Máster en Técnicas Estadísticas | Doctorado en Estadística e Investigación Operativa

UNIVERSIDADE DA CORUÑA

Facultad de Informática Coruña

Métodos no Paramétricos

Evaluación Continua

Guillermo Portela Vázquez

6 de enero de 2024

Índice

Parte A

1. Primera parte

1.1. Enunciado

He elegido hacer el ejercicio 2 del boletín. Pongo aquí su enunciado:
Probar el siguiente resultado:

Si F_x es la función de distribución de una variable aleatoria X absolutamente continua, entonces la variable aleatoria $Y = F_X(X)$ se distribuye de acuerdo a una $U[0,1]$.

Usar este resultado para generar una muestra aleatoria simple de tamaño 50 de una distribución exponencial de media 2. Chequear la bondad de ajuste de esta muestra a una exponencial de media 2 con alguna prueba apropiada.

2. Resolución

2.1. Demostración teórica

Este resultado me costó entenderlo más de lo que me gustaría admitir y considero que en parte se debe a que cuesta (o al menos personalmente me costó) entender qué es lo que se está afirmando. Se trata, esencialmente, de un resultado de transformación de variables aleatorias. Voy a realizar un enfoque que puede parecer exageradamente lento para un lector experimentado (o que simplemente ya conozca sobradamente el resultado) pero que creo que hubiera sido la mejor forma de presentármelo por primera vez.

Consideremos X una variable aleatoria que sigue una distribución exponencial de media 2. Lo cual implica:

$$f_x(x) = \begin{cases} 0 & si \quad x < 0 \\ \frac{1}{2}e^{-x/2} & si \quad x \geqslant 0 \end{cases} \qquad F_x(x) = \begin{cases} 0 & si \quad x < 0 \\ 1 - e^{-x/2} & si \quad x \geqslant 0 \end{cases}.$$

Si consideramos F_x como una función con la que transformar la V.A. X para generar la variable Y, "ignorando" que se trata concretamente de la función de distribución de X y considerando $\{X_i\}_{i=1}^n$ una M.A.S. de X tendríamos:

$$\underbrace{Y_i}_{F_x(X_i)} = \frac{1}{2}e^{-X_i/2}. \quad [1] \tag{1}$$

[1] Notar que $Y_i > 0 < X_i \ \forall i$ por ello no doy la definición de Y_i "a trozos".

Con lo cual hemos generado $\{Y_i\}_{i=1}^n$, una nueva M.A.S. pero ahora ¿Qué distribución sigue esa muestra $\{Y_i\}_{i=1}^n$? El resultado nos afirma que es una muestra de una distribución $U[0,1]$ y se puede hacer una breve simulación en R con la función ecdf para corroborar que el resultado acierta al menos en este caso. Dicho de otra forma, el resultado, para cualquier variable aleatoria absolutamente continua, afirma lo siguiente:

$$f_y(y) = \begin{cases} 0 & si \quad y \in (-\infty, 0) \bigcup (1, \infty) \\ 1 & si \qquad y \in [0,1] \end{cases} \qquad F_y(y) = \begin{cases} 0 & si \quad y \in (-\infty, 0) \\ y & si \quad y \in [0,1] \\ 1 & si \quad y > 1 \end{cases} \tag{2}$$

¿Cómo podríamos probar este resultado? Una primera idea puede ser intentar probarlo para el caso particular de la exponencial comentada previamente. $X \sim exp\left(\frac{1}{2}\right)$[2]. Esto puede parecer mucha pérdida de tiempo pero, como se ha escogido la distribución con la que se pide trabajar en la segunda parte del ejercicio, vamos a encontrarnos con algún resultado útil.Además, cuando decía que iba a hacer un desarrollo "como me hubiera gustado que se me presentase este resultado por primera vez" ese desarrollo incluye poner un ejemplo de este estilo en este punto de la presentación.
Recordemos el teorema de transformación estrictamente monótona de variables aleatorias. Si $Y = g(X)$ y g es una función estrictamente monótona entonces:

$$f_y(y) = f_x\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right| \tag{3}$$

