STAT 513- Assignment 1

Joseph David

February 18 2022

Problem 1. *Problem 1*

(a) First note that

$$\begin{aligned} E\Big[\Big(\widehat{\theta}-\theta\Big)^2\Big] &= E\Big[\Big(\widehat{\theta}-E[\widehat{\theta}]+E[\widehat{\theta}]-\theta\Big)^2\Big] \\ &= E\Big[\Big(\widehat{\theta}-E[\widehat{\theta}]\Big)^2\Big]+2E\Big[\Big(\widehat{\theta}-E[\widehat{\theta}]\Big)\Big(E[\widehat{\theta}]-\theta\Big)\Big]+E\Big[\Big(E[\widehat{\theta}]-\theta\Big)^2\Big]. \end{aligned}$$

However, since $E[\widehat{\theta}]$ and θ are constants, we see that

$$2E\Big[\Big(\widehat{\theta}-E[\widehat{\theta}]\Big)\Big(E[\widehat{\theta}]-\theta\Big)\Big]=2\Big(E[\widehat{\theta}-E[\widehat{\theta}]\Big)\Big(E[\widehat{\theta}]-\theta\Big)=0,$$

and

$$E\Big[\Big(E[\widehat{\theta}]-\theta\Big)^2\Big]=\Big(E[\widehat{\theta}]-\theta\Big)^2.$$

We also see that by definition,

$$E\Big[\Big(\widehat{\theta}-E[\widehat{\theta}]\Big)^2\Big]=Var(\widehat{\theta}),$$

so in conclusion we have that

$$E\Big[\Big(\widehat{\theta}-\theta\Big)^2\Big]=\Big(E[\widehat{\theta}]-\theta\Big)^2+Var(\widehat{\theta}).$$

(b) Since T is nonnegative, we see that

$$E[T]=\int_{-\infty}^{\infty} tf(t)\,dt=\int_0^{\infty} tf(t)\,dt.$$

For any $a>0$, we therefore see that

$$E[T]=\int_0^a tf(t)\,dt+\int_a^{\infty} tf(t)\,dt\geq a\int_a^{\infty} f(t)\,dt=aP(T\geq a),$$

implying that

$$P(T\geq a)\leq E[T]/a \text{ for all } a>0.$$

Since T/a is nonnegative still, we apply Markov's inequality to obtain

$$P\Big(\frac{T}{a}>\frac{1}{\epsilon}\Big)\leq E\Big(\frac{T}{a}\Big)\cdot\epsilon=\frac{1}{a}E[T]\cdot\epsilon\leq\frac{1}{a}\cdot a\cdot\epsilon=\epsilon.$$

(c) We see that

$$E[\widehat{\beta}] = \beta + E(X^T X)^{-1} X^T \epsilon] = \beta + (X^T X)^{-1} E[X^T \epsilon] = \beta,$$

so $\widehat{\beta}$ is a consistent estimator of β. Note that

$$\widehat{\beta} = (X^T X)^{-1} X^T y = \beta + (X^T X)^{-1} X^T \epsilon,$$

which gives

$$\widehat{\beta} - \beta = (X^T X)^{-1} X^T \epsilon \Rightarrow \sqrt{n}(\widehat{\beta} - \beta) = \sqrt{n}\left(\frac{X^T X}{n}\right)^{-1} X^T (\epsilon/n).$$

Since $\left(\frac{X^T X}{n}\right)^{-1} \to \Sigma^{-1}$ Slutsky's Theorem implies that the limiting behavior of $\sqrt{n}(\widehat{\beta} - \beta)$ only depends on $X^T(\epsilon/n)$. We see that

$$\sqrt{n} X^T (\epsilon/n) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \epsilon_i\right),$$

but we know that $(1/n)\sum_{i=1}^{n} x_i \epsilon_i \to 0$ in probability. Since $Var(x_i\epsilon_i) = \sigma^2 x_i^T x_i$, we see that

$$\begin{aligned} Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i \epsilon_i\right) &= \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i \epsilon_i) \\ &= \frac{\sigma^2}{n} x_i^T x_i \to \sigma^2 \Sigma / n. \end{aligned}$$

