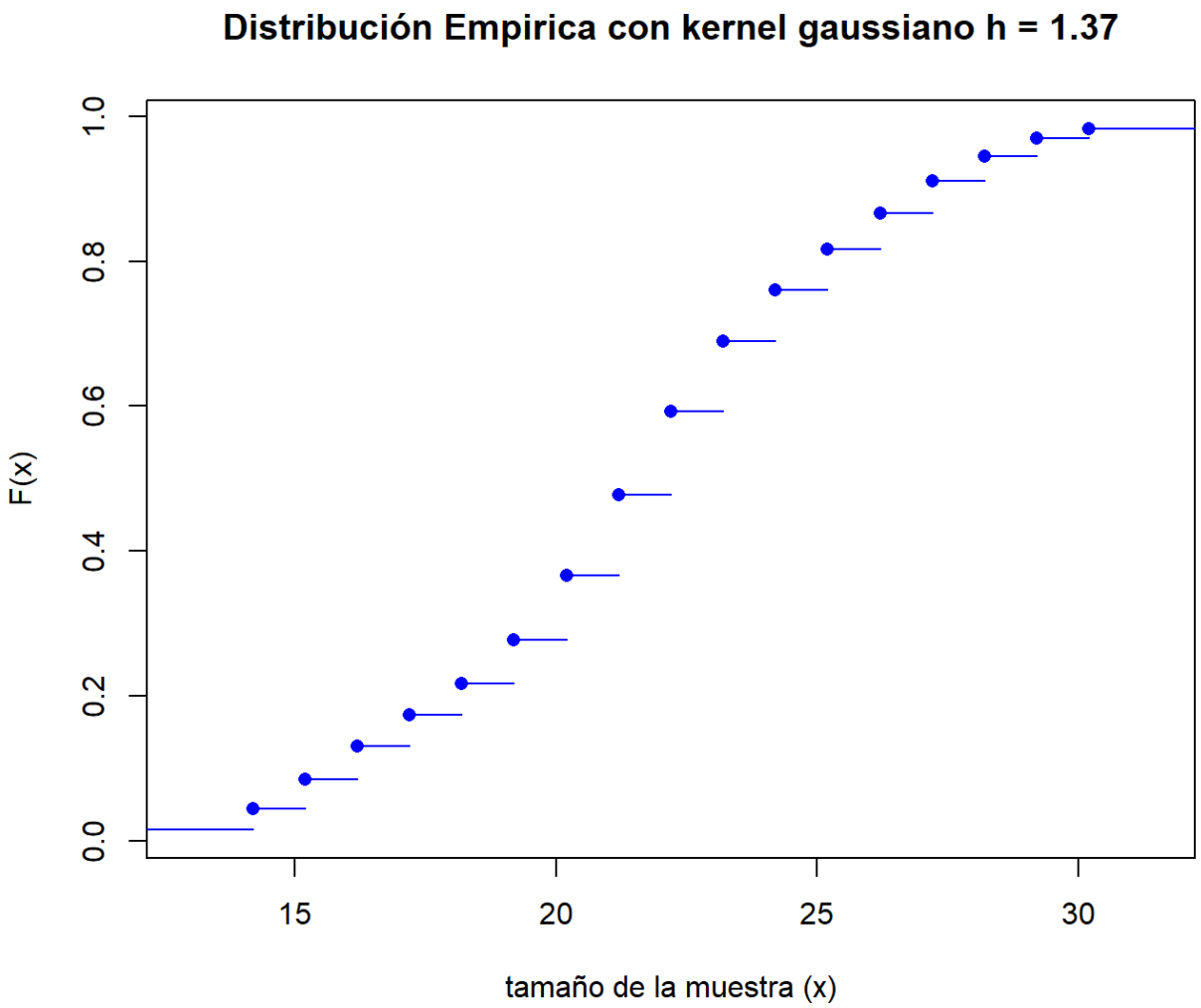
Se estima la función de distribución por dos métodos, esto solo para tener puntos de comparación en las estimaciones. La primera estimación de la distrribucion empirica se realiza al dividir las observaciones ordenadas.

Code



Por el hecho de contrastar estimaciones en la distribución emirica, se realiza otra estimación, pero ahora por el método de kernels. De acuerdo con Gramacki, A. (2018), la densidad gaussiana genera estimaciones más suavizadas que otros tipos de kernels, es por eso, que se utiliza un kernel gaussiano.

Code

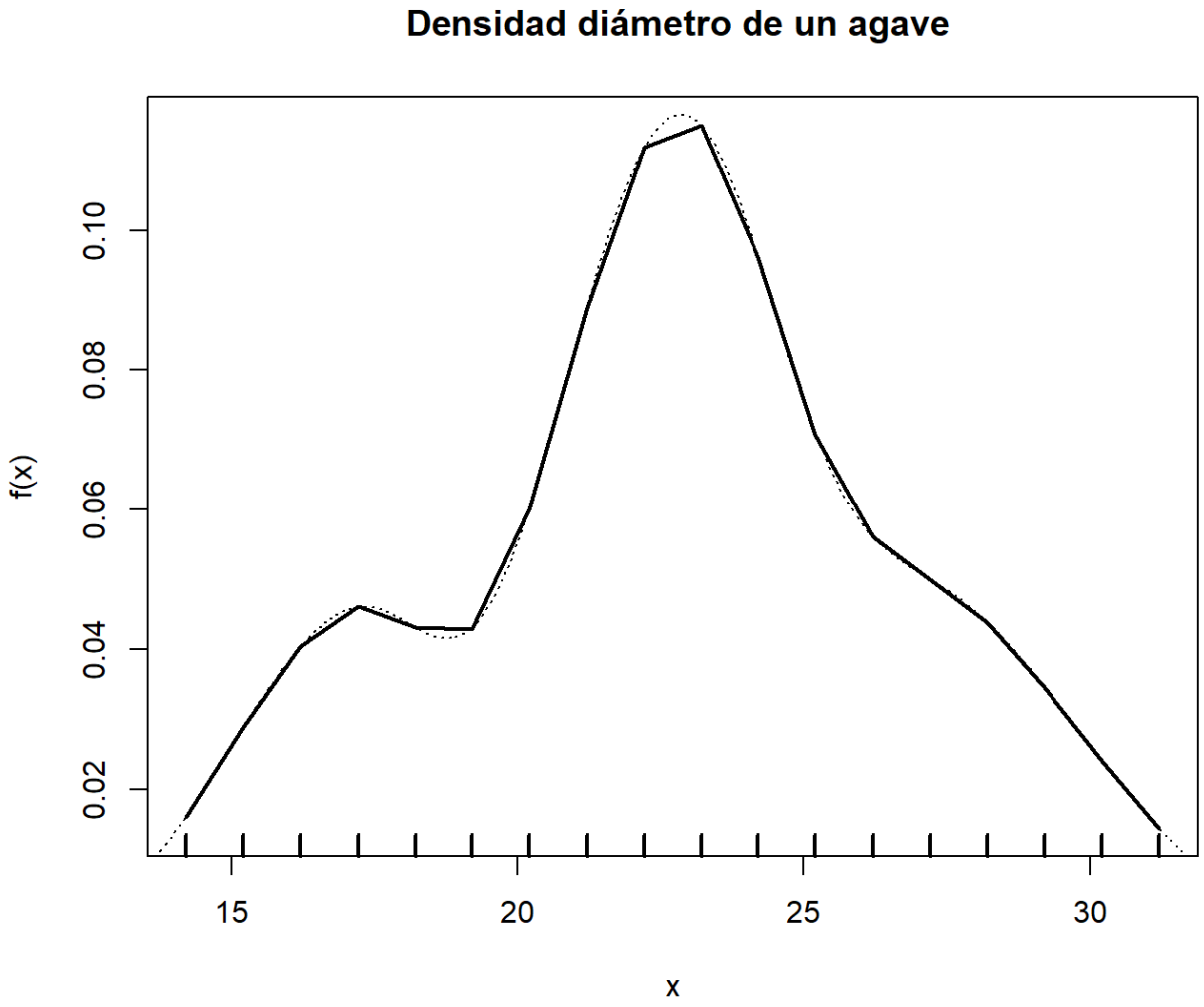


¿Qué observa?

Se observan incremento mayores al centro de las observaciones. A su vez, se observa que es más probale que el agave se encuentre entre 20 y 25 decímetros. Especialmente en el centro de la distribución.

Para observar lo que se menciona anteriormente, se gráfica la función de densidad del diámetro del agave.

Code



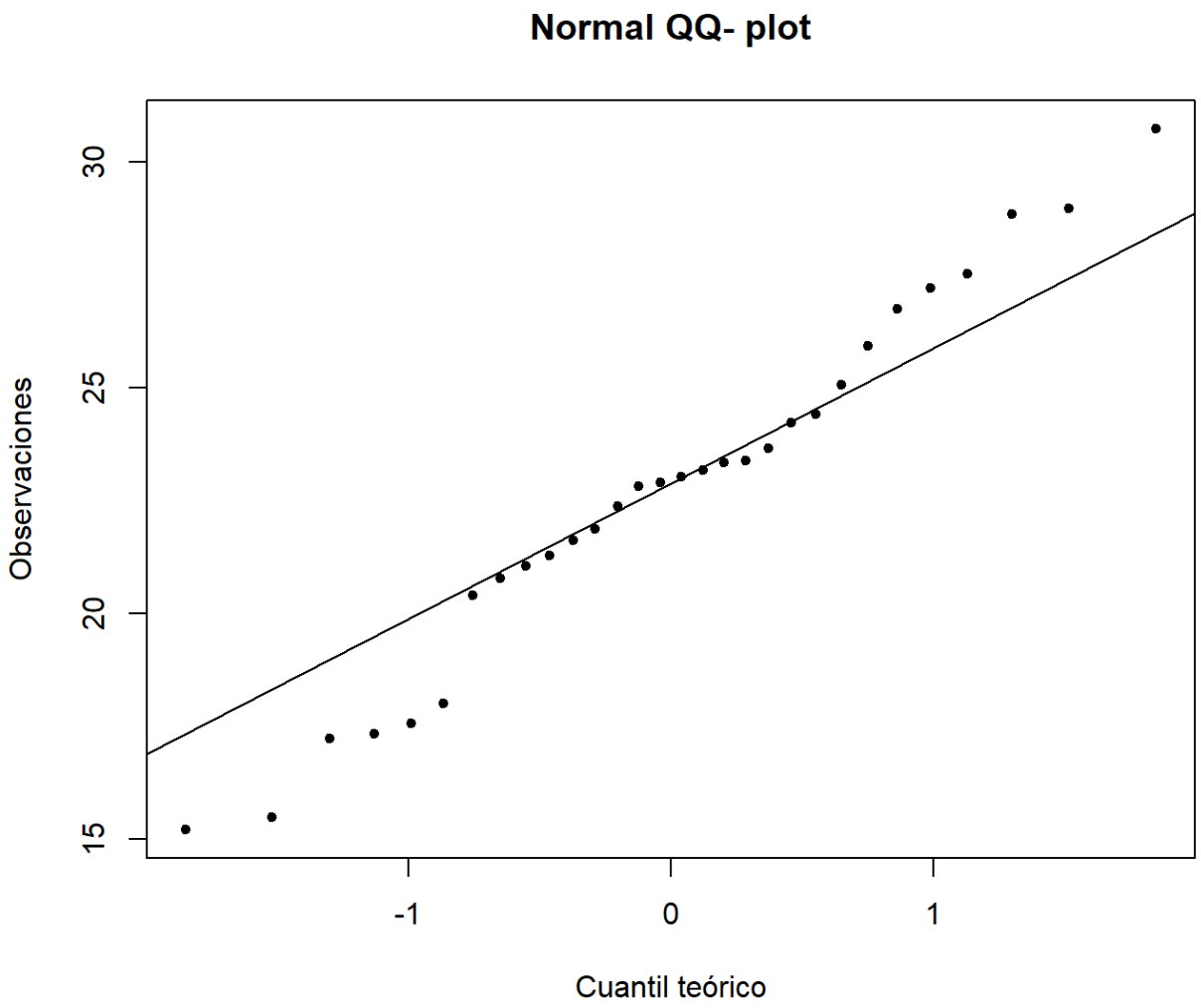
NOTA: la linea con dash es la que calcula R con density. Se traslapa dicha linea para observar como nuestra función se aproxima a density.

Como se mencionó anteriormente, la media de la distribución, como era de eseparese, se encuentra entre los 20 a 25 decimentros. Sin embargo, si se observa la cola derecha de la función de densidad, se observa con mayor claridad los brincos que aparecen en la función de distribución, los cuales, son más marcados en al inicio de la curva gráficada.

En conclusión, se podría comentar que en distintas localizaciónes los factores climaticos pueden variar influyendo en el proceso de crecimiento del agave. El proceso de crecimimento, por otro lado, puede ser factor de dichos saltos -en la función de distribución- a medida se produzcan determinada cantidad por estación.

b. Escriba una funci□on en R que determine la gr□aca Q-Q normal de un conjunto de datos. La funci□on debe tomar como par□ametro al conjunto de datos. Usando esta funci□on, determine la gr□aca Q-Q normal.

Code

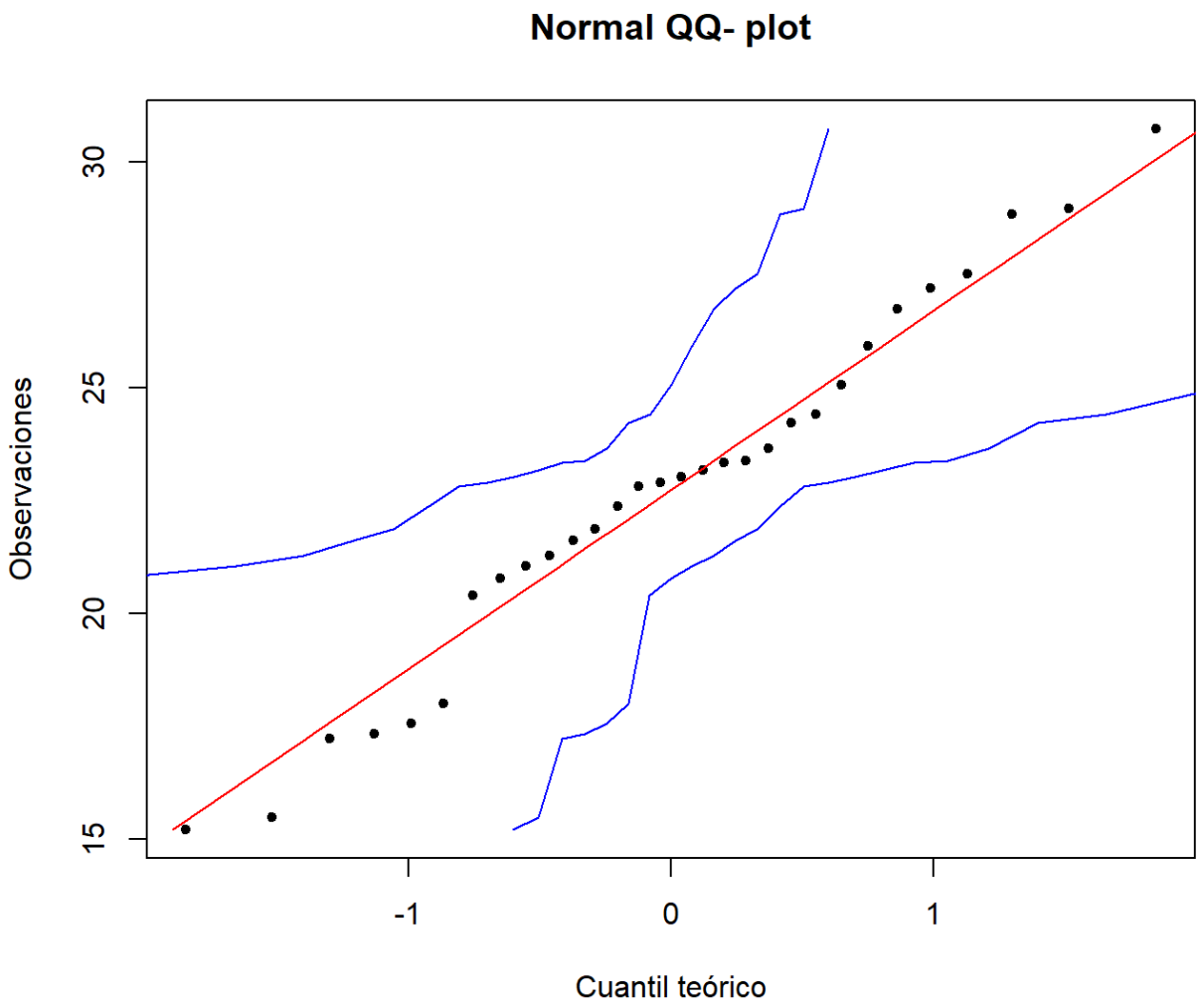


¿Qué observa?