Consideremos la variable aleatoria transformada $Y = 1 - e^{-X/2} = F_x(X)$. Notar que como $X \sim exp(1/2)$, X toma valores en $[0,\infty)$. Además $f_x(X) > 0$, $F_x(0) = 0$ y $\lim_{X\to\infty} F_x(X) = 1$. Por tanto podemos afirmar que $F_x(X) \in [0,1]\ \forall X \in \mathbb{R}^+$. Vamos a querer aplicar el resultado (3) cambiando g por F_x. Para ello busquemos los ingredientes necesarios involucrados en esa expresión. Se tiene que:

$$Y = 1 - e^{-X/2} \Leftrightarrow Y - 1 = -e^{-X/2} \Leftrightarrow 1 - Y = e^{-X/2} \Leftrightarrow \log(1-Y) = -X/2 \Leftrightarrow -2\log(1-Y) = X$$

Por tanto $F_x^{-1}(y) = -2\log(1-y)$. Notar que esta expresión tiene sentido: $y \in [0,1]$ por tanto $1-y \in [0,1]$. Además $(1-y)$ solo toma el valor 0 cuando $y = 1 \Leftrightarrow X \to \infty$. Ahora derivando obrenemos:

$$\frac{d\left(F_x^{-1}(y)\right)}{dy} = \frac{2}{(1-y)} > 0\ \forall y \in [0,1)$$

Ahora usando (3) obtenemos, para $y \in [0,1]$:

$$f_y(y) = f_x\left(F_x^{-1}(y)\right)\left|\frac{dF_x^{-1}(y)}{dy}\right| \tag{4}$$

$$= \frac{1}{2} exp\left\{-\frac{1}{2}\left(-2\log(1-y)\right)\right\}\frac{2}{(1-y)} = \frac{e^{\log(1-y)}}{(1-y)} = 1. \tag{5}$$

Como queríamos probar.

Veamos ahora como probar el resultado principal para un caso más general donde solo sabemos que F_x es monótona estrictamente creciente (notar que el enunciado únicamente nos permite afirmar

[2]En muchas referencias (sobretodo referencias online) a la hora de especificar una distribución exponencial se usa la inversa de la media como parámetro, aunque no es lo que usábamos en la USC es con lo que me manejo más cómodamente.

que F_x es absolutamente continua, aun faltaría dar un paso más después).

Sea $Y = F_x(X)$ y $u \in (0,1)$ tenemos:

$$F_y(u) = \mathbb{P}(Y \leqslant u) = \mathbb{P}(\underbrace{F_x(X)}_{Y} \leqslant u) \overset{*}{=} \mathbb{P}\left(X \leqslant F_x^{-1}(u)\right) = F_x\left(F_x^{-1}(u)\right) = u \tag{6}$$

Donde, en la igualdad marcada con $*$, se ha usado que al ser F_x monotona estrictamente creciente entonces existe F_x^{-1} también monótona estricta. Ahora trivialmente si $u \in (-\infty, 0]$ entonces $F_y(u) = \mathbb{P}\left(F_x(X) \leqslant u \leqslant 0\right) = 0$ y si $u \in [1, \infty)$ entonces $F_y(u) = \mathbb{P}\left(F_x(X) \leqslant 1 \leqslant u\right) = 1$ Por tanto F_y coincide con la expresión de la función de distribución de una variable uniforme en $[0,1]$.
Falta por probar el resultado para V.A. absolutamente continuas. En este caso se descarta la hipótesis de monotonía estricta de la función de distribución. No obstante la función de distribución sigue siendo monótona creciente. Notar que la pérdida de monotonía estricta se da cuando hay un intervalo $\mathfrak{u}$ de valores de las x para las cuales $F_x(x)\big|_{x\in\mathfrak{u}} = a \;\; cte.$ Además esto implica que $f_x(x)\big|_{x\in\mathfrak{u}} = 0$. Para este caso conviene recordar la definición de la inversa de la función de distribución de la probabilidad como:

$$F_x^{-1}(x) := \inf\left\{t \;:\; F_x(t) \geqslant x\right\} \tag{7}$$

Con esta definición se verifica que:

a. F_x^{-1} es monótona creciente (no necesariamente estricta).