Thus, by the Central Limit Theorem,

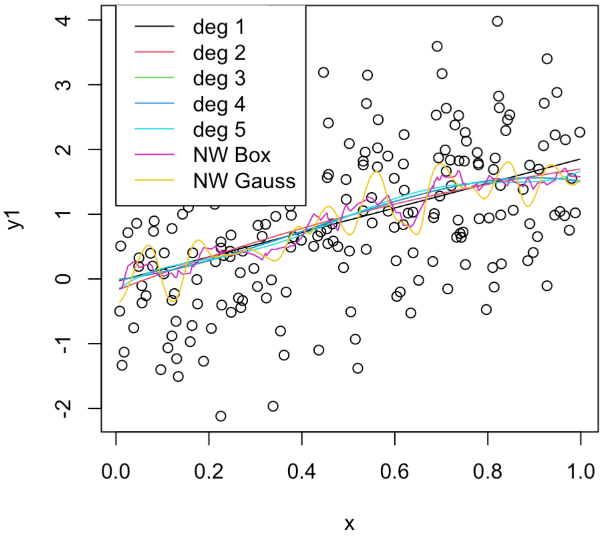
$$\begin{aligned} &\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} x_i \epsilon_i \to^d N(0, \sigma^2 \Sigma) \\ &\Rightarrow \Sigma^{-1}\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} x_i \epsilon_i \to^d N(0, \sigma^2 \Sigma^{-1}) \\ &\Rightarrow \sqrt{n}(\widehat{\beta} - \beta) \to^d N(0, \sigma^2 \Sigma^{-1}). \end{aligned}$$

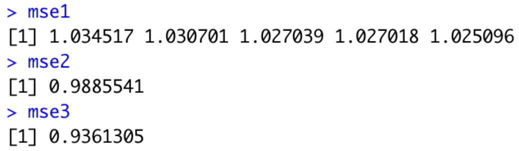
This shows that $\widehat{\beta}$ converges to β at the rate of $O(1/n)$.

Problem 2.

For each of the below, `mse1` gives the MSE using polynomials of degree 1 through 5, respectively, and `mse2` and `mse3` give the MSE for Nadaraya-Watson using box and gaussian kernel respectively.

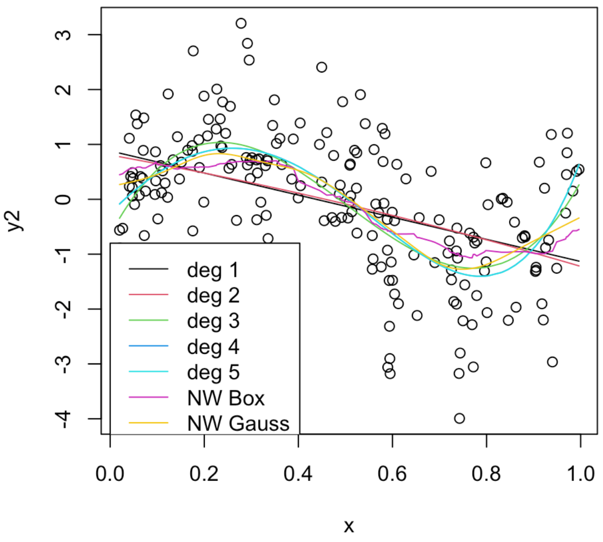
(a) We found that NW methods resulted in lower MSE and all polynomials performed similarly regradless of their degree. This is because the truth is linear and the NW regressors overfit the error in the data. The NW methods used a bandwidth of 0.05, since the function is linear we would like it to be small but we don't want to make it too small or else the errors might create bias.

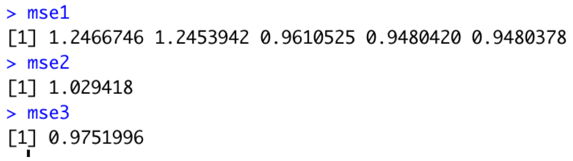




```
> mse1
[1] 1.034517 1.030701 1.027039 1.027018 1.025096
> mse2
[1] 0.9885541
> mse3
[1] 0.9361305
```

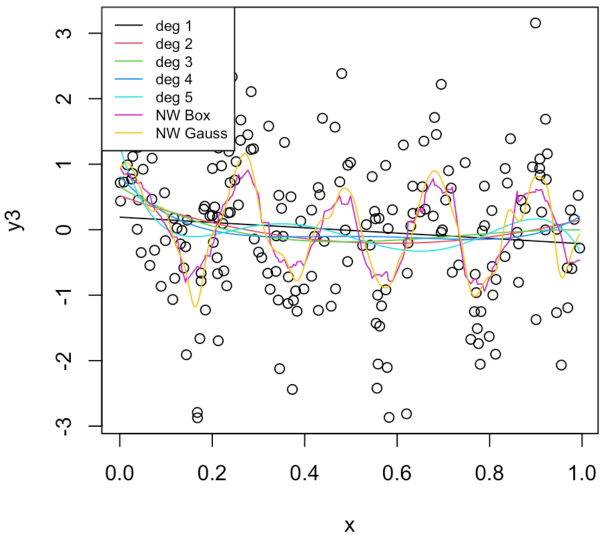
(b) Here, we used a bandwidth of 0.2 since the function has period 1. Higher polynomial degrees resulted in lower MSE, as well as NW methods.