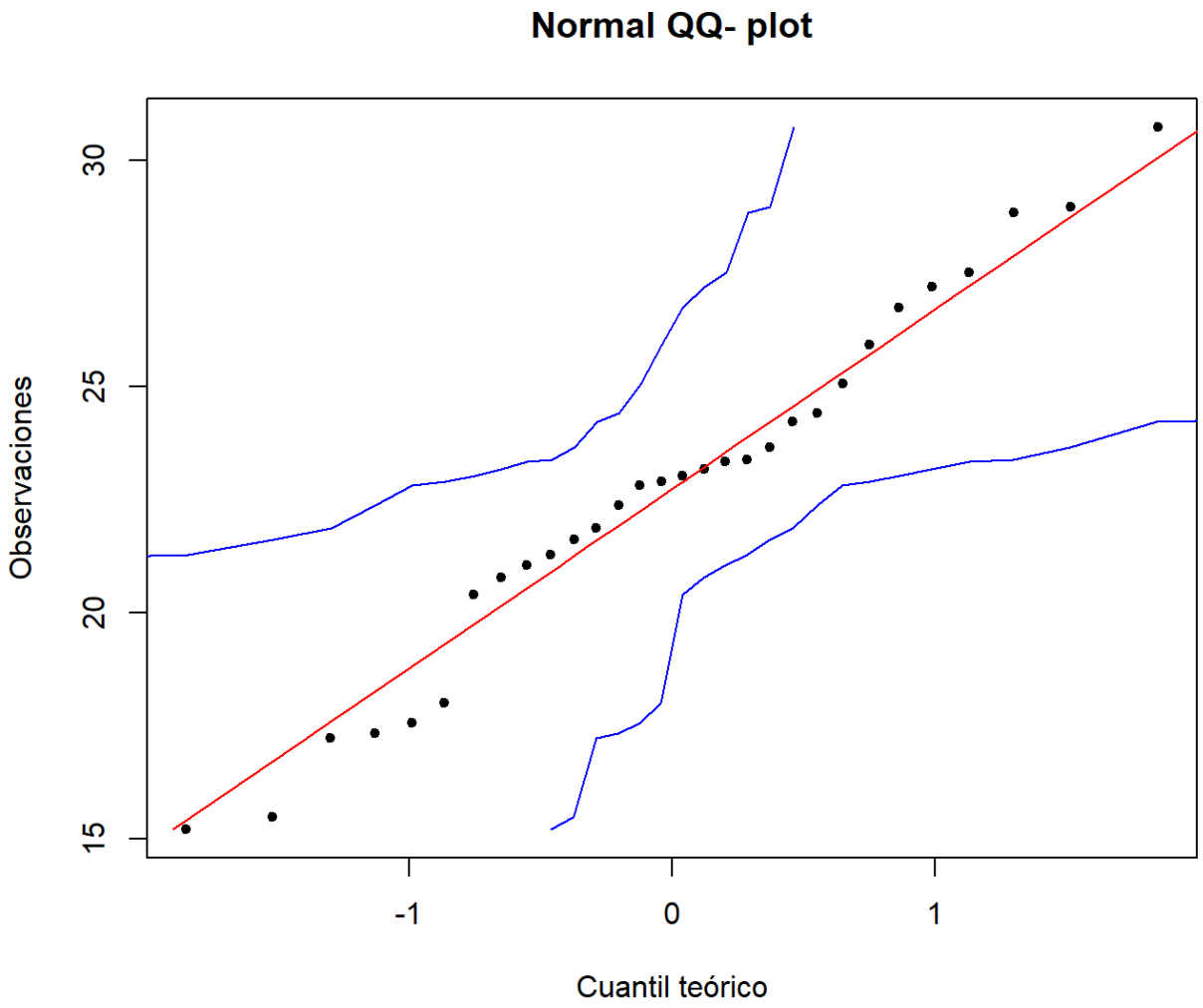
Se observa que las colas tanto superior como inferior son muy pesadas. A primera impresión la distrubución de los datos no es normal. Sin embargo, el centro de la distribución aproxima a una normal. La presencia de colas pesadas es el gran número de datos a tipicos, esto es, que hay una cantidad considerada de agavez que su altura -en decímetros- , es o muy pequeña o muy grande, claro está, que en su totalidad la mayoría se encuntra en promedio entre los 20 a 25 decímetros, aproximadamente.

c. Añada a la funci□on anterior la opci□on de que graque la banda de confíanza, de cobertura $1 - \alpha$, basada en el estadístico de Kolmogorov-Smirnov. La función debe tomar como parámetros al conjunto de datos y el nivel de confíanza $1 - \alpha$,. Aplique esta función al conjunto de datos para un nivel de conanza $1 - \alpha$□ = 0:95; 0:9 Qu□e observa?

Code


Normal QQ- plot
Observaciones
15
20
25
30
-1
0
1
Cuantil teórico

Code



NOTA: El valor d se calcula por medio de las tablas de kolmogorov-Smirknov. Las cuales se encuentran disponibles en : Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner. © 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

Para una $N = a30$, con un alpha de .05, el estadístico d es 0.24170 Para una $N = a30$, con un alpha de .01, el estadístico d es 0.28987

¿Qué observa?

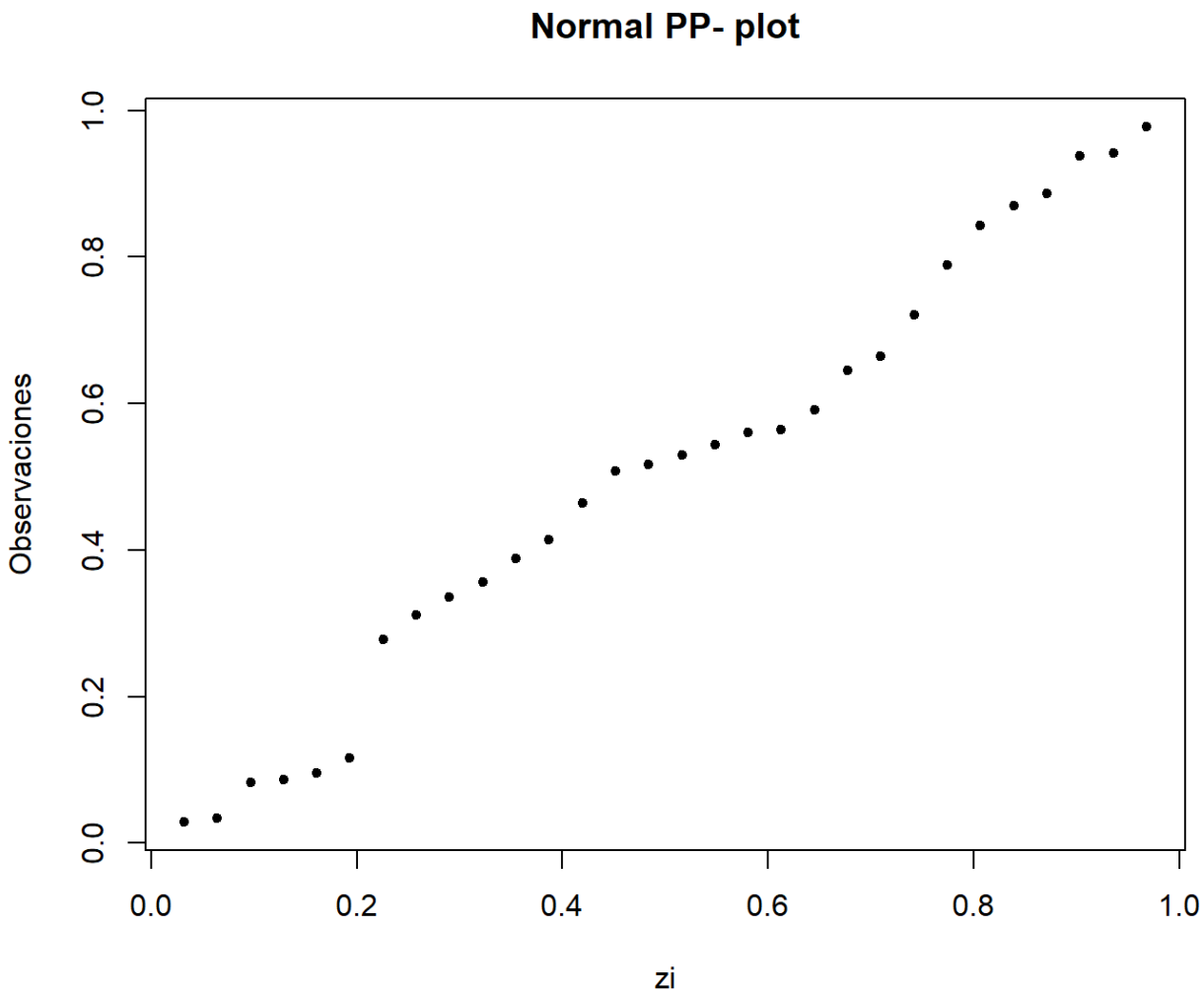
Ahora, se le anexa al QQ-plot un intervalo de confianza. Dicho intervalo, se contruye mediante el estadístico de Kolmogorov-Smirnov. En el primer gráfica se lacula el intervalo de confianza al 95%, en el cual se observa como todos los valores se encuentran dentro del intervalo de confianza.Sin embargo, la distribución del diametro del agave en decímetros se aproxima en las colas a una normal. Pero, el centro, las dos distribuciones si parecen que se aproximan.

Para el segundo gráfico, el intervalo de confianza es del 99%. En el cual, la mayoria de las observaciones caen dentro de dicho intervalo. No obstante, al tener niveles de insignificancia (α) tan pequeños, es decir la probabilidad de que la distribución estimada

no se ajuste bien a la teórica, será muy baja, dejando la posibilidad de que dicha distribución mientas sobre si es o no es la normal.

d. Escriba una función en R que determine el gráfico de probabilidad normal. La función debe tomar como parámetro al conjunto de datos. ¿Qué observa?

Code



¿Qué observa?

Se gráfica el pp-plot de los datos, en el cual, al igual que con el qq-plot, se observa que los datos presentan colas pesadas y algo de sesgo.

e. Los datos anteriores se distribuyen normalmente? Argumente.

En conclusión, la función de densidad que se estimó se aproxima a la normal sólo en el centro, sin embargo, tanto el QQ y PP plot nos hace ver que las colas son muy pesadas, inclusive presentando sesgo. Al introducir el intervalo de confianza, para cada punto se construye un intervalo de 95% de confianza, esto es que tienden a mentir más por ser intervalos puntuales, ya que se encuentran subestimando la variablidad de los datos. Lo mejor sería contrastar con unas bandas globales. A su vez, los datos a típicos puede que no se estén capturando por la distribución normal. De esta manera, se concluye que dado a

las estimaciones que se realizaron y a la información del qq- plot dado su intervalo, se puede decir que dichas bandas mienten mucho, que la normal no captura los datos a tipicos, y por el sesgo que presenta la distribución no se considerará que os datos se distribuyan normales.

EJERCICIO 6

a. Escriba una funci□on en R que calcule el estimador de la densidad por el m□etodo de kerneles. La funci□on deber□a recibir al punto x donde se eval□ua al estimador, al par□ametro de suavidad h, al kernel que se utilizar□a en la estimaci□on y al conjunto de datos.

Code

```
##    X1
## 1  25
## 2  40
## 3  83
## 4 123
## 5 256
## 6   1
```

Code

```
##        X1
##  Min.   :  1.0
##  1st Qu.: 32.0
##  Median : 79.0
##  Mean   :123.8
##  3rd Qu.:144.0
##  Max.   :737.0
```

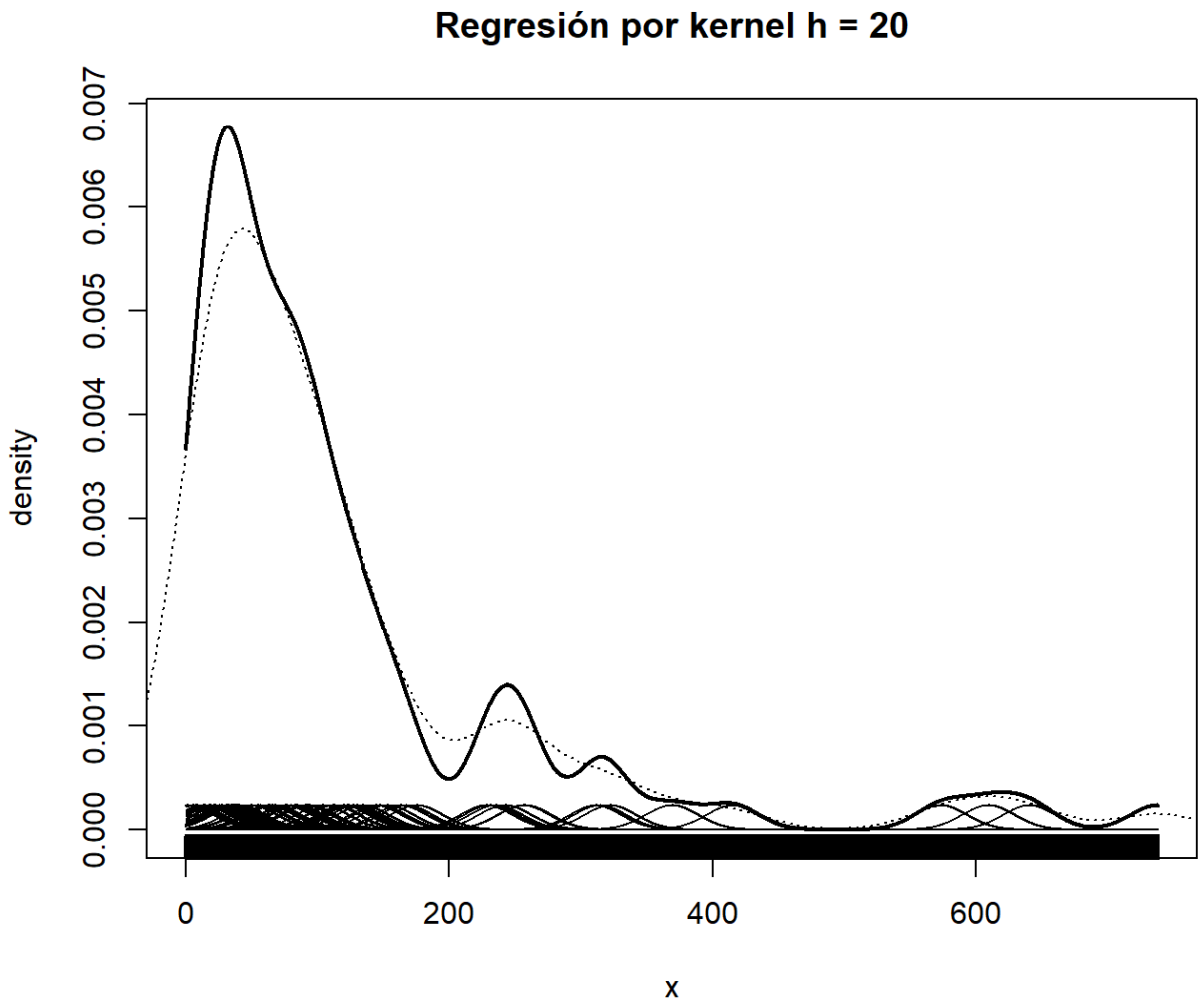
Code

b. Cargue en R al archivo .csv“, el cual contiene la duraci□on de los per□□odos de tratamiento (en días) de los pacientes de control en un estudio de suicidio. Utilice la funci□on del inciso anterior para estimar la densidad del conjunto de datos para $h = 20, 30, 60$. Graque las densidades estimadas. >Cu□al es el mejor valor para h? Argumente.

