Homework 10

Rafael Espinosa Castañeda

1)

Consider the nonlinear regression model,

$$y_i = m(x_i) + \sigma\epsilon_i, \quad i = 1, 2, ..., n,$$

where $n = 100$ and the data (x_i, y_i) are provided in "HW10_data.csv"

(1) (10 pts) Using the uniform kernel giving weights of the form

$$w_{i,h}(x) = \frac{\mathbf{1}(|x_i - x| < h)}{\sum_{i=1}^{n} \mathbf{1}(|x_i - x| < h)},$$

find the estimator of $m(0)$; i.e. $\widehat{m}(0)$, obtained by minimizing

$$\sum_{i=1}^{n} w_{i,h}(0)\,(y_i - \theta)^2$$

with $h = 0.1$.

The Nadaraya - Watson estimator is given by

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} y_i\, K((x - x_i)/h)}{\sum_{i=1}^{n} K((x - x_i)/h)}$$

where we will use the uniform kernel

$$K(u) = \frac{1}{2}\,\mathbb{1}(|u| < 1)$$

and at $x = 0$

$$\hat{m}(0) = \frac{\sum_{i=1}^{n} y_i\, K(-x_i/h)}{\sum_{i=1}^{n} K(-x_i/h)}$$

So $K(-x_i/h) = \frac{1}{2}\mathbb{1}\left(\frac{|x_i|}{h} < 1\right)$

Using R code, we obtain

$$\hat{m}(0) = -0.111118$$

```
df<-read.csv("D:/MAESTRIA_AUSTIN/RegressionAndPrediction/hw10/Hw_10_data.csv")
x<-0
y<-0
y<-df$y

x<-df$x
h<-0.1
k<-0

for (i in 1:100)
{
  if ( abs(-x[i]/h)< 1)
  {
      k[i]<-0.5
  }
  else{
    k[i]<-0
  }
}

print(k)

m_hat<-sum(y*k)/sum(k)

print (m_hat)
```

(2) (10 pts) Write down an expression for and hence find the value of the bias of $\widehat{m}(0)$ if the true function $m(x)$ is linear with slope 1.2.

$m(x)$ is linear so it has the form

$$m(x) = ax + b \text{, where } a = 1.2$$

So $m'(x) = a = 1.2$ $m''(x) = 0$

We use that

$$\text{Bias}(\hat{m}(x)) = E[\hat{m}(x)] - m(x)$$

$$= m'(x) \frac{\sum_{|x-x_i|<h} (x_i - x)}{\sum_{|x-x_i|<h} 1} + \frac{1}{2} m''(x) \frac{\sum_{|x-x_i|<h} (x_i-x)^2}{\sum_{|x-x_i|<h} 1}$$

0

So,

$$\text{Bias}(\hat{m}(x)) = 1.2 \frac{\sum_{|x-x_i|<h} (x_i - x)}{\sum_{|x-x_i|<h} 1}$$

For $\hat{m}(0)$, we have

$$\text{Bias}(\hat{m}(0)) = 1.2 \frac{\sum_{|-x_i|<h} x_i}{\sum_{|-x_i|<h} 1}$$

Using R, we obtain

$$\text{Bias}(\hat{m}(0)) = -0.030515562$$

```
sum_x<-0
sum_1<-0
for (i in 1:100)
{
   if ( abs(-x[i])< h)
   {
      sum_x<-sum_x+x[i]
      sum_1<- sum_1+1
   }
}


bias_m<- 1.2*(sum_x/sum_1)
```

(3) (10 pts) What is the variance of $\widehat{m}(0)$? Leave your answer in terms of the unknown σ^2.

We use that

$$\mathrm{Var}(\hat{m}(x)) = \sigma^2 \sum_{i=1}^{N} w_{i,h}^2(x),$$

where

$$w_{i,h}(x) = \frac{1(|x_i - x| < h)}{\sum_{i=1}^{n} 1(|x_i - x| < h)}$$

Hence

$$\mathrm{Var}(\hat{m}(0)) = \sigma^2 \sum_{i=1}^{N} w_{i,h}^2(0)$$

with

$$w_{i,h}(0) = \frac{1\!1(|x_i| < 0.1)}{\sum_{i=1}^{N} 1\!1(|x_i| < 0.1)}$$

So using R, we have

$$\mathrm{Var}(\hat{m}(0)) \simeq 0.09090909$$

```
sum_x_ones<-0
for (i in 1:100)
{
  if ( abs(x[i])< h)
  {
    sum_x_ones<-sum_x_ones+1
  }
}

variance_factor<-0
for (i in 1:100)
{
  if ( abs(x[i])< h)
  {
    variance_factor<-variance_factor+1/sum_x_ones^2
  }
}


print(variance_factor)
```

4)

(4) (10 pts) Without actually calculating the variance, write down an expression for estimating σ^2.