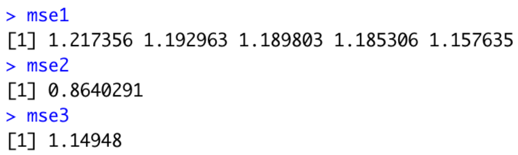




```
> mse1
[1] 1.2466746 1.2453942 0.9610525 0.9480420 0.9480378
> mse2
[1] 1.029418
> mse3
[1] 0.9751996
```

(c) Since the truth has short period, we used a bandwidth of 0.04. With such high frequency, using only up to polynomial of degree 5 did not make much of a difference. In this case, using box kernel was optimal.

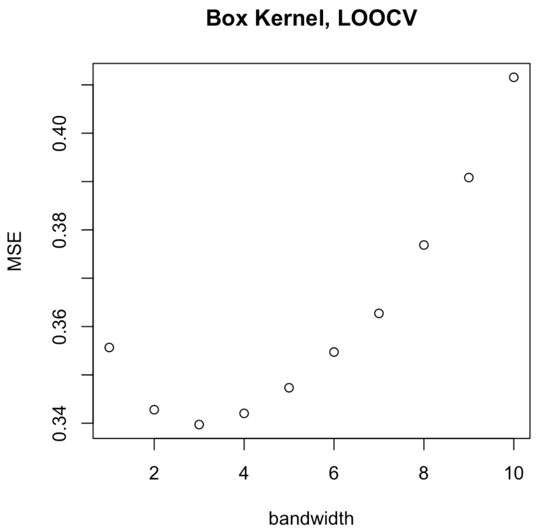


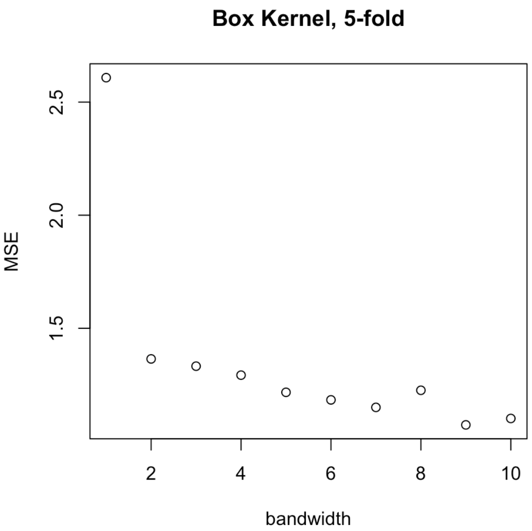


```
> mse1
[1] 1.217356 1.192963 1.189803 1.185306 1.157635
> mse2
[1] 0.8640291
> mse3
[1] 1.14948
```

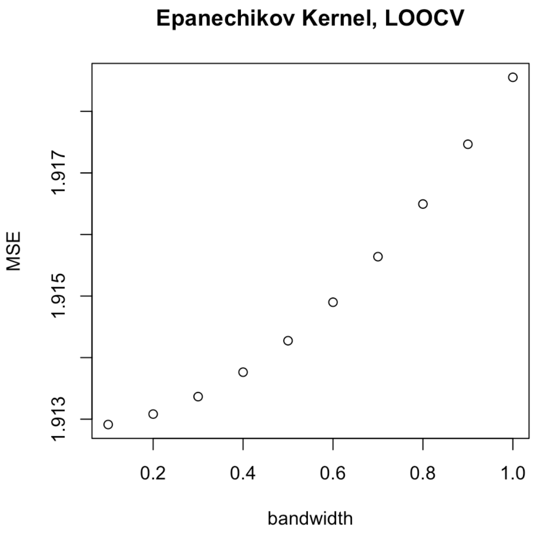
Problem 3.

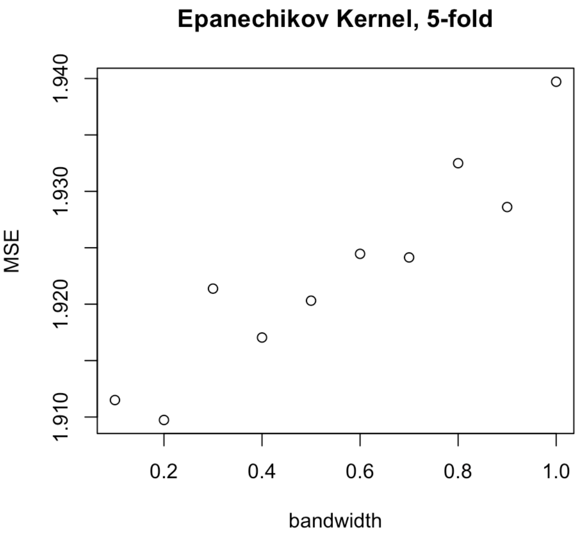
(a) We tested using bandwidths of $1, 2, \ldots, 10$. Below are the resulting MSEs using a Box Kernel with Leave One Out and 5-fold cross validation. In the LOOCV case, we found an optimal bandwidth to be 3, whereas in the 5-fold case the optimal bandwidth is 9.