b. Si F_x es continua, entonces $F_x\left(F_x^{-1}(x)\right) = x, \;\forall x \in [0,1]$.

Veamos una **demostración** de estas dos afirmaciones:

a. Es prácticamente trivial por definición:

$$F_x^{-1}(a) \leqslant F_x^{-1}(b) \Leftrightarrow \inf\left\{t_a \;:\; F_x(t_a) \geqslant a\right\} \leqslant \inf\left\{t_b \;:\; F_x(t_b) \geqslant b\right\} \overset{*}{\Leftrightarrow} a \leqslant b.$$

Donde en $*$ he usado la monotonía de F_x.

b. Dado $x \in [0,1]$ se tiene que $F_x\left(F_x^{-1}(x)\right) = F_x\left(\inf\left\{t \;:\; F_x(t) \geqslant x\right\}\right)$. Si denotamos como $\widetilde{t}$ a ese ínfimo. Es decir $\widetilde{t} := \inf\left\{t \;:\; F_x(t) \geqslant x\right\}$. Tenemos, por definición, $F_x\left(\widetilde{t}\right) \geqslant x$ pero la desigualdad estricta no podría darse porque rompería con la hipótesis de continuidad de F_x. Veamoslo: por ser F_x monótona creciente se cumple que : $\lim_{t\to\widetilde{t}^-} F_x(t) = a \leqslant x$. Ahora, si suponemos que $F_x\left(\widetilde{t}\right) = b > x$ llegamos a lo siguiente:

$$F_x\left(\widetilde{t}\right) - \lim_{t\to\widetilde{t}^-} F_x(t) = b - a \geqslant b - x > 0 \implies F_x \text{ presenta una discontinuidad en el punto } \widetilde{t}.$$

□

Vistas estas dos propiedades de F_x^{-1} podemos obtener el resultado pedido en el enunciado repitiendo la cadena de argumentos dados en (6) de la siguiente forma:

$$F_y(u) = \mathbb{P}(Y \leqslant u) = \mathbb{P}(\underbrace{F_x(X)}_{Y} \leqslant u) \overset{\text{a.}}{=} \mathbb{P}\left(X \leqslant F_x^{-1}(u)\right) = F_x\left(F_x^{-1}(u)\right) \overset{\text{b.}}{=} u. \tag{8}$$

□

2.2. Generando una muestra de la exponencial

Vamos a aprovechar la expresión $F_x^{-1}(y) = -2\log(1-y)$ obtenida en la sección anterior, junto a una M.A.S. uniforme en $[0,1]$ para generar la muestra de $Y \sim exp(1/2)$. Basta considerar $\{Y_i\}_{i=1}^{50}$ una muestra de $Y \sim U[0,1]$ como si se hubiera obtenido como resultado de hacer la transformación $Y = F_x(X)$ donde $X \sim exp(1/2)$ por tanto:

$$F_x^{-1}(Y_i) = -2\log(1-Y_i) = F_x^{-1}(\underbrace{F_x(X_i)}_{Y_i}) = X_i$$

donde al final obtenemos $\{X_i\}_{i=1}^{50}$ que se trata de una M.A.S. de $X \sim exp(1/2)$. Una vez hecho esto utilizando distintos tests (Kolmogorov-Smirnov, Anderson-Darling, Cramer von Mises) podemos realizar las pruebas de bondad de ajuste que se nos piden en el enunciado. Recordar que el test que queremos realizar es:

$$\begin{cases} H_0: & F \sim exp(1/2) \\ H_1: & F \nsim exp(1/2). \end{cases} \tag{9}$$