We can get

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{m}(x_i))^2}{N-1}$$

5) We use

$$\hat{m}(0) \pm C_{\alpha/2} \sqrt{Var(\hat{m}(x))}, \quad \alpha/2 = 0.025$$

Using $\sigma = 0.1$

$$Var(\hat{m}(0)) = (0.1)^2 \sum_{i=1}^{N} w_i^2 h(0)$$

So, using R-code we obtain the confidence interval.

$$(-0.1702131, -0.0520228)$$

```
alpha_2<-abs(qnorm(0.025))
print(alpha_2)
sqrt_var<-sqrt(variance_factor*0.1*0.1)



print(m_hat-sqrt_var*alpha_2)
print(m_hat+sqrt_var*alpha_2)
```

Problem 2

Problem 2

Consider the nonlinear regression model, with $n = 100$ and data (x_i, y_i) provided in "HW10_data.csv" (the same dataset as Problem 1),

$$y_i = m(x_i) + \sigma\epsilon_i, \quad x_i \in (-1, 1), \quad i = 1, ..., n,$$

and the aim is to split the range of $x \in (-1, 1)$ into two intervals; the start of a regression tree. That is, we are looking for the estimator

$$\widehat{m}(x) = \begin{cases} m_1 & x \leq r \\ m_2 & x > r. \end{cases}$$

(1) (10 pts) Find the optimal choice of r which minimizes the sum of square errors. Show your working and algorithm.

1) We sort the data with respect the x-values.

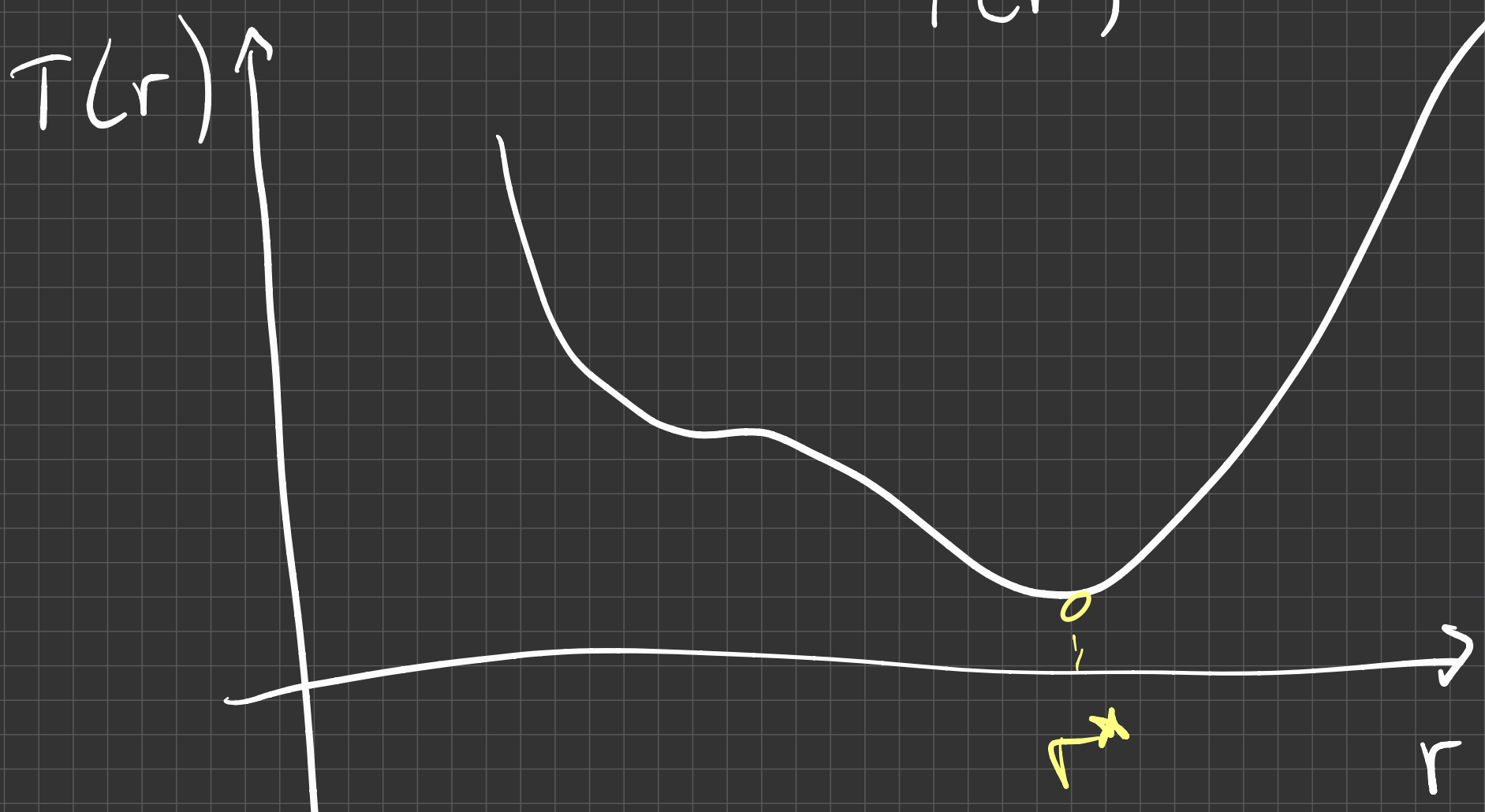
2) we create N-different 2-partitions, with $[x_i, x_{i+1}]$ midpoint r_i

