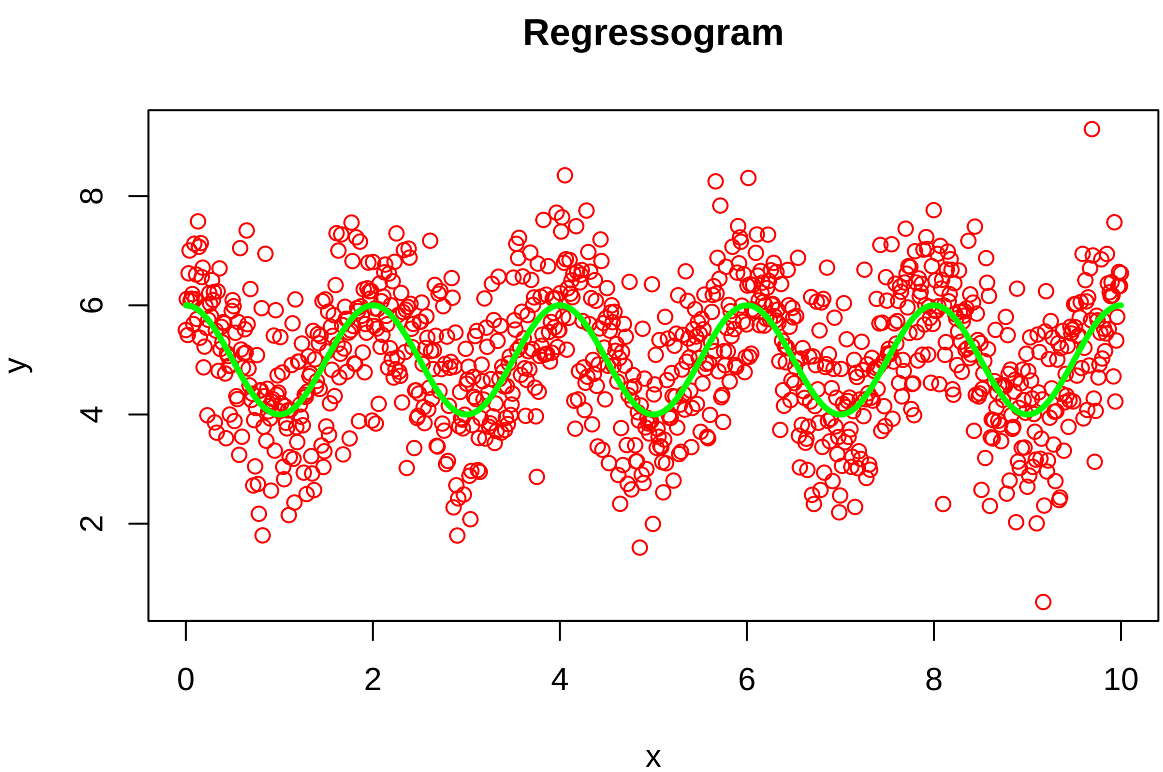
Regressogram Implementation & Analysis

Stephen J George

2024-03-12

Let us assume the function (on which we want to perform regression) as 5+cos(pi*x)+ noise and the domain to be [0,10]