Donde F representa la función de distribución que sigue la muestra $\left\{F_x^{-1}(Y_i)\right\}_{i=1}^{50}$.
Las siguientes sencillas lineas de R llevan a cabo el proceso descrito y, a su vez, realizan 3 contrastes de bondad de ajuste.

```
library(goftest)
set.seed=(1234)
Y=runif(50)
X=(-log(1-Y)*2)
stats::ks.test(x=X, y="pexp", rate = 0.5)
goftest::cvm.test(x = X, null = "pexp", rate=1/2)
goftest::ad.test(x=X, null = "pexp", rate=1/2)
```

Donde se han obtenido los siguientes p-valores:

$$\text{KS test: } 0{,}8636 \quad \text{CvM test: } 0{,}6858 \quad \text{A-D test: } 0{,}6473 \tag{10}$$

Por lo que concluyo que el proceso ha llegado al resultado esperado.

Parte B

3. Segunda Parte

3.1. Enunciado

He decidido realizar el ejercicio 4 del boletín. Pongo aquí su enunciado:

Considérese la base de datos del archivo **Boston Housing Data** con información diversa sobre la vivienda en 506 barrios del área metropolitana de Boston en 1978 y que está disponible en el archivo **BostonHousing.rda**. Se pide:

a. Estimar la densidad de la variable "porcentaje de población con estatus social en categoría inferior" (lstat) con: (i) un histograma con origen en 20 y ancho de clase 1,5, (ii) estimadores núcleo con selectores de banda obtenidos por la regla del dedo, la regla plug-in de Sheater-Jones y por validación cruzada (chequear que se minimiza adecuadamente la función de validación cruzada), y (iii) el mejor ajuste normal. Mostrar todos los ajustes en un único gráfico y comentar los resultados.

b. Segmentar el archivo en dos grupos $\mathcal{C}_1$ y $\mathcal{C}_2$ caracterizados porque la variable "número medio de habitaciones por vivienda" (rm) sea menor o mayor que su mediana muestral, respectivamente. Chequear analítica y gráficamente la hipótesis nula de que la densidad de la variable lstat es idéntica en ambos grupos. Comentar los resultados.