3) We calculate N $\hat{m}_1$ and N $\hat{m}_2$

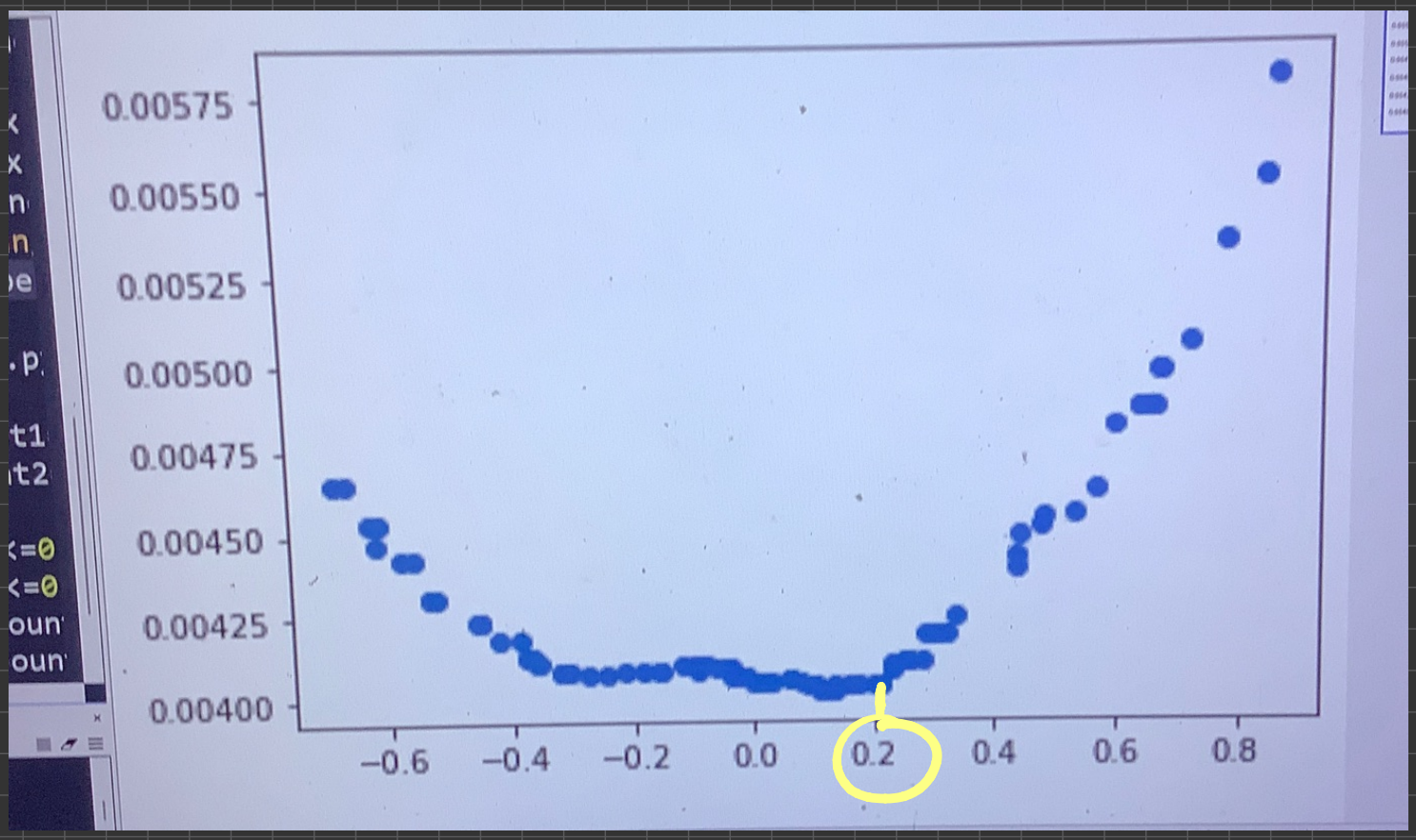
4) We see which value of $\hat{r}_k$ minimize.

$$T(r) = \sum_{x_i \le r} (y_i - \hat{\mu}_1)^2 + \sum_{x_i > r} (y_i - \hat{\mu}_2)^2$$

To find the minimum we create a plot, and visually value of r state which minimizes $T(r)$



2)



$$\hat{m}(x) = \begin{cases} -0.650014 & x \leq 0.2 \\ -0.732599 & x > 0.2 \end{cases}$$

3) $\hat{m}(0.5) = -0.732599$

4) $Var(\hat{m}(0.5)) = 0$ because

$Var(\text{constant}) = 0$

5)

Confidence interval is just

$\hat{\mu}$, because there is

no variance