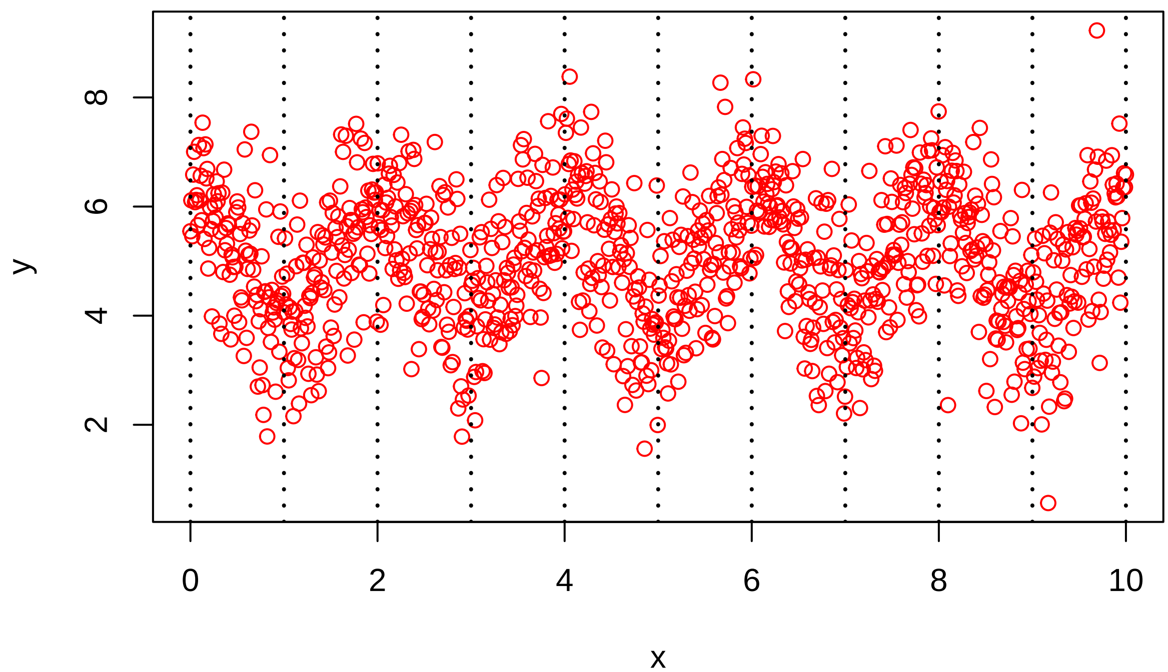
```
n = 1000
x = seq(0,10,length.out=n)
y = 5 + cos(pi*x) + rnorm(n)
#y = 5 + cos(pi*x)
plot(x,y,col="red",main="Regressogram")
curve(5+cos(pi*x),min(x),max(x),n,add=TRUE,col="green",lwd=3)
```



Let's say the number of bins we want is 10.We plot the bin boundaries as well using vertical lines,splitting the x axis into 10 different rectangles/bins.

```
m = 11
bins = 0:10
plot(x,y,col="red")
abline(v = bins,lty = 3,lwd=2)
```

```
```