Code

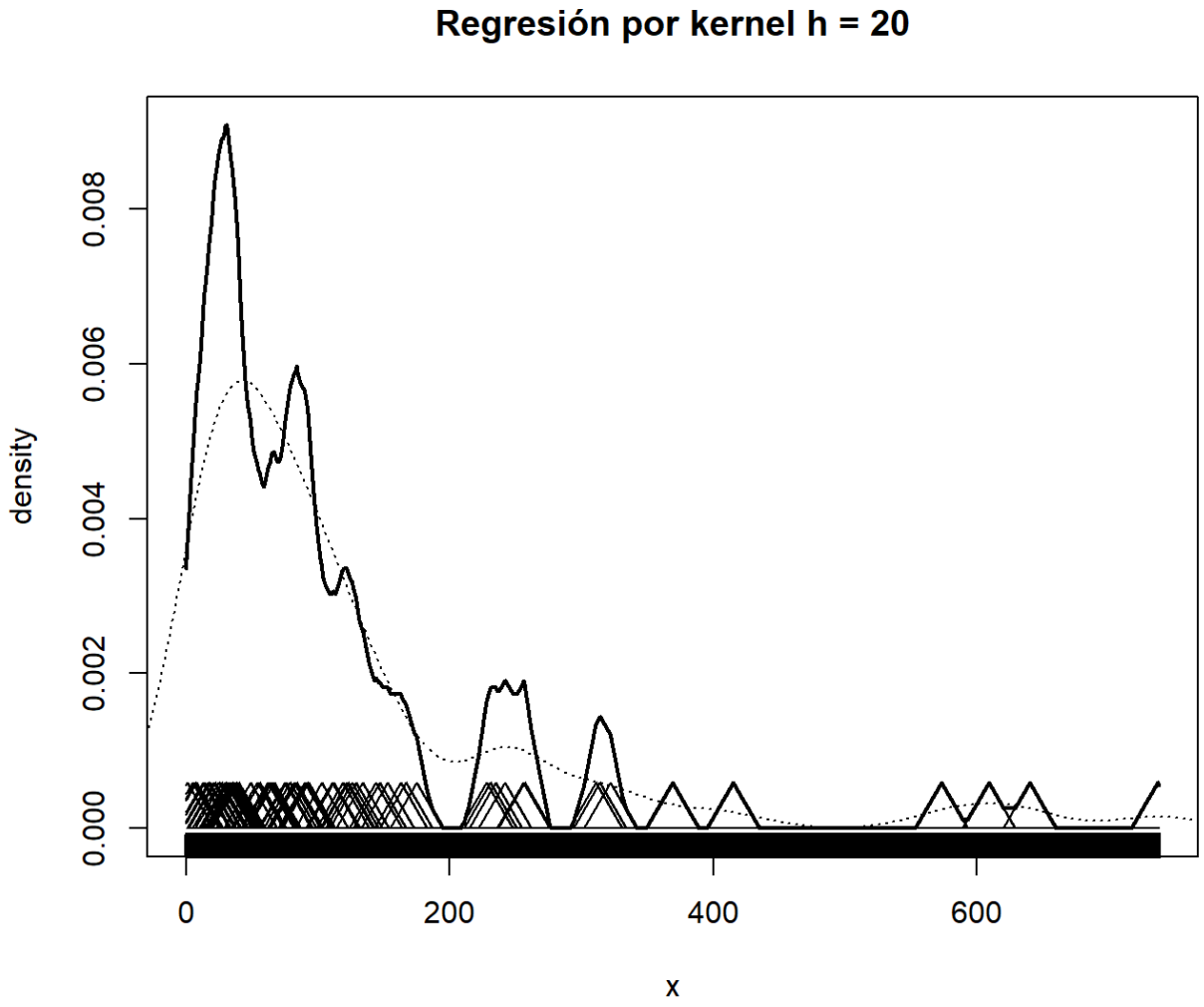
Code

```
## [1] "Seleccionó Un kernel Normal para la Estimación"
```



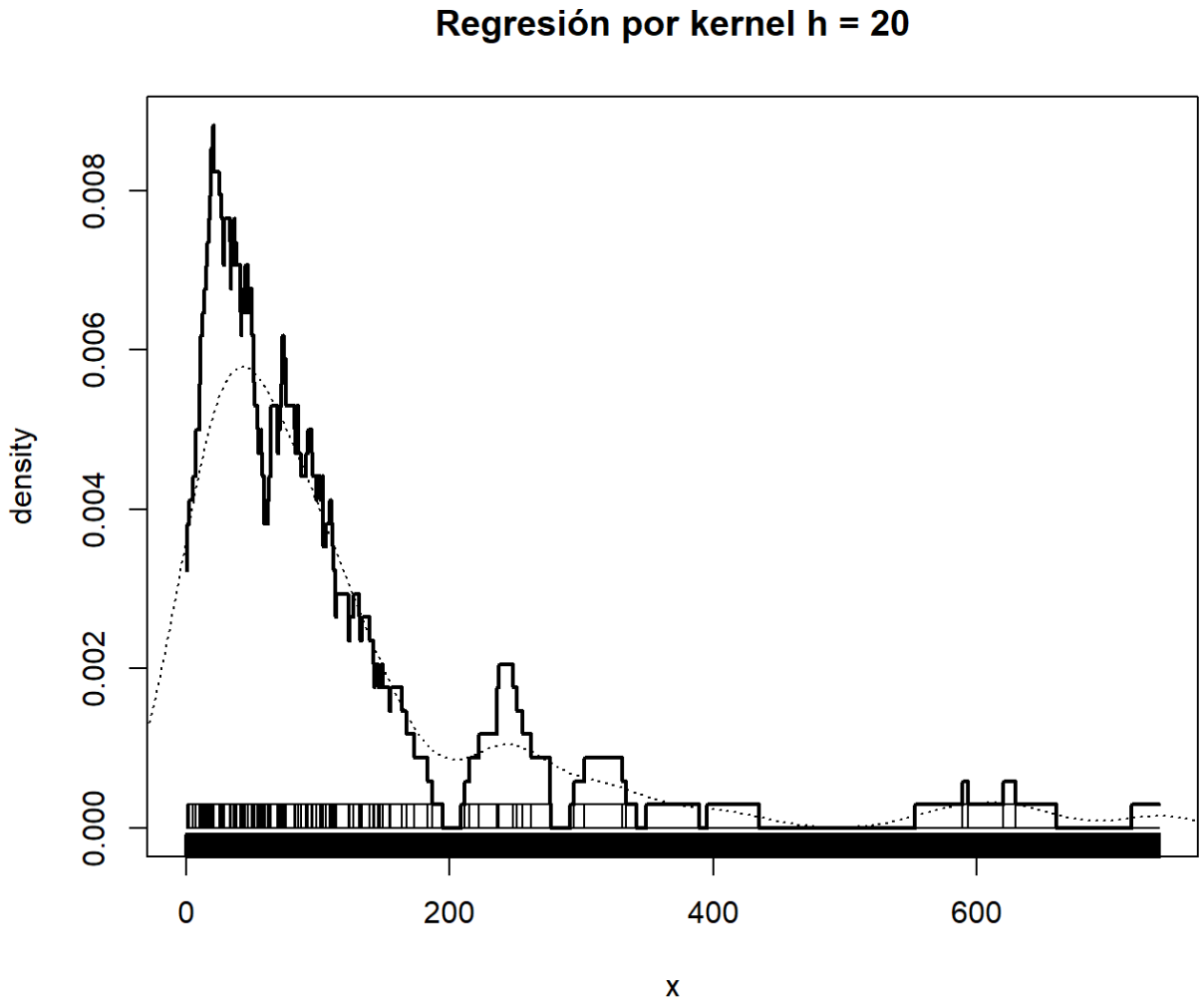
Code

```
## [1] "Seleccionó Un kernel Triangular para la Estimación"
```



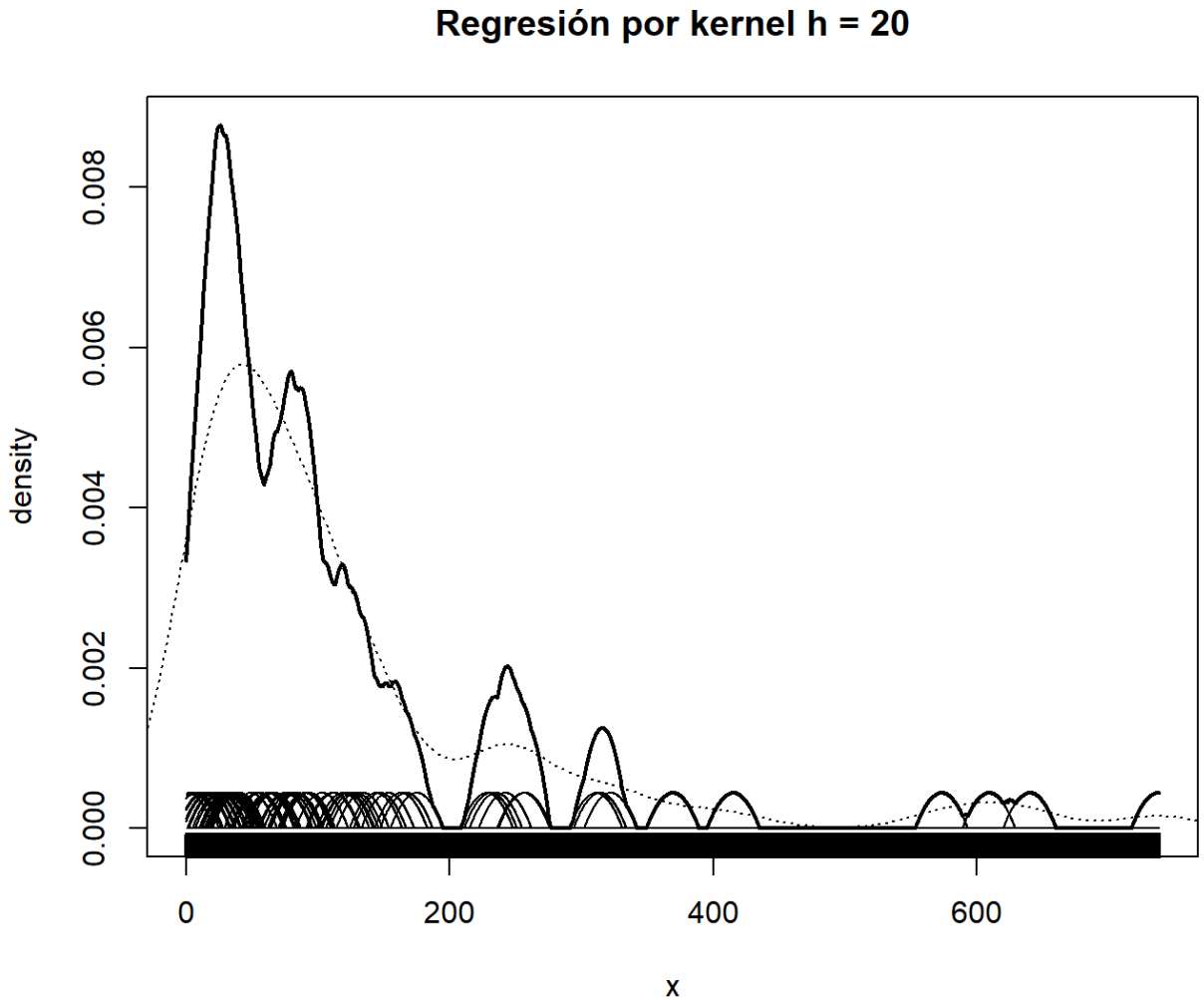
Code

```
## [1] "Seleccionó Un kernel Uniforme para la Estimación"
```



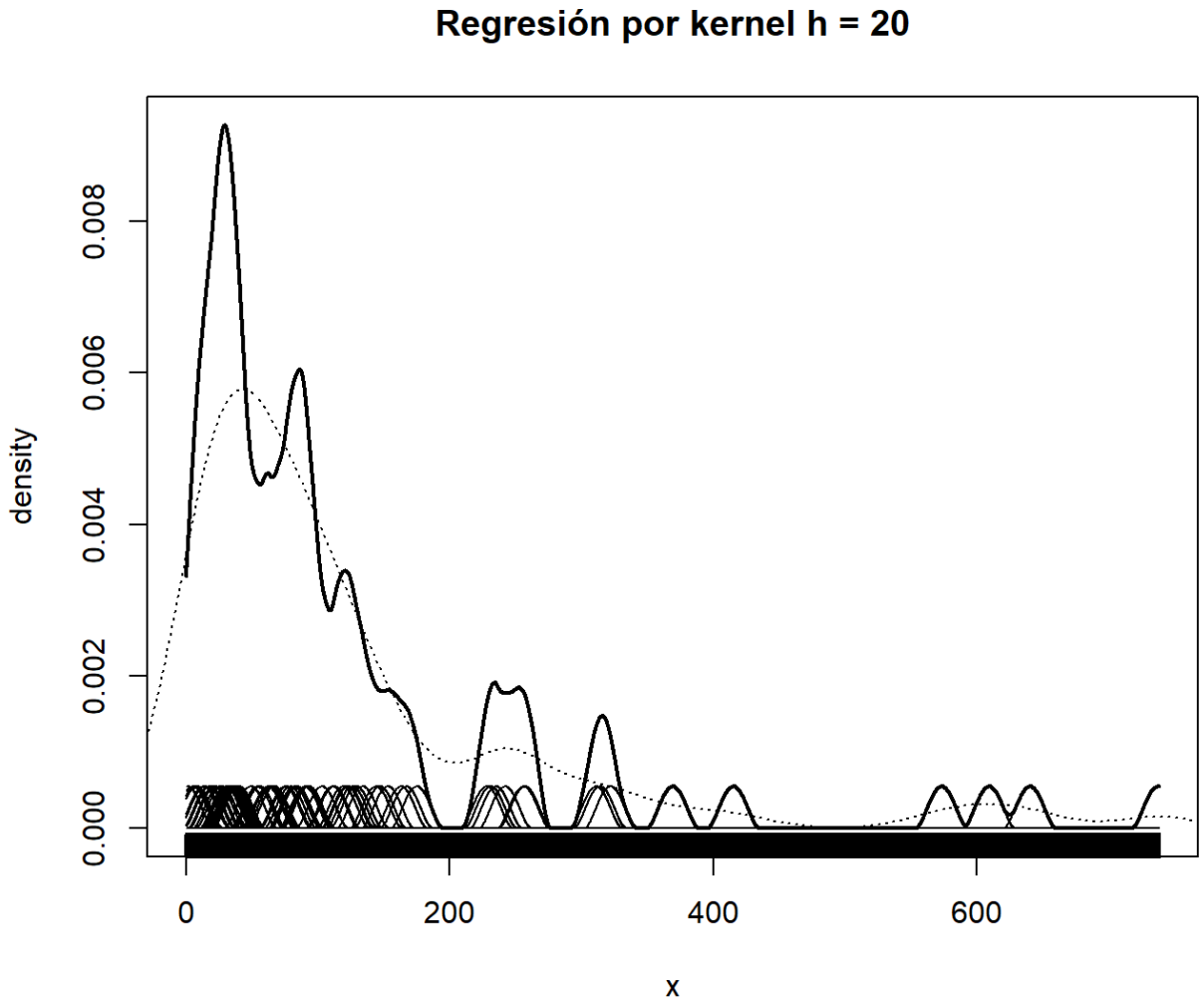
Code

```
## [1] "Seleccionó Un kernel Epanechnikov para la Estimación"
```



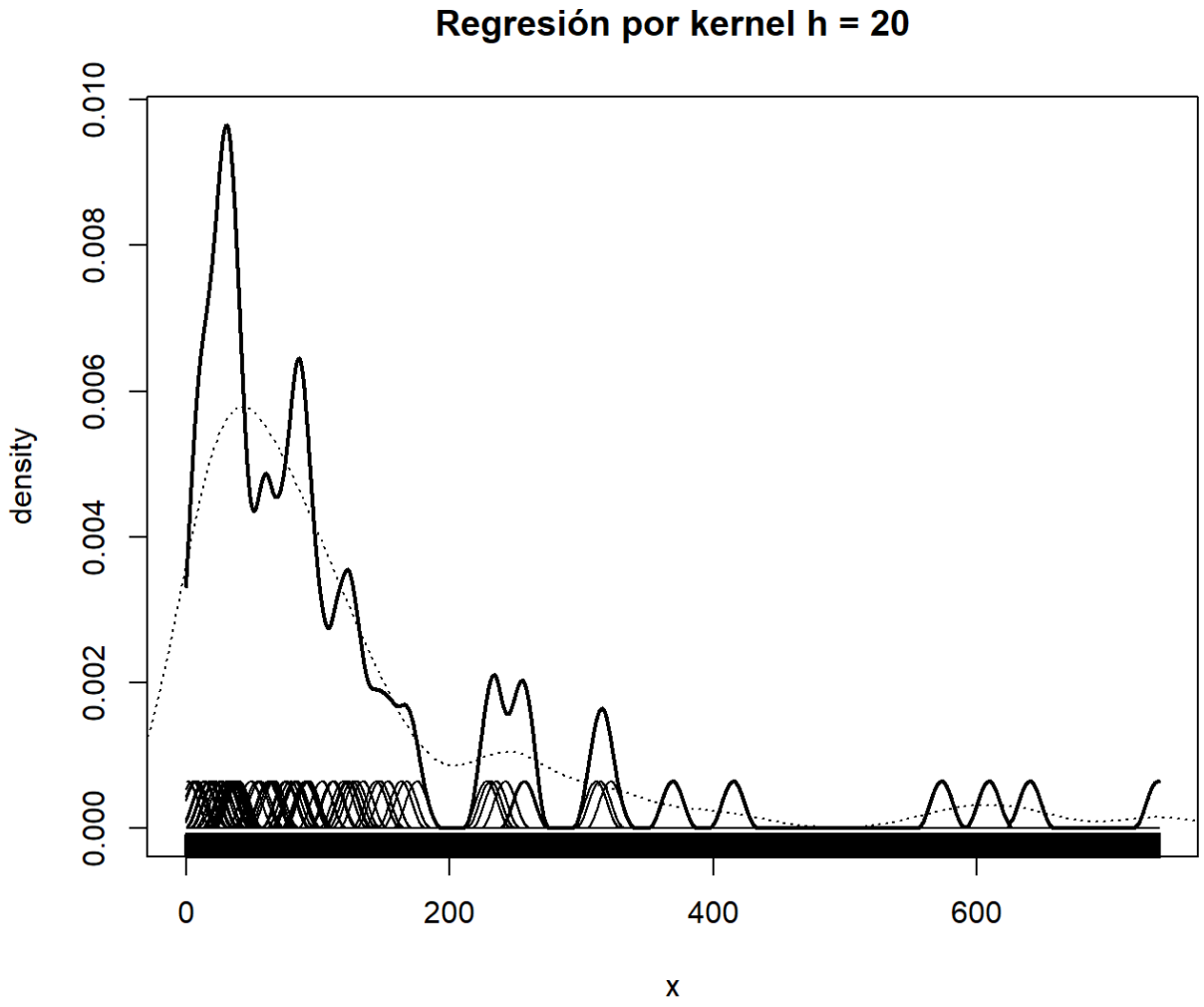
Code

```
## [1] "Seleccionó Un kernel Biweight para la Estimación"
```