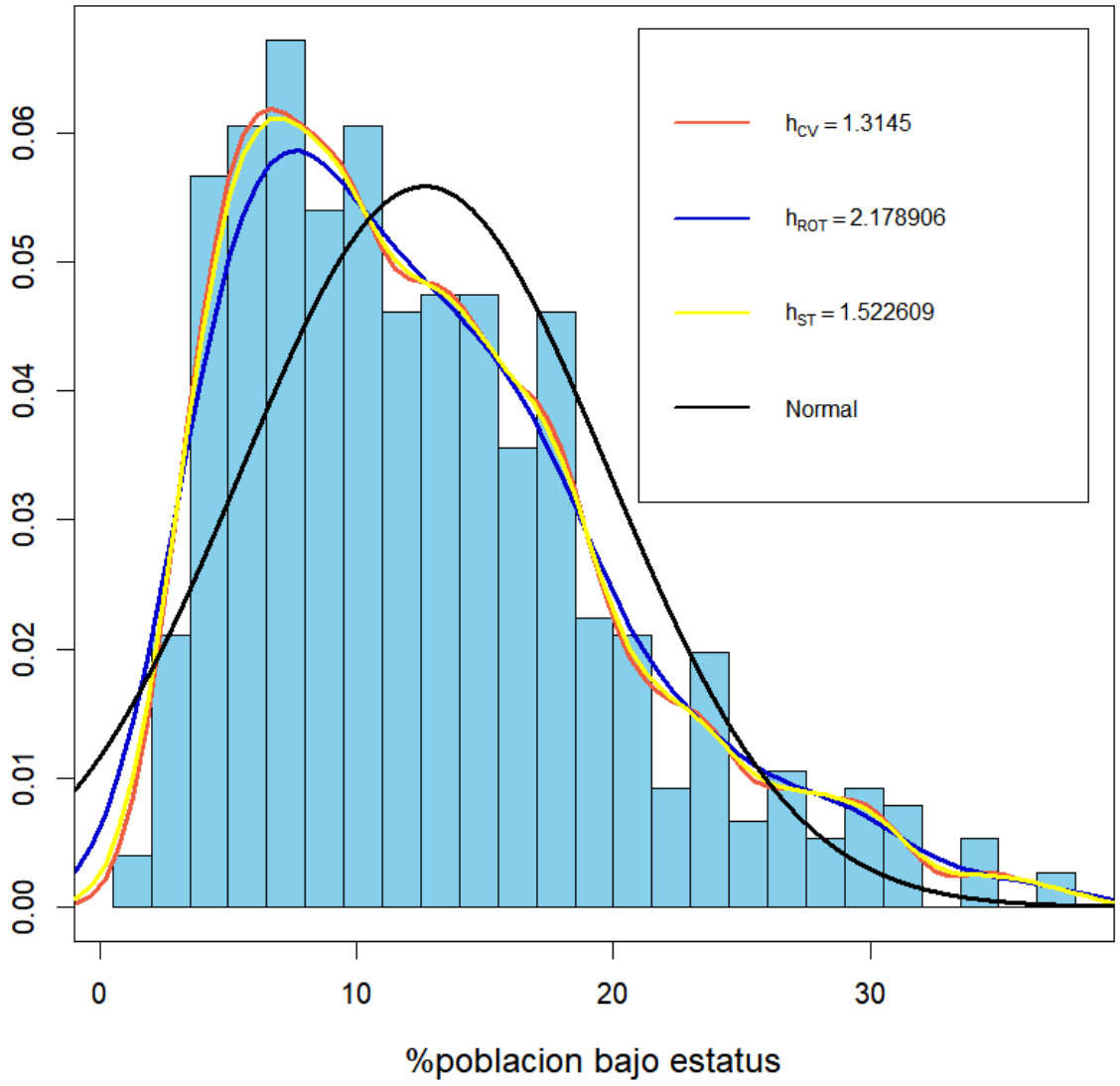
c. Explorar el comportamiento de ajustes polinómicos de hasta orden 2 para explicar el comportamiento de la variable rm en función de la variable lstat. Comparar su comportamiento con los ajustes no paramétricos de tipo kernel local lineal y local cúbico. Discutir los resultados.

d. Emplear el algoritmo loess para predecir los valores de rm sobre los cuartiles muestrales de lstat.