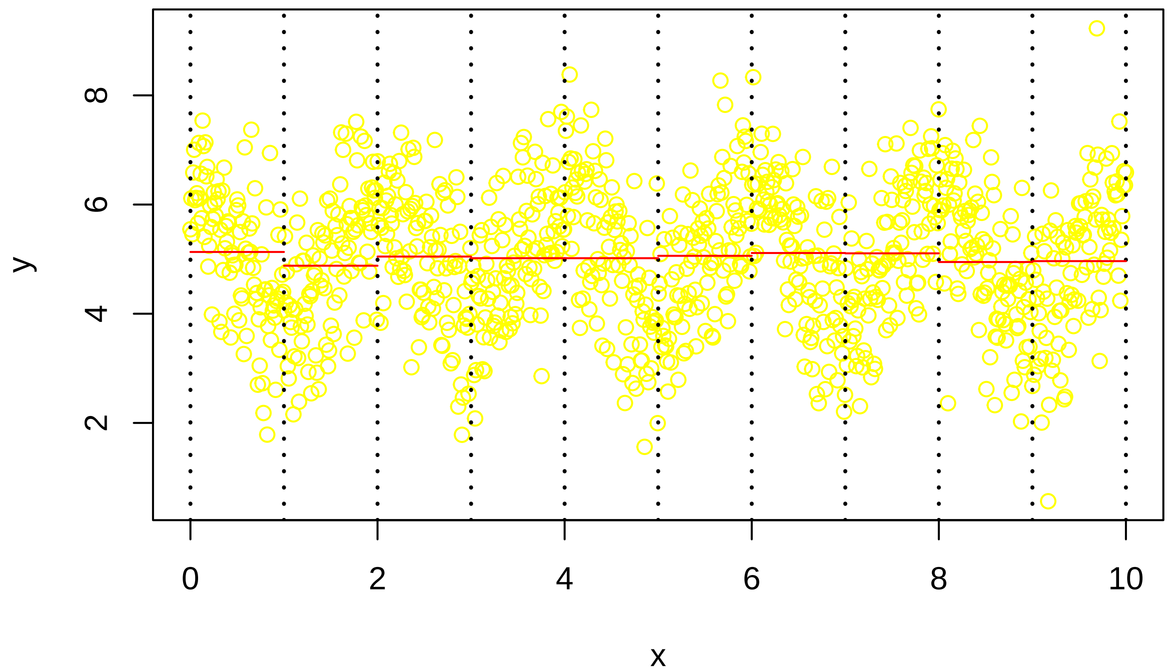


We initialize all the estimates for the m different Y's to 0.We then evaluate the estimates as the mean of all the y values in a particular bin. For the last bin,we calculate the mean separately as both the bin endpoints are included.We then plot horizontal lines,across each bin width,representing the mean calculated (i.e the y estimate).

```
y.hat=rep(0,m-1)
for (i in 1:(m-2)){
    # y.hat[i] <- mean(y[which((x<bins[i+1]) & (x>=bins[i]))])
    y.hat[i] <- mean(y[which((bins[i]<=x)&(x<bins[i+1]))])
}

#for the last bin
y.hat[m-1] = mean(y[which((x<=bins[m]) & (x>= bins[m-1]))])

plot(x,y,col="yellow")
abline(v = bins,lty = 3,lwd=2)
for (i in 1:(m-1)){
    lines(bins[i:(i+1)],rep(y.hat[i],2),col="red")
}
```



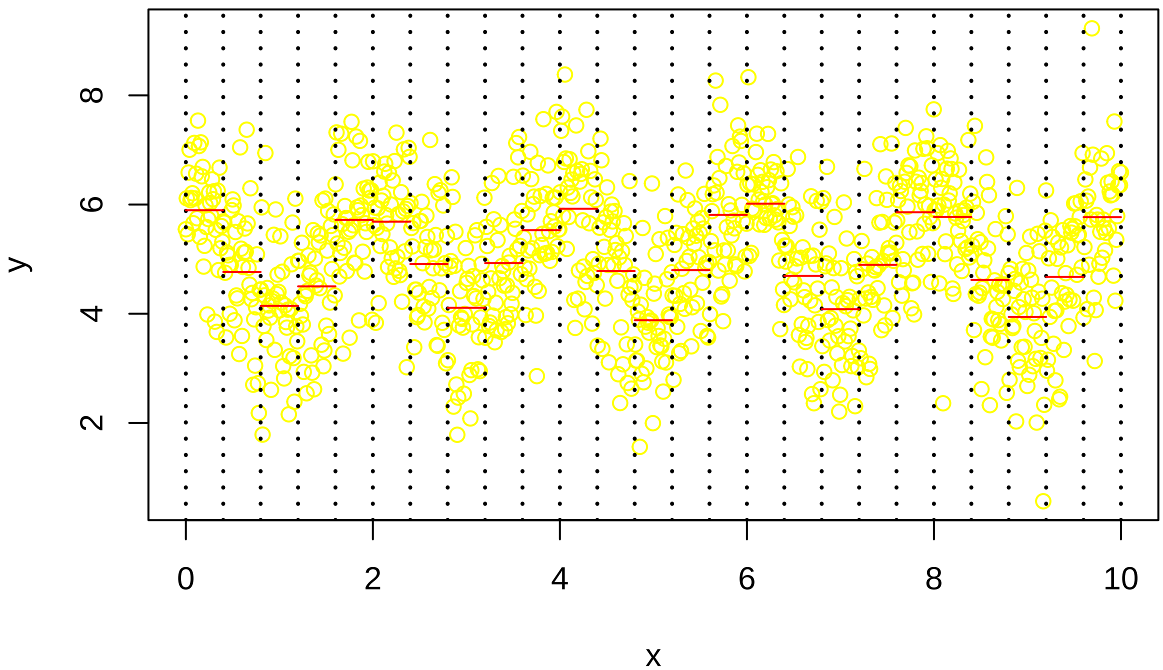
Let's now say the number of bins we want is 26 and repeat the same procedure.

```
m = 26
bins = seq(0,10,by=0.4)

y.hat=rep(0,m-1)
for (i in 1:(m-2)){
    # y.hat[i] <- mean(y[which((x<bins[i+1]) & (x>=bins[i]))])
    y.hat[i] <- mean(y[which((bins[i]<=x)&(x<bins[i+1]))])
}

#for the last bin
y.hat[m-1] = mean(y[which((x<=bins[m]) & (x>= bins[m-1]))])

plot(x,y,col="yellow")
abline(v = bins,lty = 3,lwd=2)
for (i in 1:(m-1)){
    lines(bins[i:(i+1)],rep(y.hat[i],2),col="red")
}
```



We can thus conclude that increasing the number of bins gives us a better overall fit that matches the true shape of the function (in question) in a much improved manner.The advantage is that we do this without assuming the functional form for the true original function.