Code

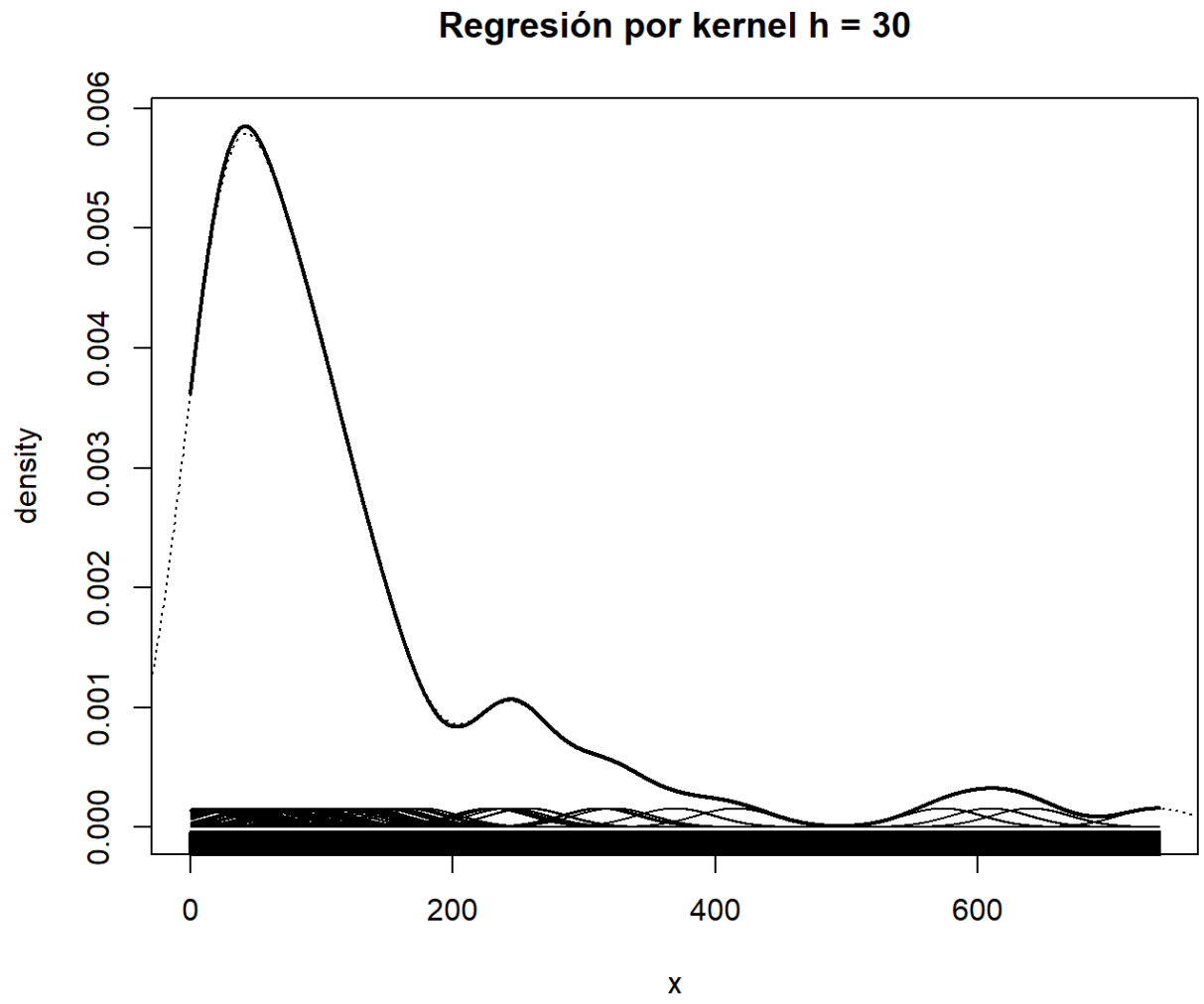
```
## [1] "Seleccionó Un kernel Triweight para la Estimación"
```



Antes del análisis, se vuelve a recordad que la densidad contra la que se traslapa la estimación kernel, es contra la función density, la cual calcula un h óptimo. Esto con el fin de saber que la función que se programó se aproxima a una ya realizada en R. La comparación de prámetro de suavidad se centrará en los Kernel gaussianos. Sin embargo, se discute un poco sobre la estimación con los distintos tipos de kernel para cada valor de h. A manera muy general, entre los kernels con los que se puede estimar la densidad, se observa, que el gaussiano con un suavizador de 20 se aproximá más que los otros posibles tipos de kernel. No obstante, con un h igual a 20 se puede observar que con ese valor aún se observa un poco de variabilidad. mostrando algunos intervalos con algunos picos.

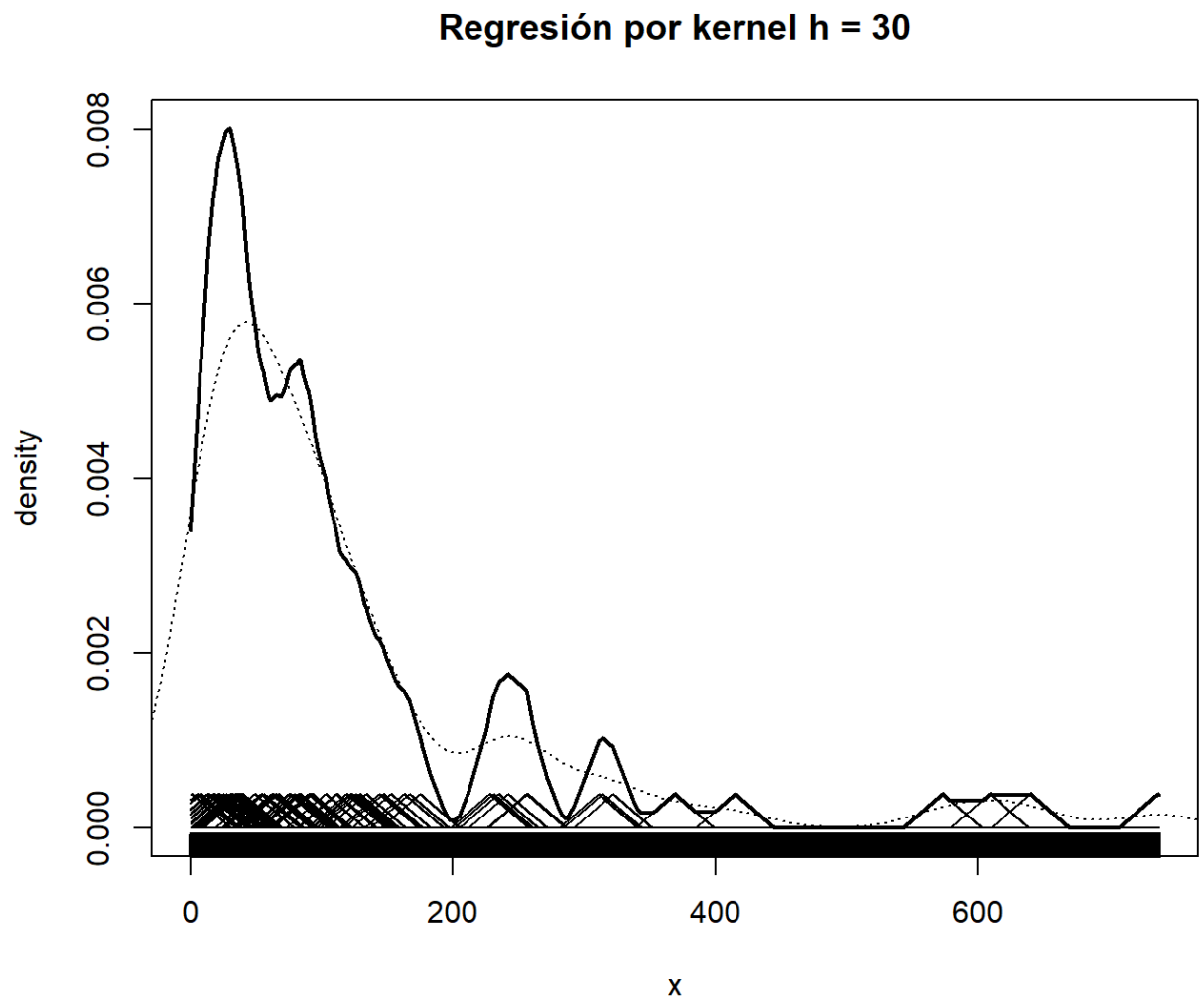
Code

```
## [1] "Seleccionó Un kernel Normal para la Estimación"
```



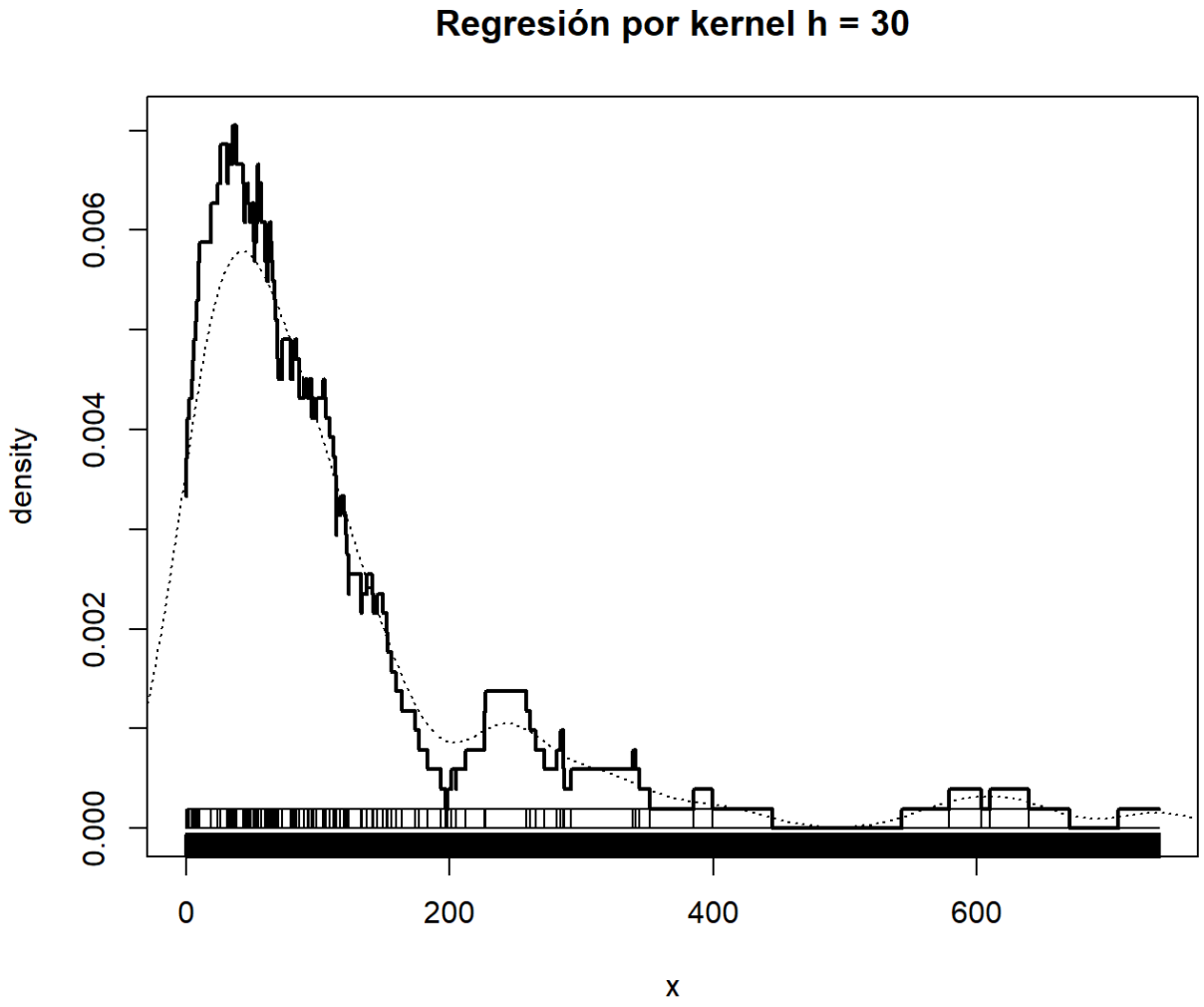
Code

```
## [1] "Seleccionó Un kernel Triangular para la Estimación"
```



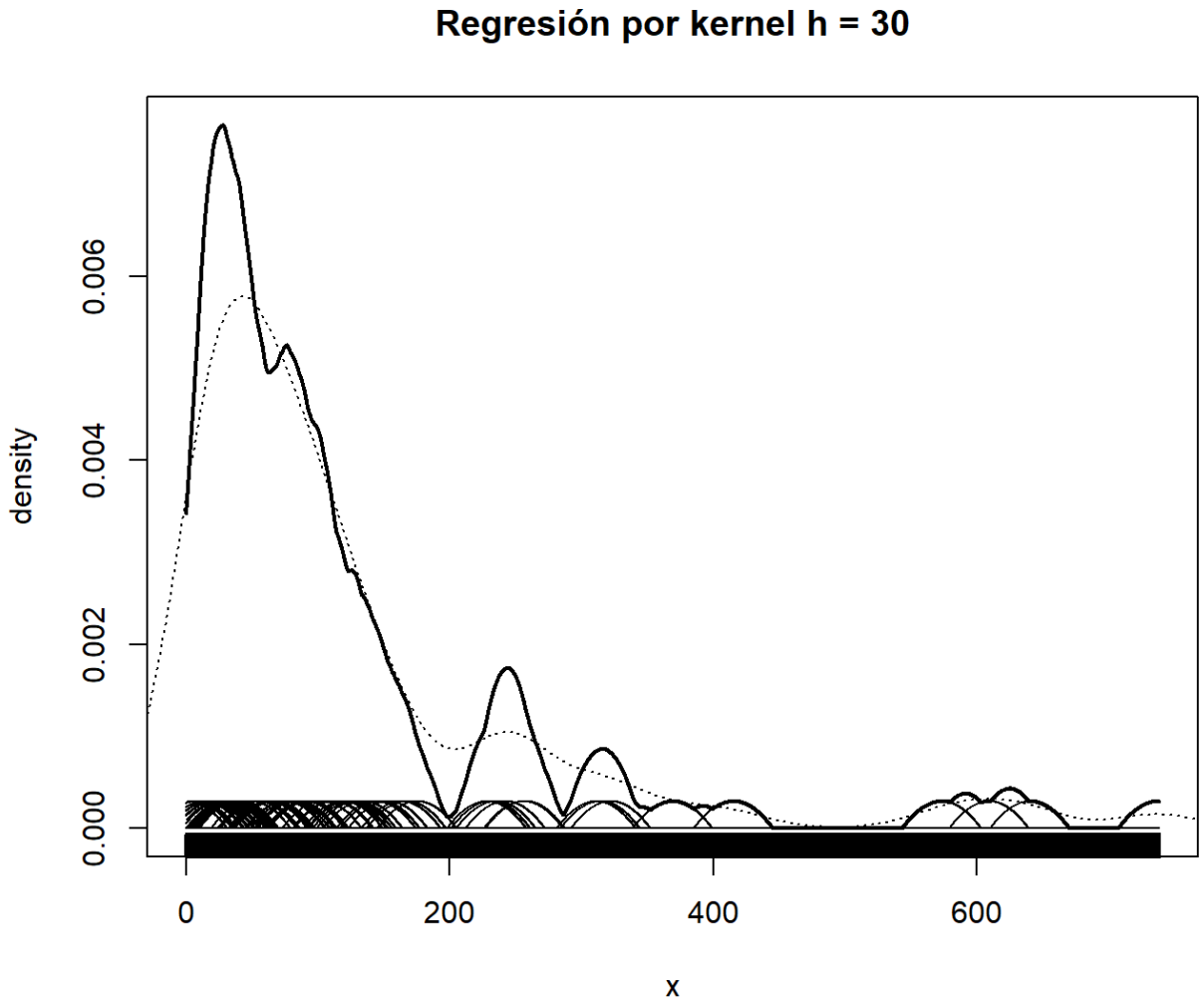
Code

```
## [1] "Seleccionó Un kernel Uniforme para la Estimación"
```