4. Resolución

4.1. Apartado a.

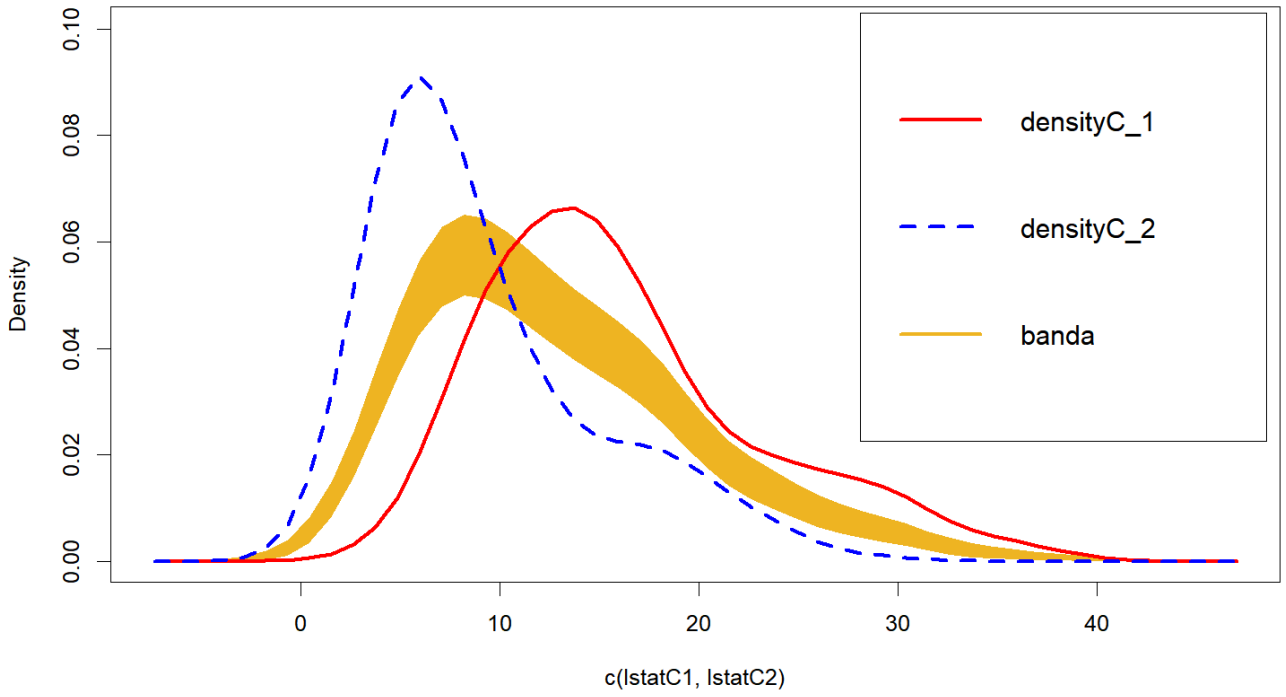
Haciendo uso de los distintos comandos de los scripts de R vistos en clase podemos generar la siguiente gráfica:



Se aprecia que una normal no se ajusta para nada a los datos (el máximo de la densidad está desplazado hacia la derecha) además de sufrir efecto frontera. De los 3 estimadores no paramétricos el que peor lo hace es el que utiliza parámetro de suavizado por la regla del dedo, es demasiado suave y el que más sufre del efecto frontera. Entre selector plug-in y validación cruzada están más parejos los resultados: ambos tienen poco efecto frontera aunque mayor variabilidad, sin embargo el estimador de validación cruzada aproxima más el histograma que el de plug-in en máximos y valles sin tener siginificativamente más variabilidad que este último.

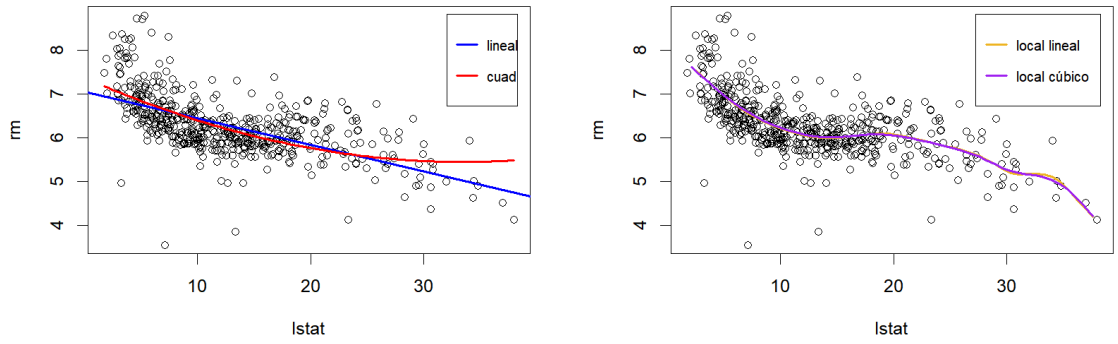
4.2. Apartado b.