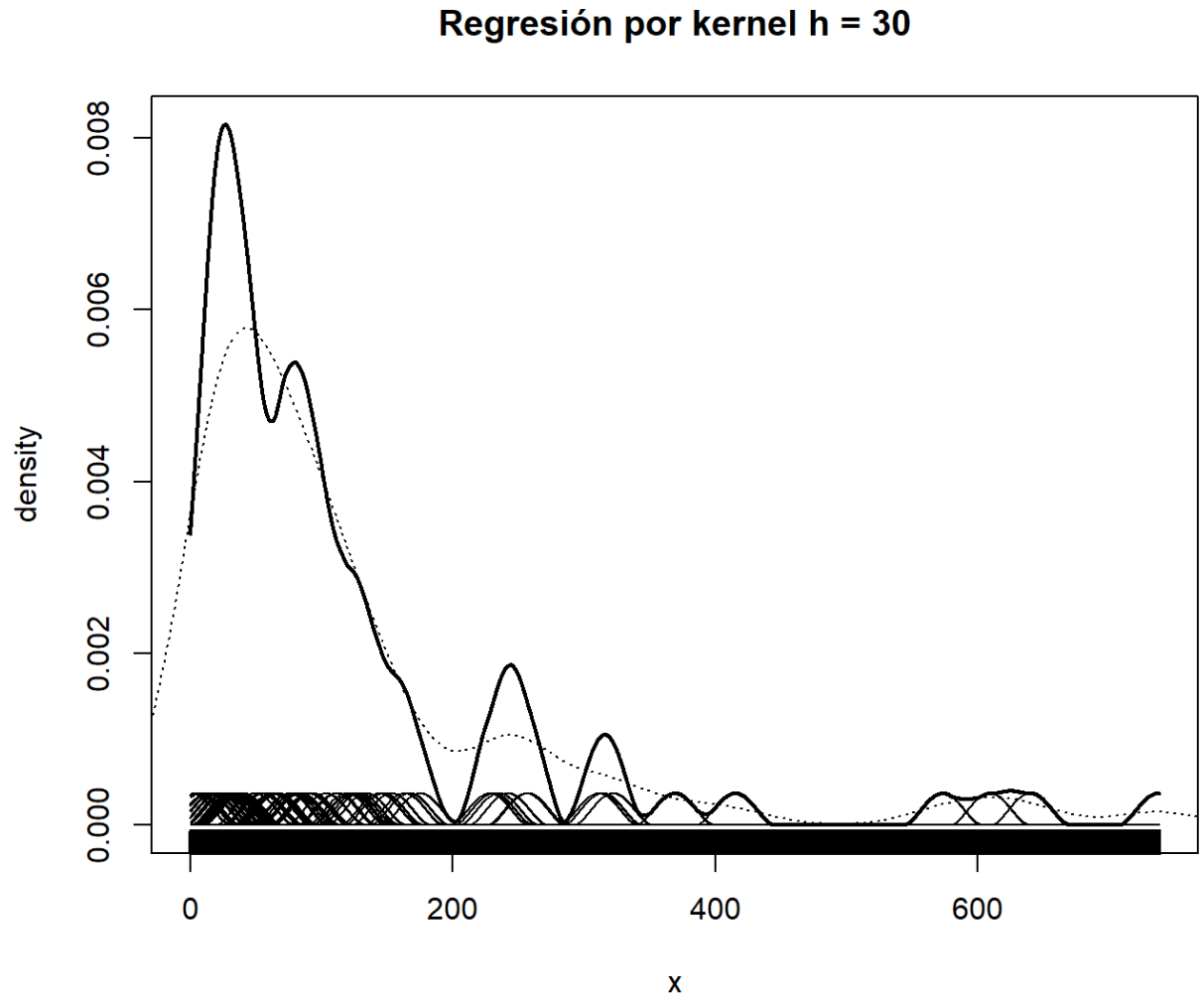
Code

```
## [1] "Seleccionó Un kernel Epanechnikov para la Estimación"
```



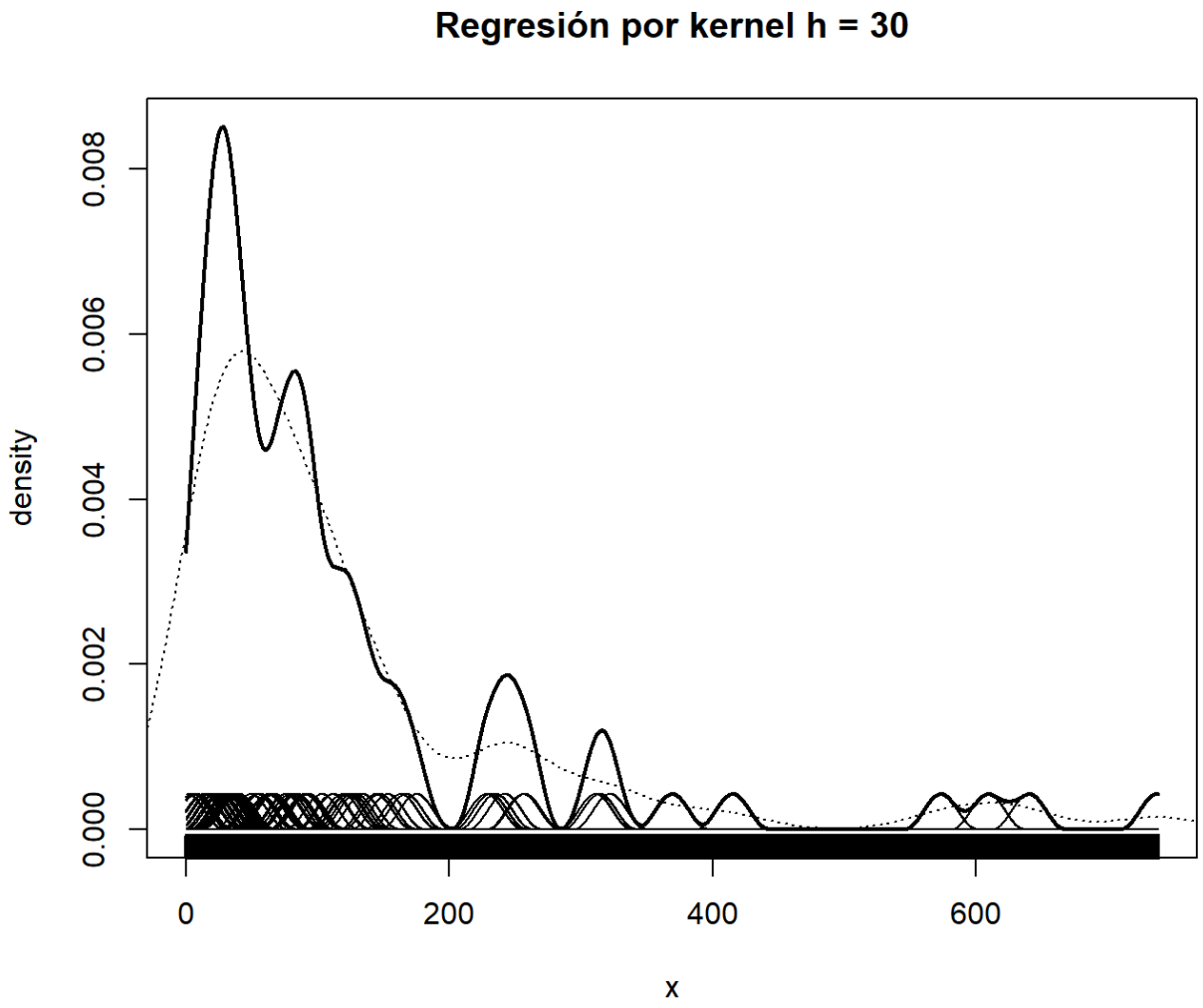
Code

```
## [1] "Seleccionó Un kernel Biweight para la Estimación"
```



Code

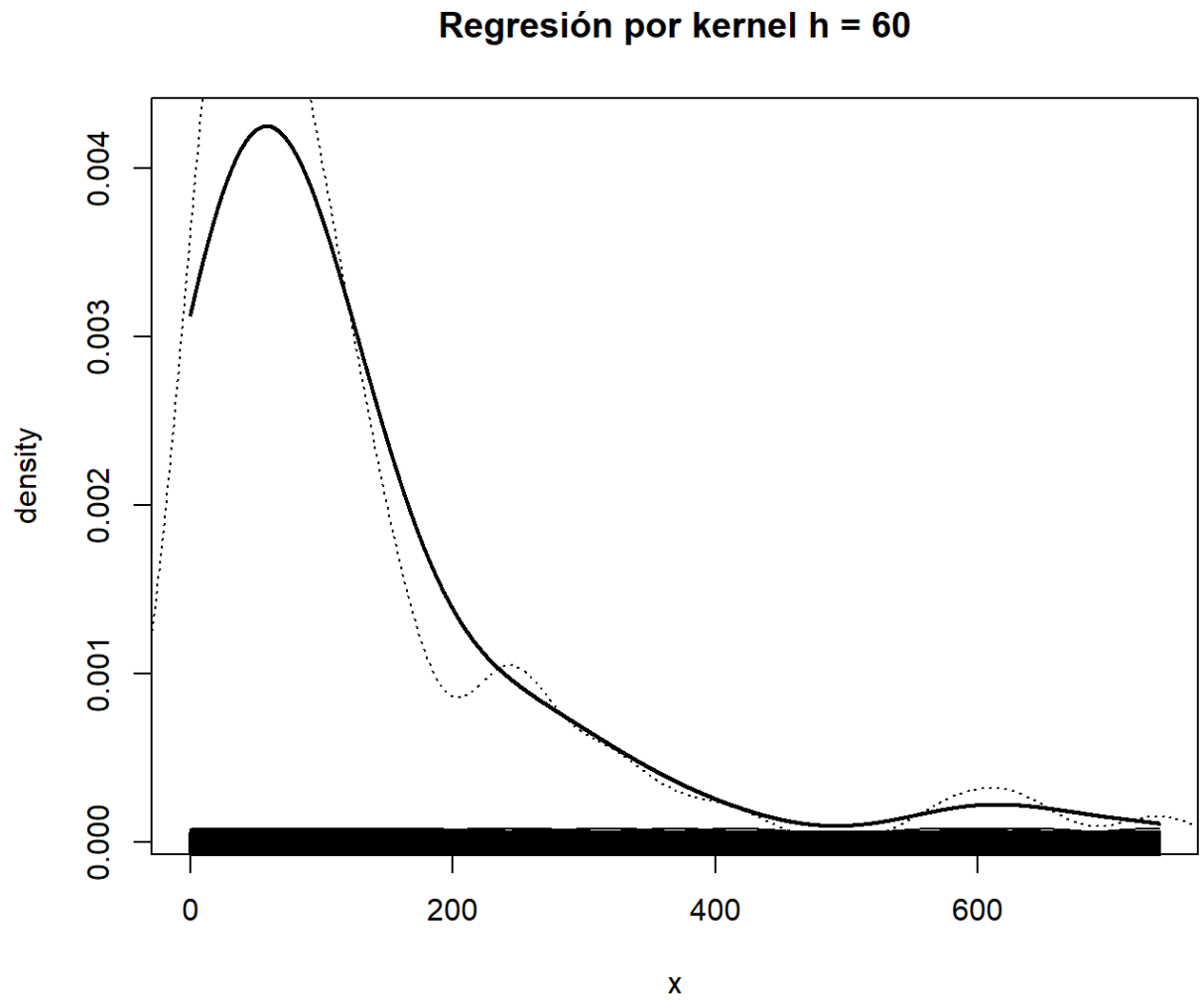
```
## [1] "Seleccionó Un kernel Triweight para la Estimación"
```



Con un h igual a 30, la dessidad estimada con un kernel gaussiano parece lo suficientemente suavizada. A lo que respectan los otros tipos de kernel, no onstante, la variabilidad es muy marcada a ese valor del parámetro, haría falta un h más grande para que tendieran a la gaussiana.

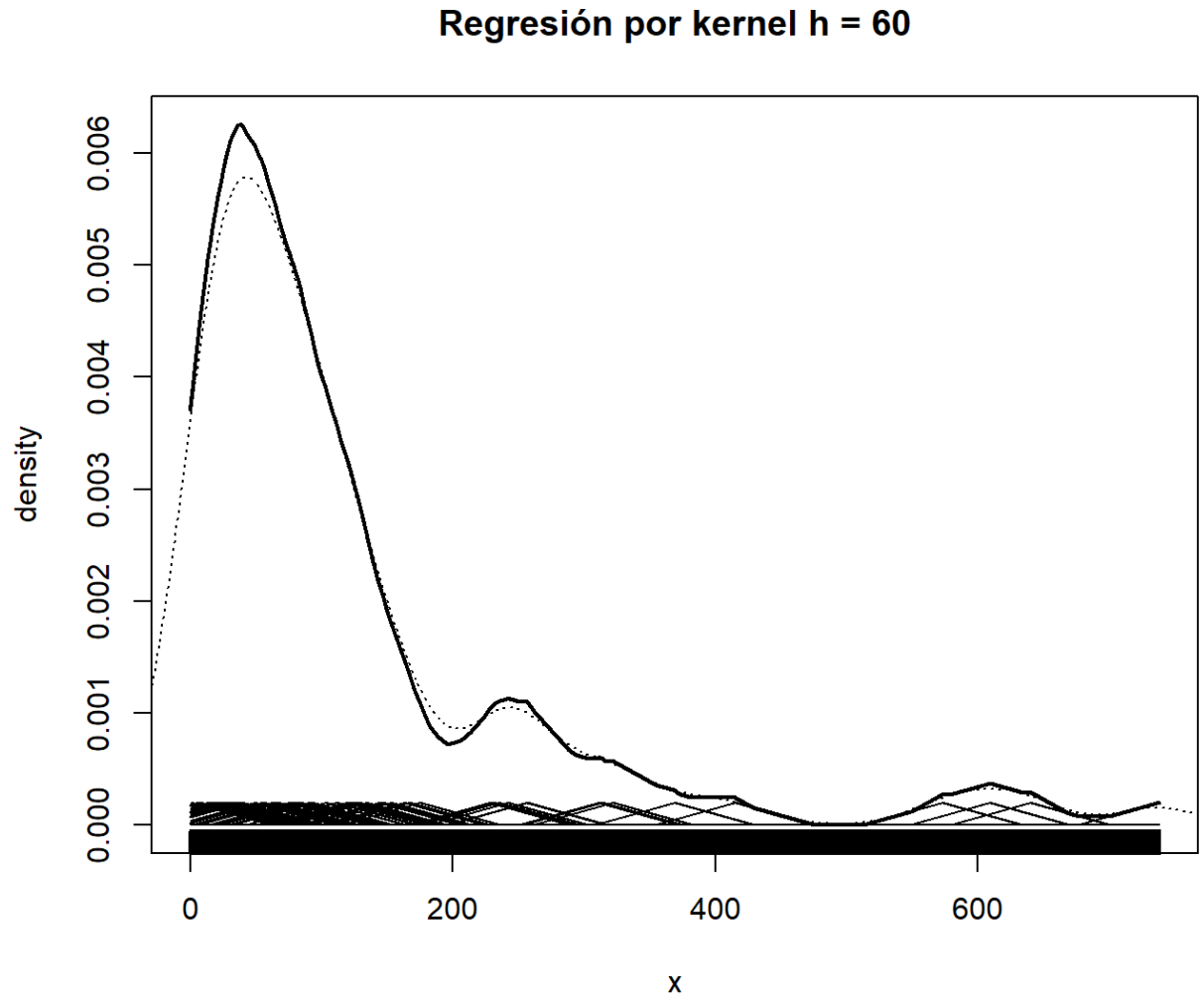
Code

```
## [1] "Seleccionó Un kernel Normal para la Estimación"
```



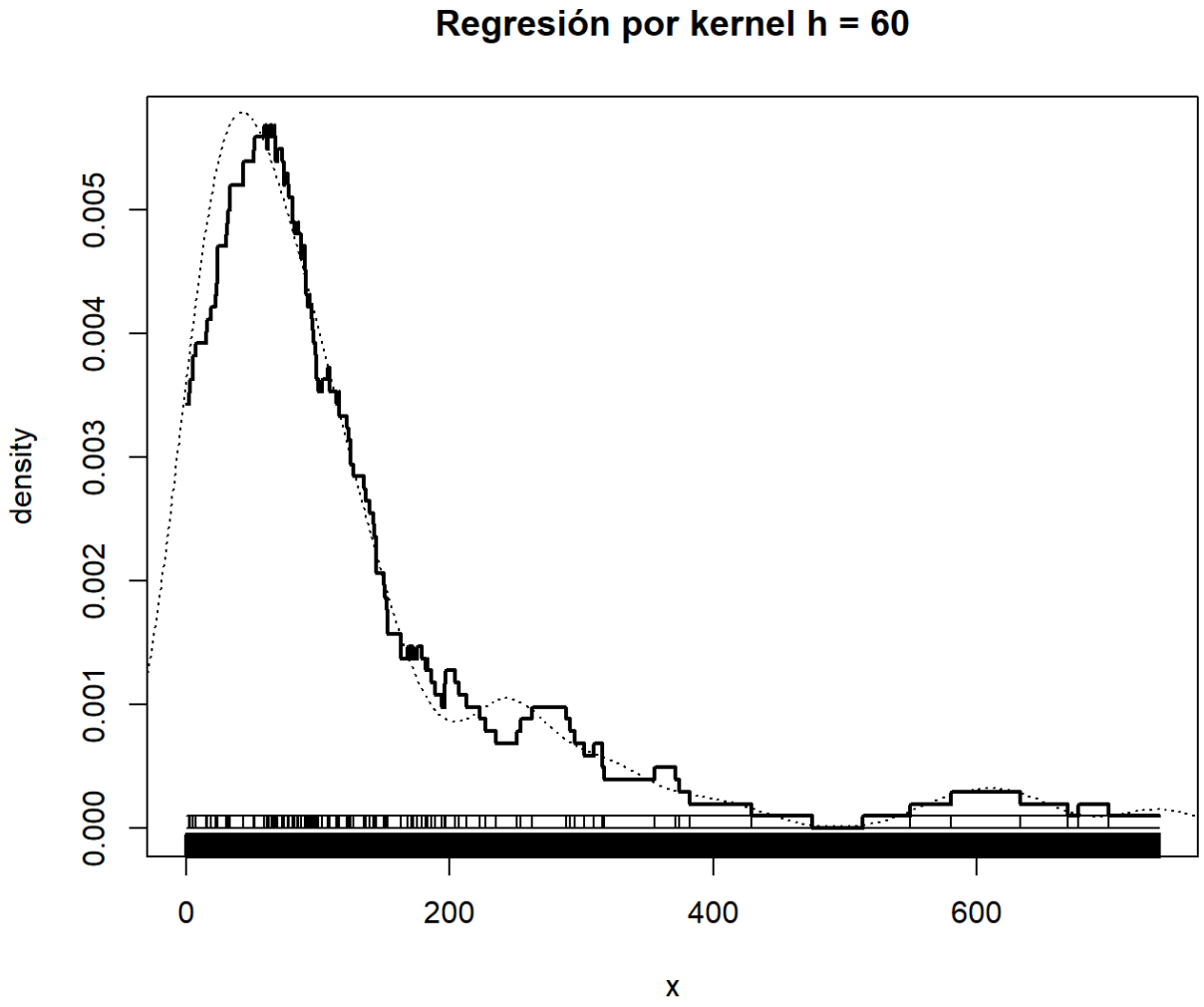
Code

```
## [1] "Seleccionó Un kernel Triangular para la Estimación"
```



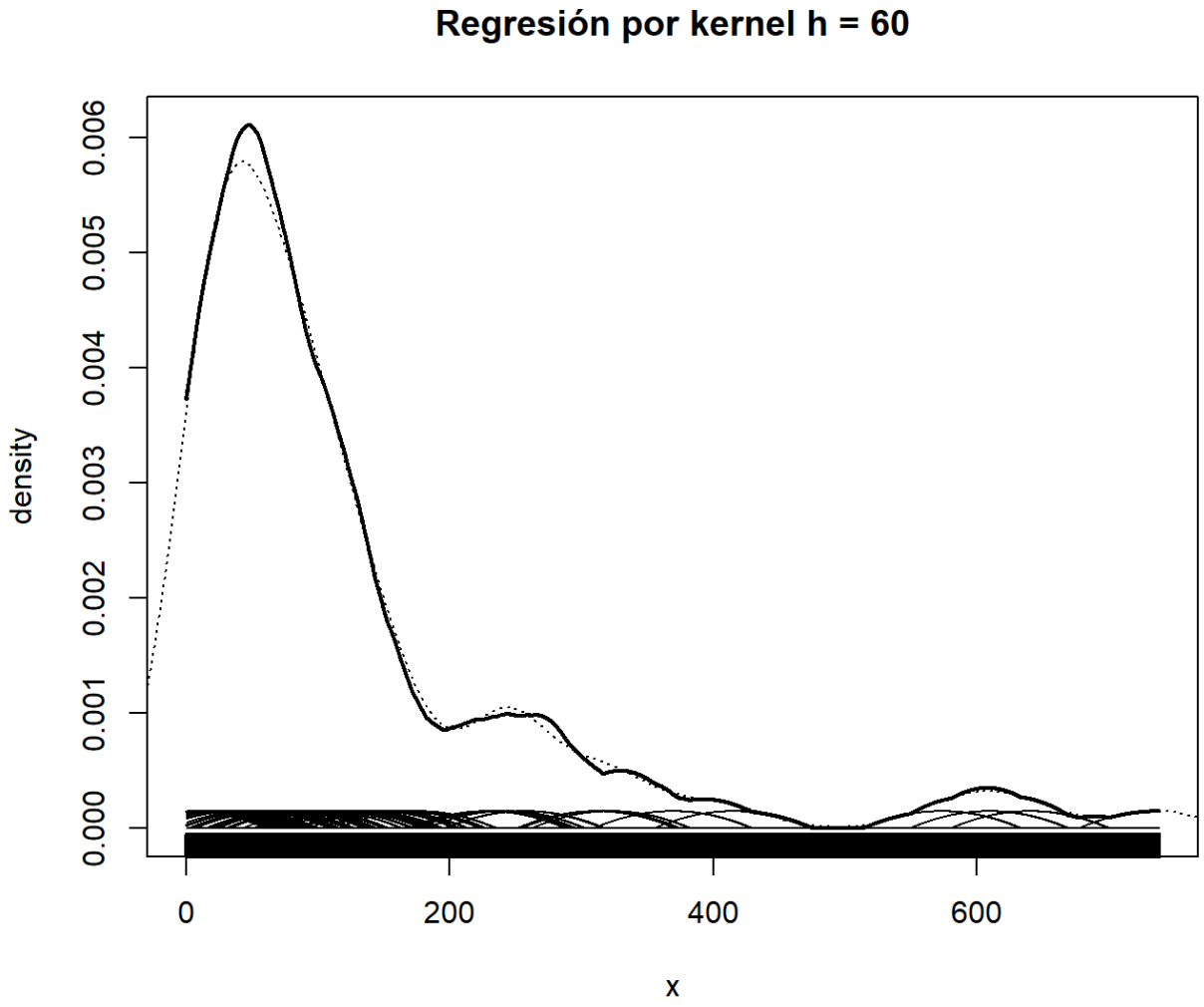
Code

```
## [1] "Seleccionó Un kernel Uniforme para la Estimación"
```



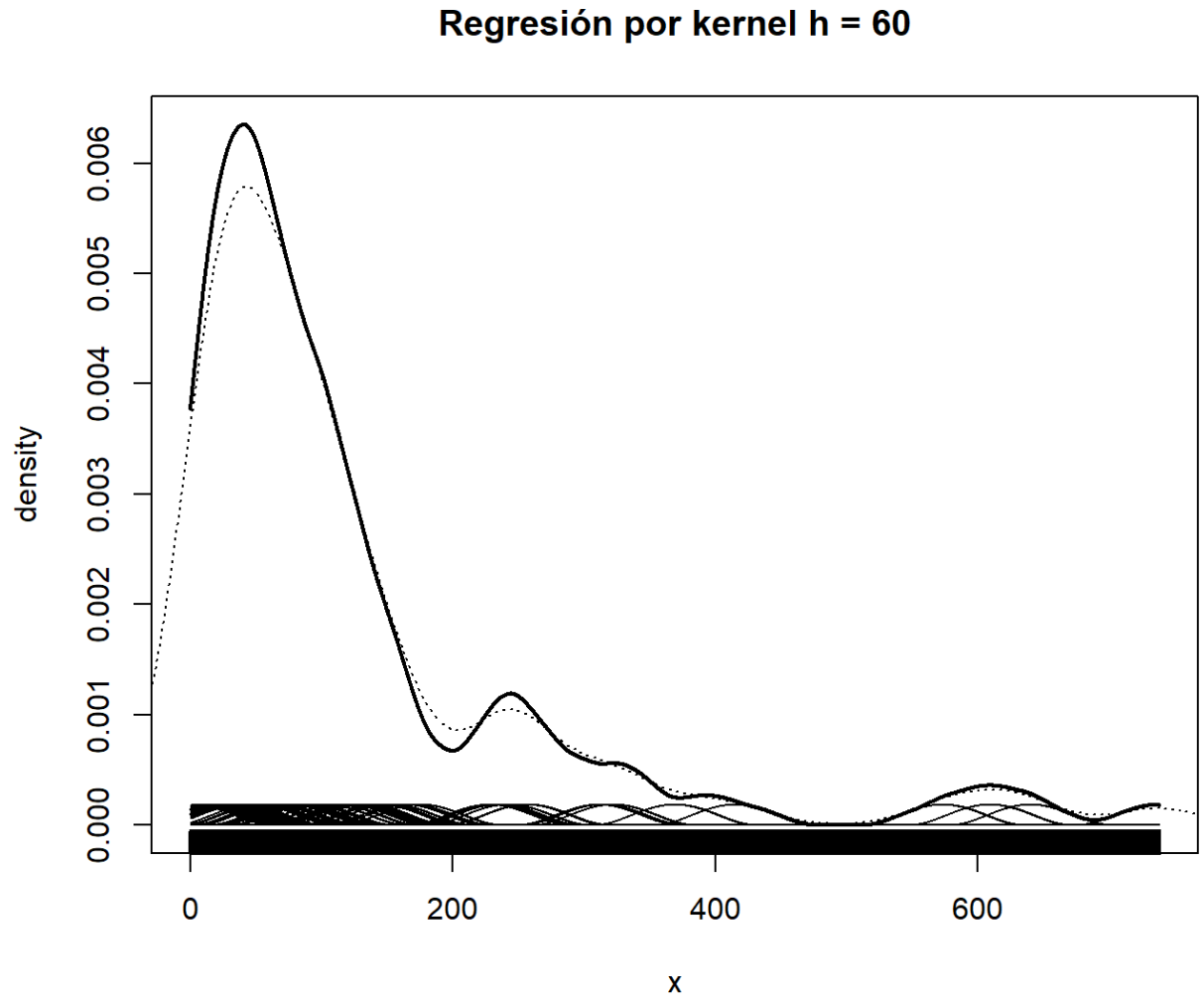
Code

```
## [1] "Seleccionó Un kernel Epanechnikov para la Estimación"
```



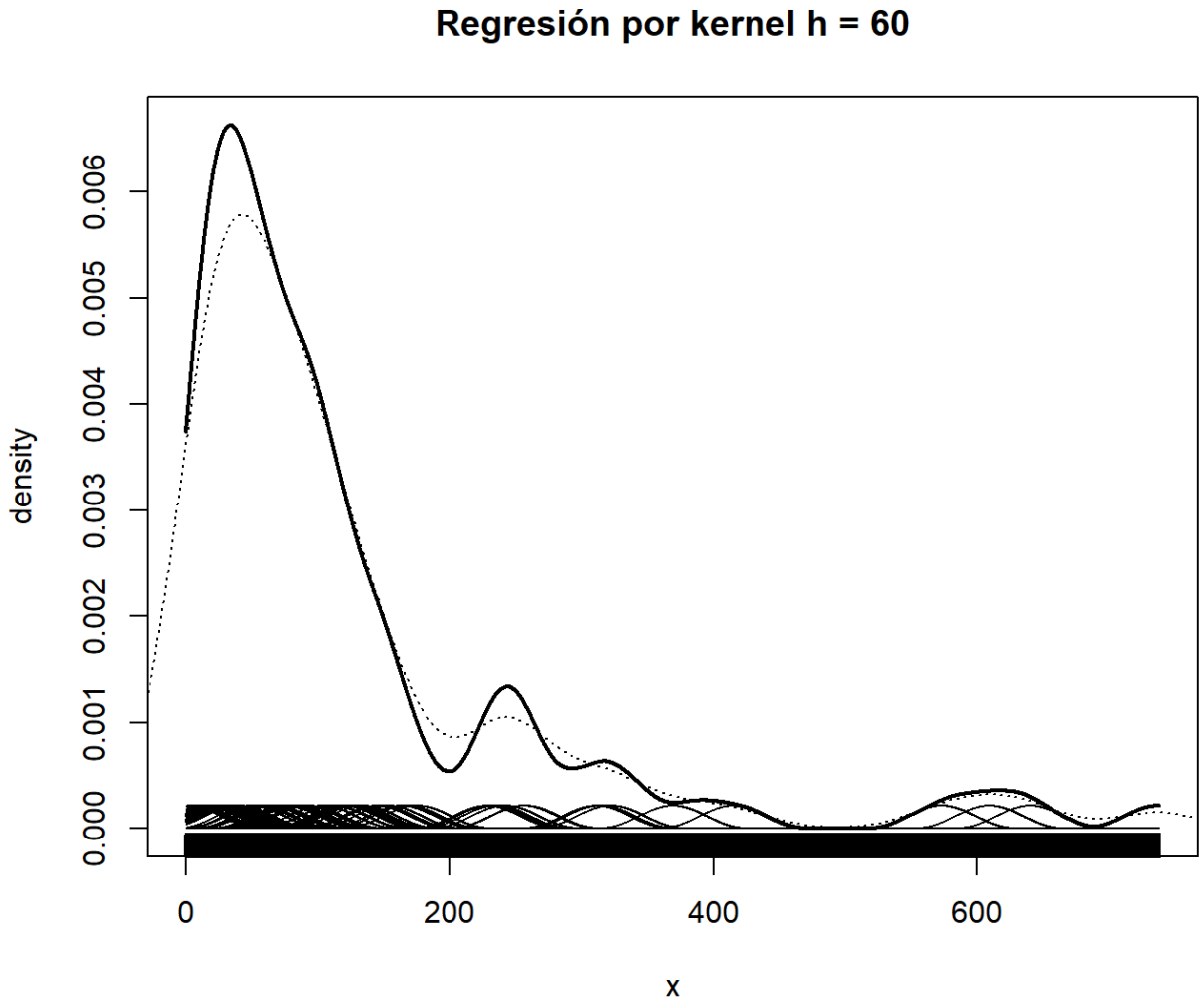
Code

```
## [1] "Seleccionó Un kernel Biweight para la Estimación"
```



Code

```
## [1] "Seleccionó Un kernel Triweight para la Estimación"
```



Un valor del parámetro de suavidad de 60, genera que la estimación con el kernel gaussiano se encuentra muy suavizado. El tener una h muy grande implica el trade off que ha mayor h, nayor sesgo, y a menor h, mayor variabilidad En este sentido, la densidad estimada se encuentra muy suavizada. Para los otros tipos de kernel, lo interesante es que presentan un comportamiento aproximado a un kernel gaussiano, sin embargo, no tan suavizado con un h de 60 como el kernel gaussiano.

Conclusion, el parámetro de suavizamiento de 30, se va a considerar el mejor entre el de 20 y el de 60. Con el de 20, h es muy chico,lo cual deja ver curvas no lo suficientemente suavizadas, y con ello muchas curvas espurias. Con un h de 60, por otro lado, genera una suavización de más, ocultando de más información sobre los datos. Recordando, el trade off entre el sesgo y la variabilidad, que a mayor h tenemos más sesgo, y a menor h mayor variabilidad. De esta forma, con un parámetro de suavidad de 30, la densidad estimada parecería estár mejor estimada que con los anteriores valores de h, dejandola lo suficientemente suave para obtener una buena estructura de información sobre los datos.

EJERCICIO 7

7. Cargue en R al conjunto de datos “Ma´iz.csv”, el cual contiene el precio mensual de la tonelada de ma´iz y el precio de la tonelada de tortillas en USD. En este ejercicio

tendr´a que estimar los coeficientes de una regresi´on lineal simple.

a. Calcule de forma expl´icita la estimaci´on de los coeficientes via m´inimos cuadrado y ajuste la regresi´on correspondiente. Concluya.

Code

Conociendo los datos

Code

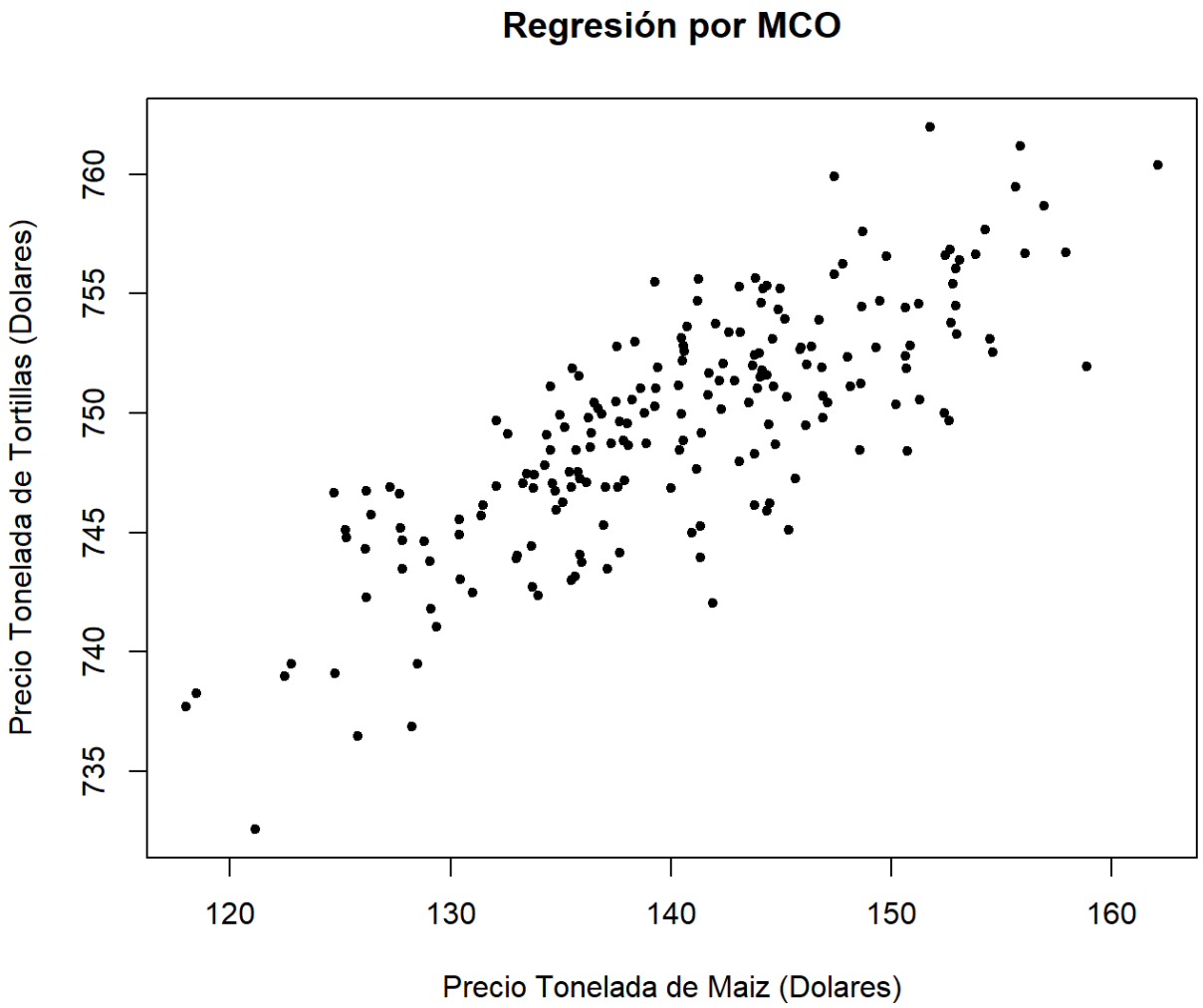
```
##   P..Tonelada.Maíz P..Tonelada.Tortilla
## 1         138.9115             748.7452
## 2         144.9533             755.2285
## 3         122.7757             739.5067
## 4         152.4164             750.0170
## 5         133.0140             744.0338
## 6         148.6396             751.2484
```

Code

```
##
##  Pearson's product-moment correlation
##
## data:  data$P..Tonelada.Tortilla and data$P..Tonelada.Maíz
## t = 18.602, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7407565 0.8429957
## sample estimates:
##       cor
## 0.7975325
```

Como era de esperarse se tiene un coeficiente de corelación lineal positiva mayor al .5, expresando la relación lineal entre ambas variables, la cual es prudente por ser maíz insumo de algun tipo de tortillas

Code



Code

Con el fin de contrastar mi estimación explicita de los coeficientes de regresión Se ajusta una regresión en la cual modela la esperanza condicional de las toneladas de tortilla dado as toneladas de máíz

Code

```
##
## Call:
## lm(formula = data$P..Tonelada.Tortilla ~ data$P..Tonelada.Maíz)
##
## Coefficients:
##           (Intercept)  data$P..Tonelada.Maíz
##                684.95                  0.46
```

Ahora, se estima los ceoficiente explicitamente por el método de mínimos cuadrados ordinarios.