Tras segmentar el dataframe según el criterio especificado y usar la función *sm.density.compare* del paquete "sm" obtenemos un tajante p-valor de 0 y la siguiente gráfica:



El resultado no debería ser sorprendente. Al fin y al cabo el conjunto $\mathcal{C}_1$ es la mitad "baja" de los barrios en cuanto a número de habitaciones por vivienda. Es de esperar que en los barrios con las casas más pequeñas las personas que los habitan tengan menos recursos económicos y por tanto la densidad aparezca más desplazada a la derecha en el gráfico (teniendo en cuenta que cuanto más a la derecha aparece un valor más porcentaje de población tiene "estatus social de categoría inferior" en ese barrio).

4.3. Apartado c.



Vemos a la izquierda el resultado al intentar ajustar un modelo no local lineal y un modelo no local

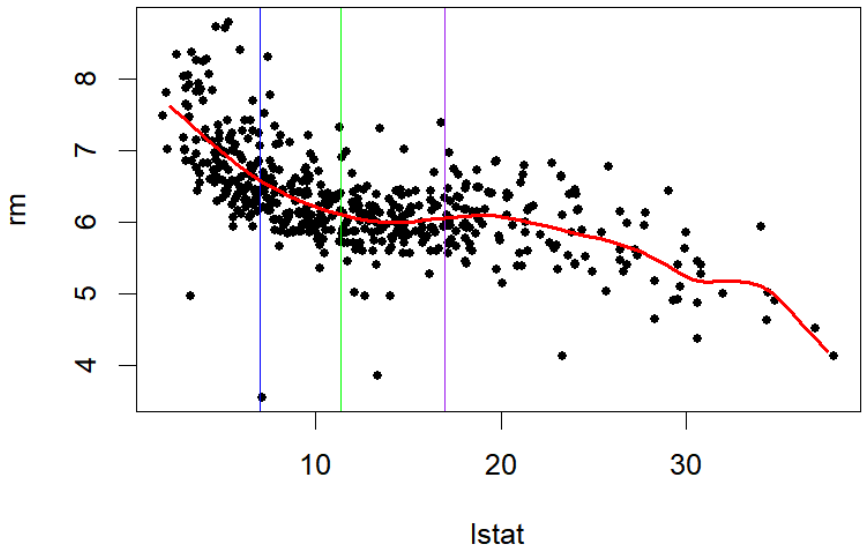
cuadrático. Ambos con errores notables: El cuadrático se comporta muy mal a la derecha del gráfico y el lineal muy mal a la izquierda. En la imagen de la derecha están **ambos** ajustes polinómicos locales pero como resulta que pintan casi la misma curva están muy superpuestos, curiosamente esta pequeña diferencia desaparece completamente si elegimos la ventana con *cross validation* para ambas estimaciones, parece que se modifican más las curvas al cambiar el método de selección de ventana que al cambiar el grado del polinomio local al que ajustarlas. En la imagen, el método escogido fue plug-in Ruppert, Sheater y Wand para la local lineal y *cross validation* para local cúbico. Si hubiera usado sm.regression con la ventana por defecto también saldría una curva (distinta a las 2 de la imagen) idéntica para ambos grados de polinomio pues se superpondrían totalmente ambos ajustes. Sea como fuere parece acertar mucho mejor el ajuste lineal polinómico (de cualquier grado) que el ajuste no local.

4.4. Apartado d.

Utilizando la función loess del paquete stats de R hice las siguientes predicciones para cada cuartil de **lstat**

Primer cuartil	Mediana	Tercer cuartil
6.559848	6.167779	6.067860

Que sigue con la idea de que en barrios con mayor porcentaje de vecinos con menor estatus social baja el número medio de habitaciones por casa ademas:



Vemos que los 3 cuartiles están en el grueso de los valores a la derecha y que la mayoría de esos valores oscilan entre 5,5 y 7 por lo que son predicciones bastante razonables.