Code

$$y = \beta_0 + \beta_1 x + \epsilon$$

Code

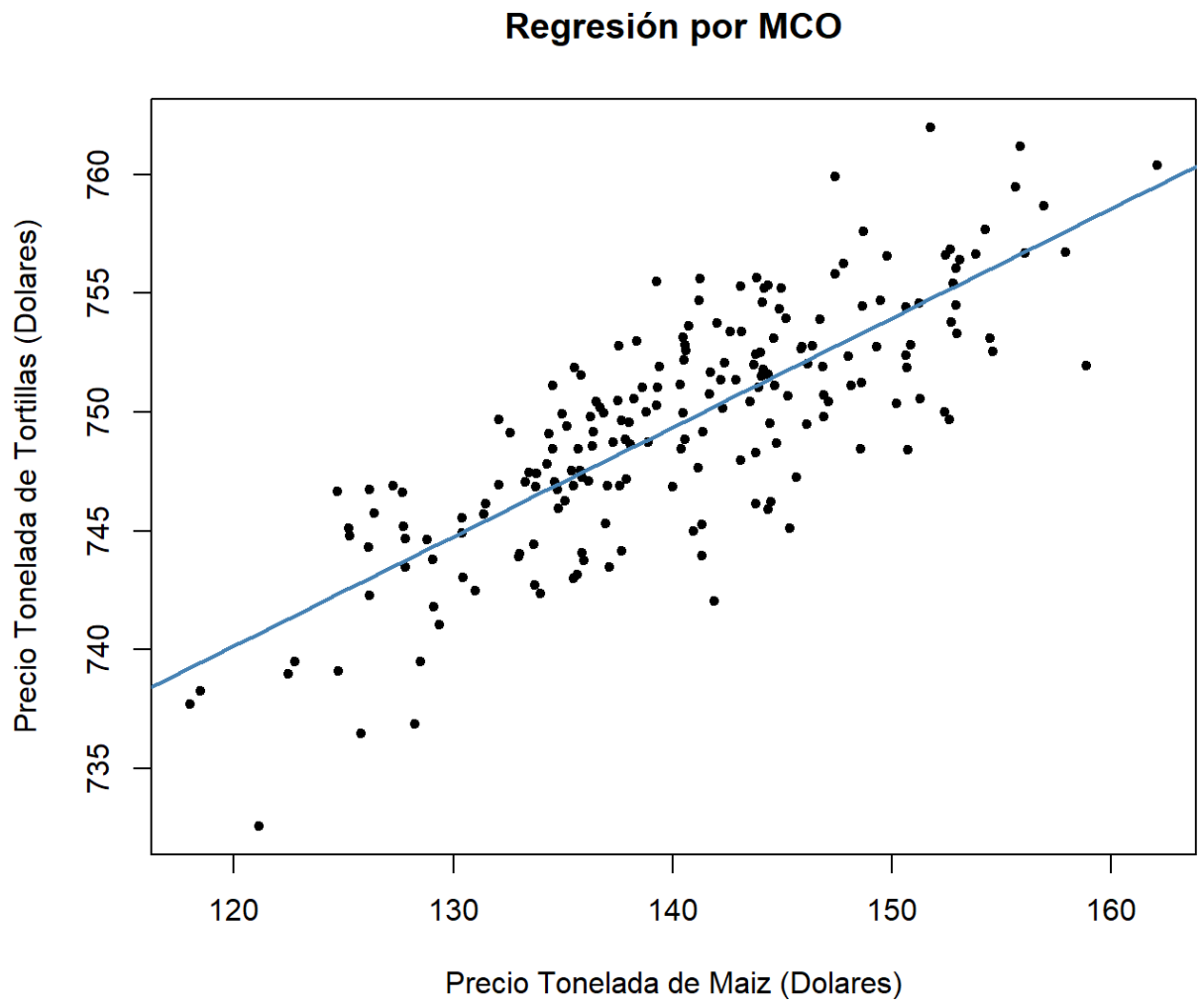
```
## [1] 684.9545
```

Code

```
## [1] 0.4600343
```

se observa que son los mismos que los que genera R con la función lm Interpretación, por cada incremento de una tonelada en el maíz genera un incremento de .4600, más su valor constante de 684.9545, en el caso de que querer hacer predicción sobre algún valor en especifico de x.

Code

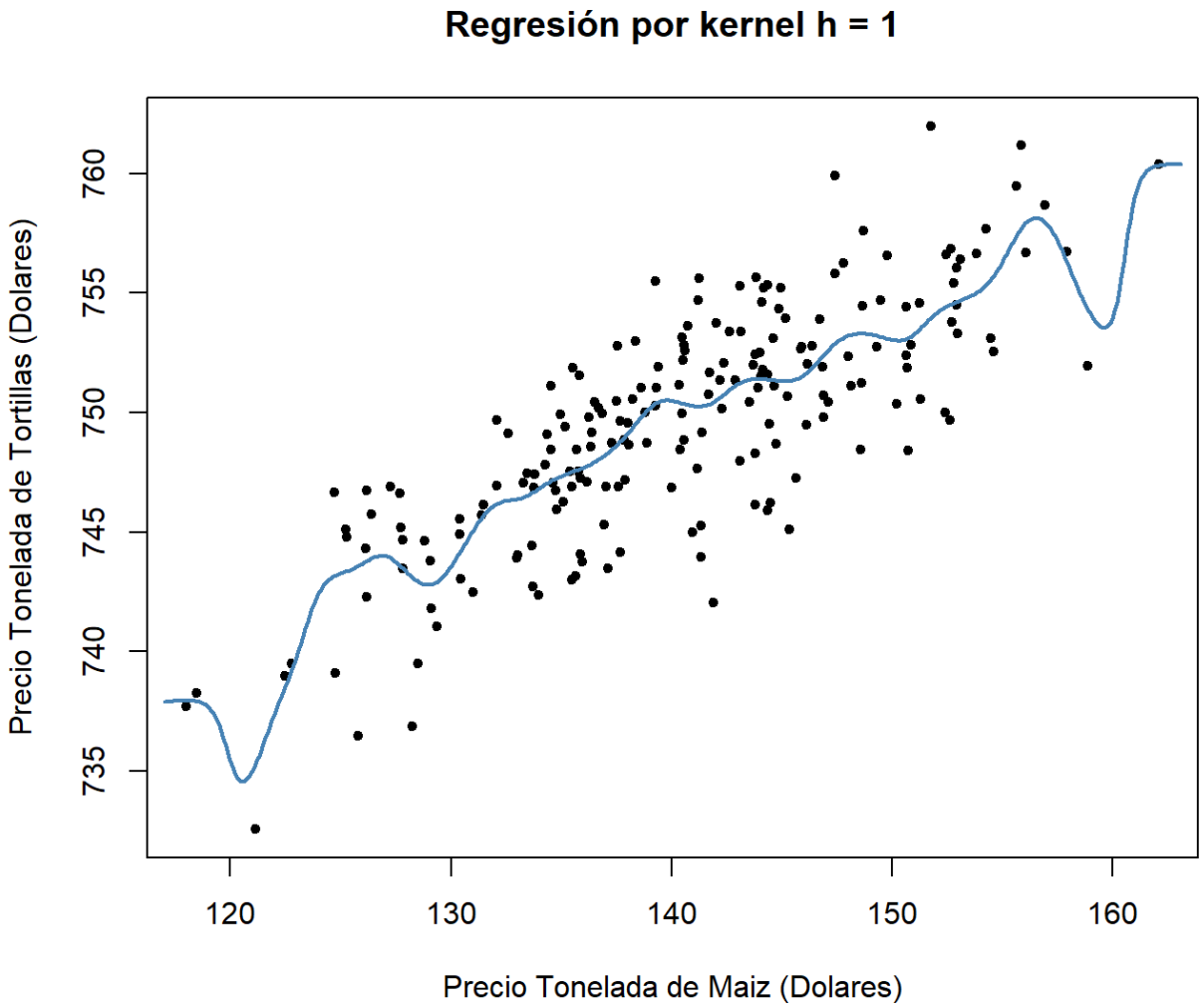


b) Calcule de forma expl□□cita la estimaci□on de los coecientes via regresi□on no-param□etrica tipo kernel (ver Nadaraya, E. A. (1964). Estimating Regression“. Theory of Probability and its Applications. 9 (1): 141{2. doi:10.1137/1109020 (doi:10.1137/1109020)) y ajuste la regresi□on correspondiente. Concluya.

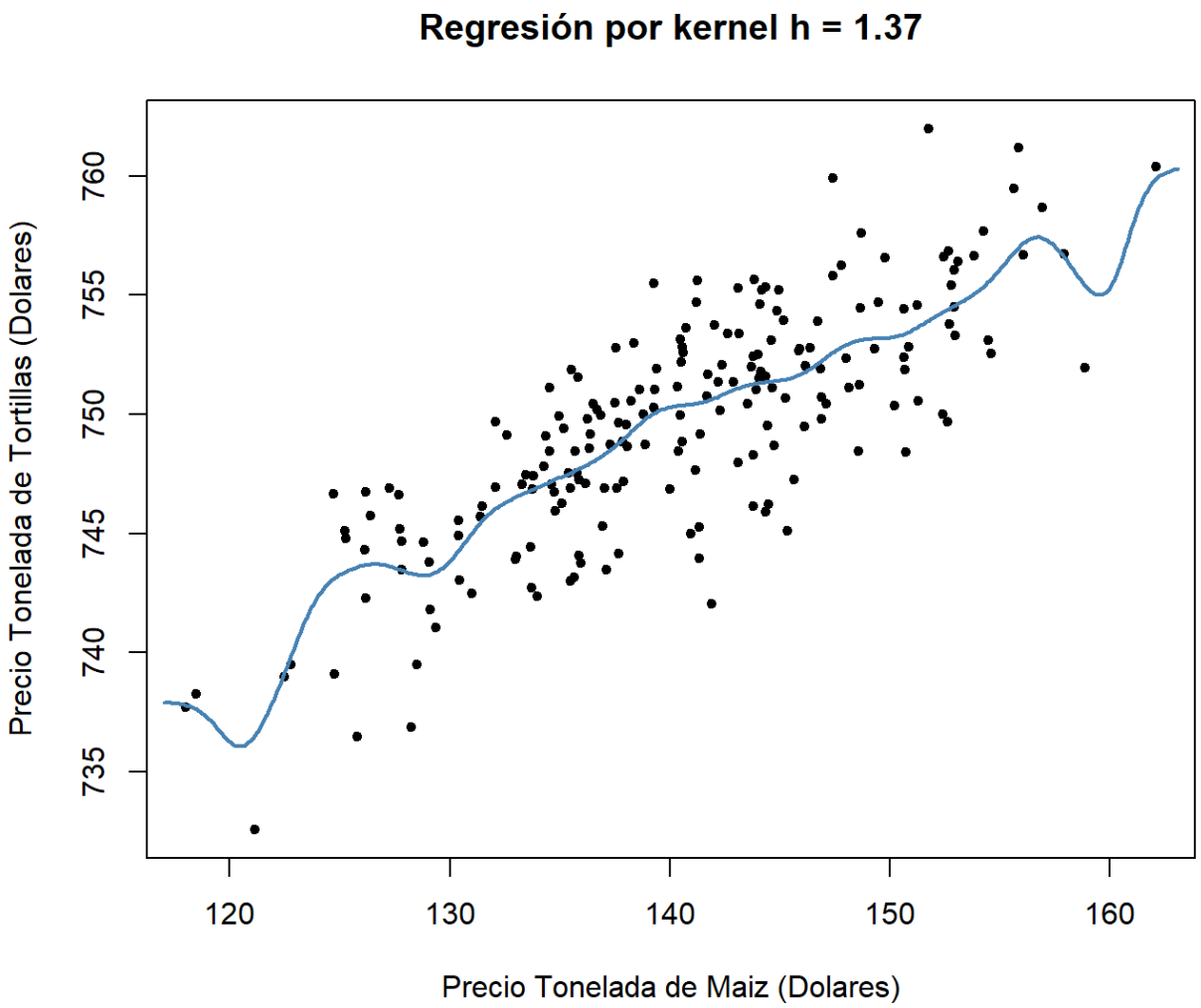
Code

Code

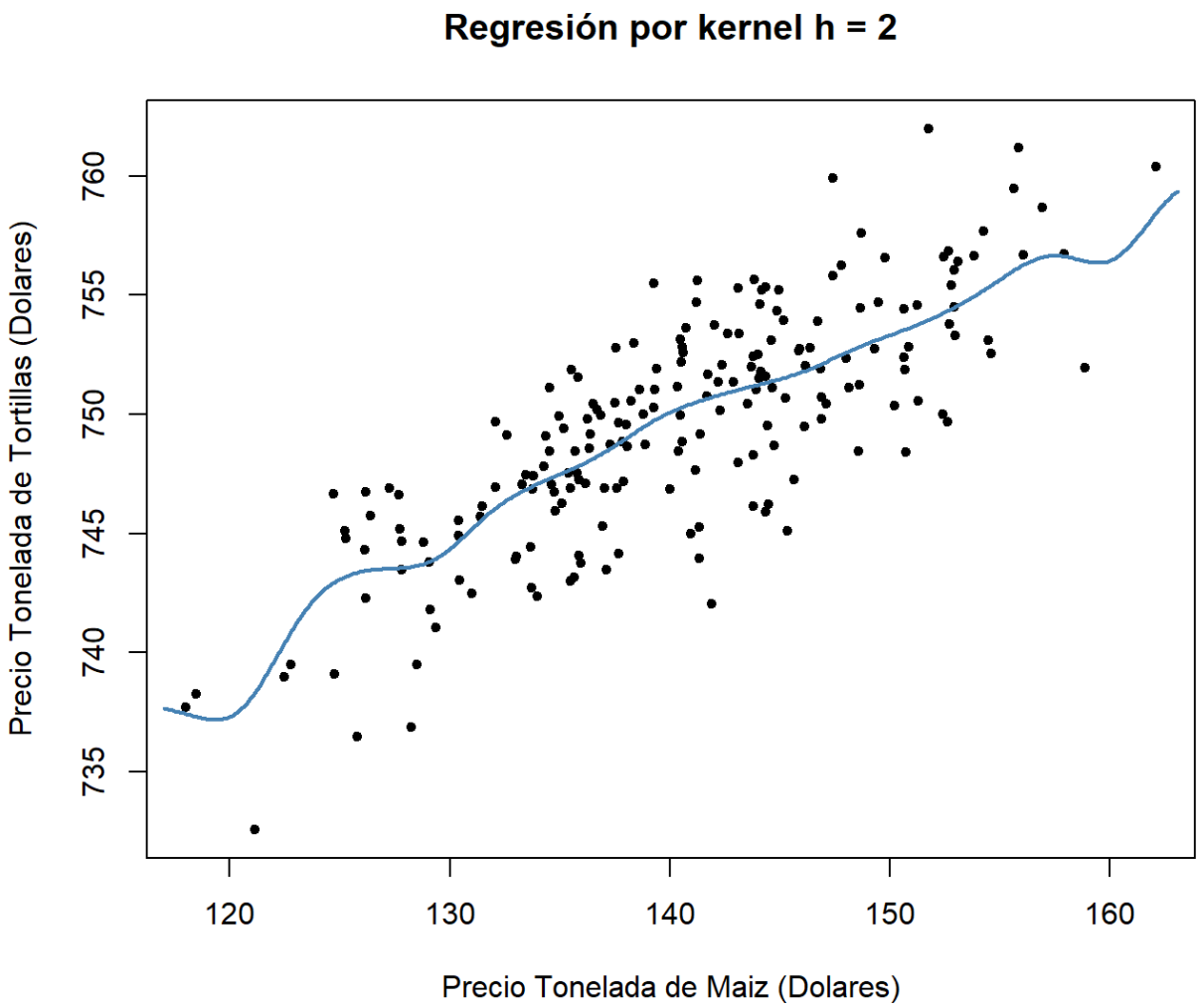
Code



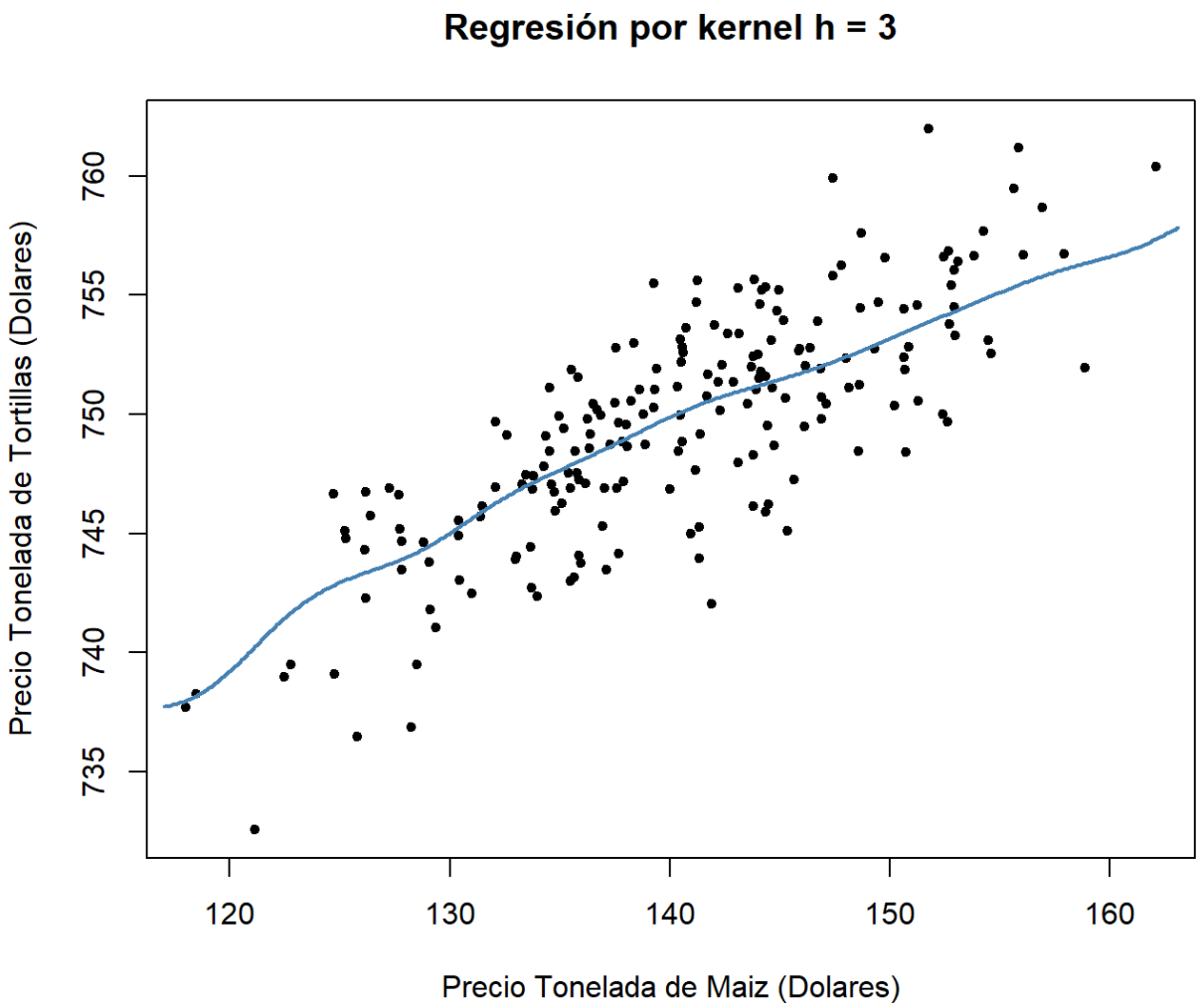
Code


Regresión por kernel h = 1.37
Precio Tonelada de Tortillas (Dolares)
735
740
745
750
755
760
120
130
140
150
160
Precio Tonelada de Maiz (Dolares)

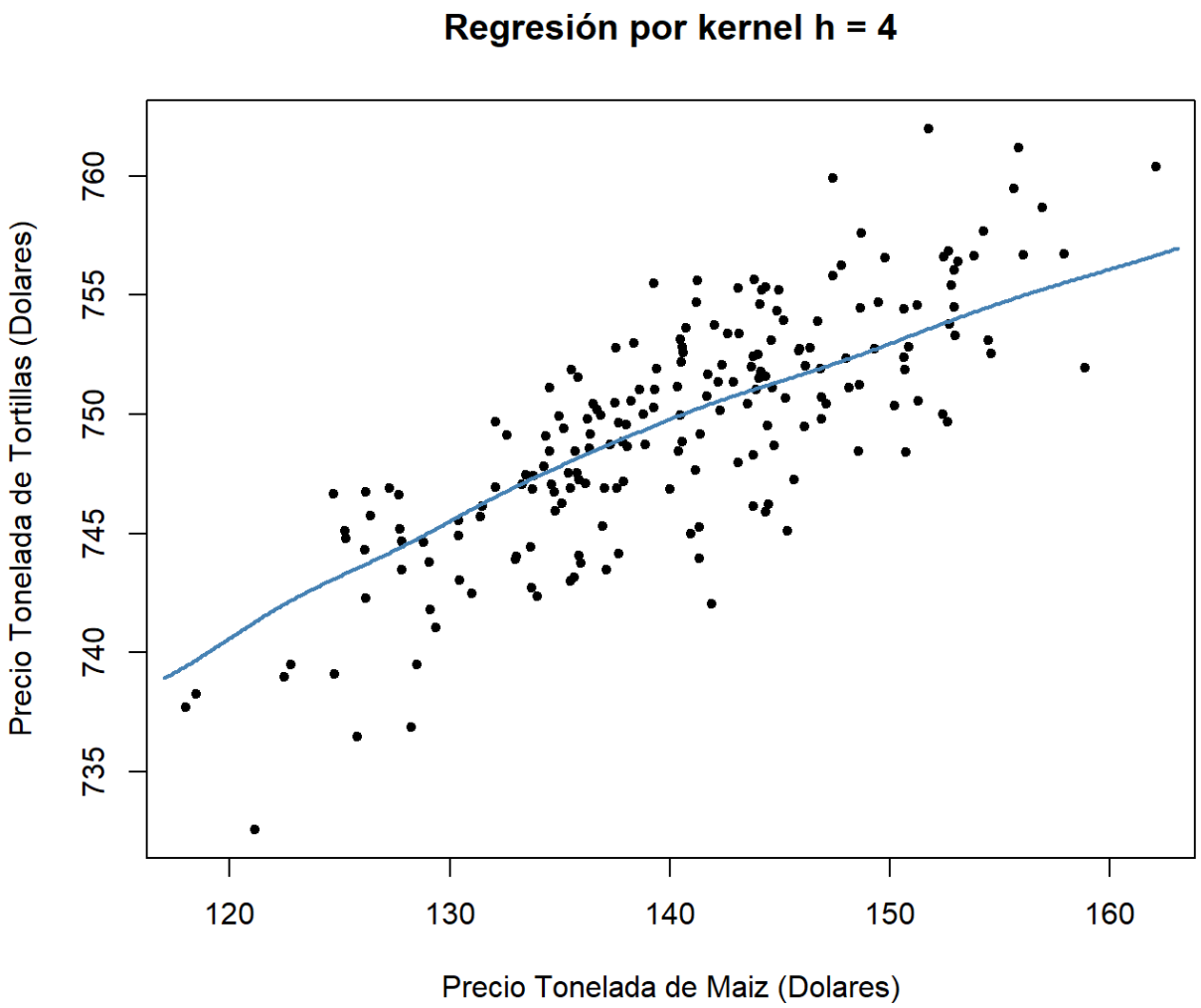
Code


Regresión por kernel h = 2
Precio Tonelada de Tortillas (Dolares)
735
740
745
750
755
760
120
130
140
150
160
Precio Tonelada de Maiz (Dolares)

Code

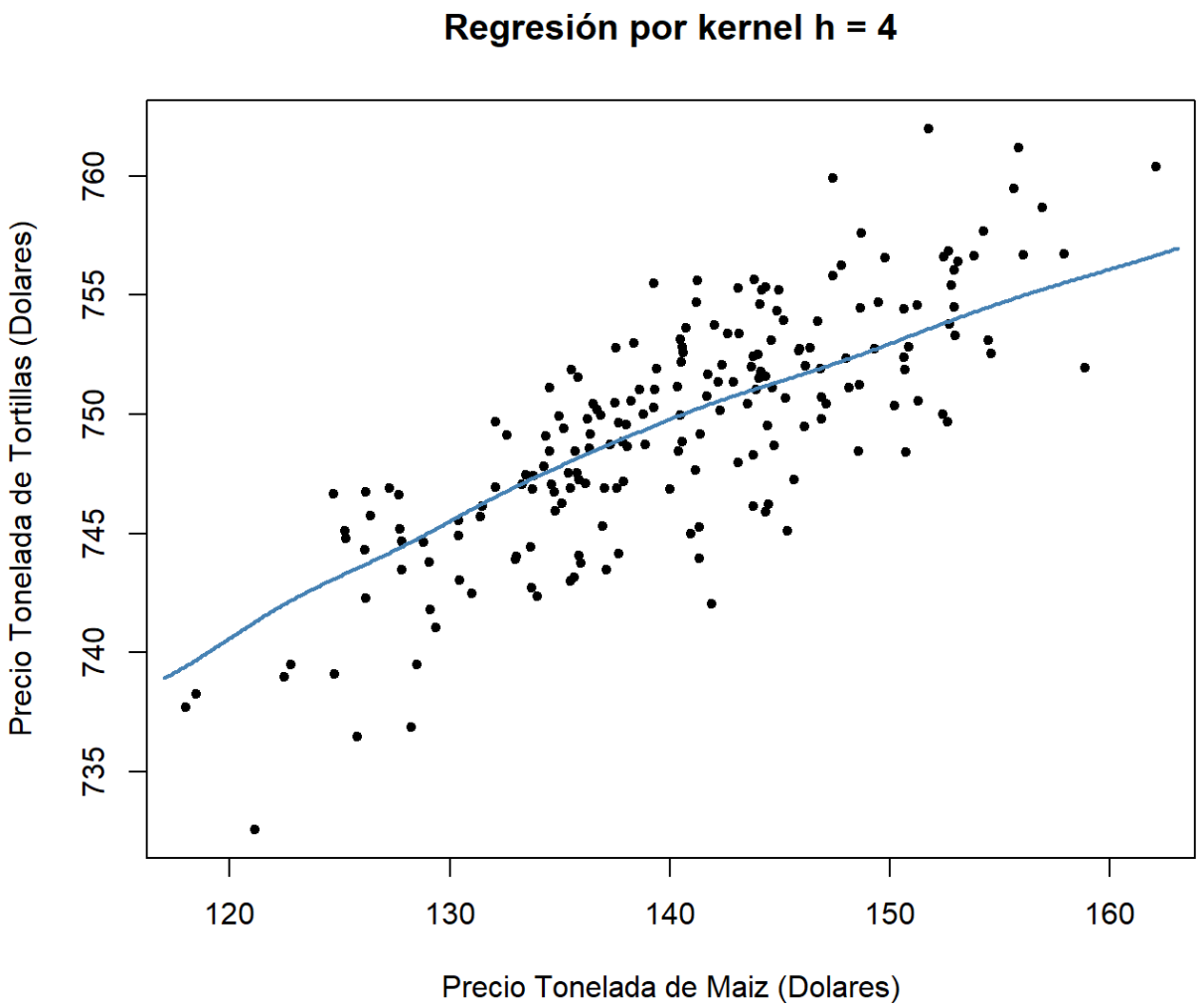

Regresión por kernel h = 3
Precio Tonelada de Tortillas (Dolares)
735
740
745
750
755
760
120
130
140
150
160
Precio Tonelada de Maiz (Dolares)

Code

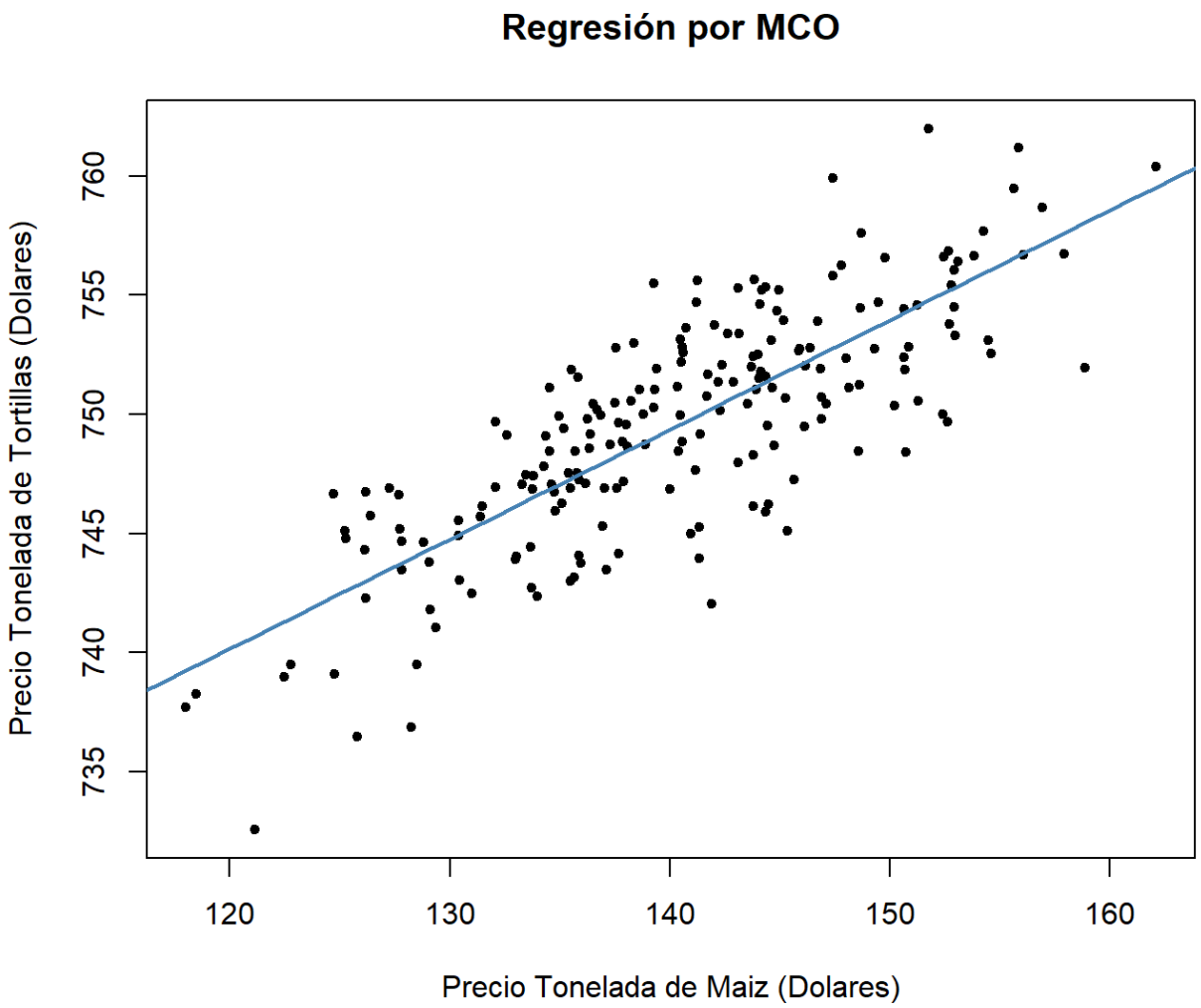


Comparando con Con un h = 4 Kernel regresión y MCO.

Code


Regresión por kernel h = 4
Precio Tonelada de Tortillas (Dolares)
735
740
745
750
755
760
120
130
140
150
160
Precio Tonelada de Maiz (Dolares)

Code



c. Compare ambos resultados. >Qu□e diferencias observa?

Concluya. Cuando se utiliza la regresión con el método de kernel el parámetro de suavidad al modificarlo, puede encontrar un mejor ajuste que la estimación con mínimos cuadrados ordinarios. En cuanto la comparación, se puede observar, que si se utiliza un parámetro de suavidad de cuatro (h = 4), la regresión por kernel ajusta de una forma similar a la regresión con MCO.