(b) Here we used the same bandwidths of $0.1, 0.2, \ldots, 1$. Since the Epanechikov Kernel K_n is defined as $D(|x - x_0|/\lambda)$ where $D(t) = \frac{3}{4}(1 - t^2)$ for $|t| \leq 1$ and $D(t) = 0$ otherwise, a large bandwidth results in assigning almost equal weight to all points, thus reducing any local effects. In the case of LOOCV, it appears smaller bandwidths are better, signifying that it is best to stay very local. In 5-fold, a similar pattern emerges where smaller bandwidths improve the fit.





4 R code

#Assignment 1

```
#Problem 2
n=200
x = sort(runif(n))
y1 = 2*x + rnorm(n)
y2 = sin(2*pi*x) + rnorm(n)
y3 = sin(30*x) + rnorm(n)

#polynomial with degrees 1-5
plot(x,y3)
mse1 = numeric(5)
for(d in 1:5){
  reg = lm(y3 ~ poly(x,degree=d))
  lines(x,fitted(reg),col=d)
  mse1[d] = (1/n)*sum((fitted(reg) - y3)^2)
}

#NW box kernel
y3_nw_box = numeric(n)
box_width = .04
for(i in 1:n){  #create predictions vector
  y3_nw_box[i] = mean(y3[abs(x-x[i])<box_width])
}
mse2 = mean((y3 - y3_nw_box)^2)
lines(x,y3_nw_box,col=6)

#NW gaussian kernel
y3_nw_gauss = ksmooth(x,y3,bandwidth = 0.04,kernel='normal')
lines(y3_nw_gauss,col=7)
mse3 = mean((y3 - y3_nw_gauss$y)^2)
legend('topleft',legend=c('deg␣1', 'deg␣2', 'deg␣3', 'deg␣4', 'deg␣5', 'NW␣Box',
                          'NW␣Gauss'),col=1:7,lty=rep(1,7),cex=0.7)

#Problem 4
#Read data
fev = read.table('fev', header = TRUE, sep = "", dec = ".")
x = fev[,5]
y = fev[,4]
plot(x,y)
bandwidths = 1:10

#Box Kernel LOOCV
mse_box_loo = rep(0,length(bandwidths))
for(b in bandwidths){
  for(i in 1:length(x)){
    y_box_loo = ksmooth(x[-i],y[-i],bandwidth=b,kernel='box',x.points=c(x[i]))$y
    mse_box_loo[b] = mse_box_loo[b] + (y[i] - y_box_loo)^2
  }
}
```

```
mse_box_loo = mse_box_loo / length(x)

#Box Kernel 5-fold
mse_box_5 = rep(0,length(bandwidths))
for(b in bandwidths){
  indices = 1:length(x)
  for(i in 1:5){
    new_i = sample(indices, 130, replace=F)
    print(new_i)
    indices = setdiff(indices, new_i)
    y_box_5 = ksmooth(x[-new_i],y[-new_i],bandwidth=b,kernel='box',
                      x.points=x[new_i])$y
    mse_box_5[b] = mse_box_5[b] + mean((y[new_i] - y_box_5)^2)
  }
}
mse_box_5 = mse_box_5 / 5

#Epanechikov LOOCV
mse_ep_loo = rep(0,10)
for(b in bandwidths){
  for(i in 1:length(x)){
    y_ep_loo = sum((3/4)*(1-(abs(x[i]-x[-i])/(b/20))^2) * y[-i])/sum((3/4)*(1-(abs(
    mse_ep_loo[b] = mse_ep_loo[b] + (y[i] - y_ep_loo)^2
  }
}
mse_ep_loo = mse_ep_loo / length(x)

#Epanechikov 5-fold
mse_ep_5 = rep(0,10)
for(b in bandwidths){
  indices = 1:654
  for(i in 1:5){
    new_i = sample(indices, 130, replace=F)
    indices = setdiff(indices, new_i)
    y_ep_5 = numeric(130)
    for(j in 1:130){
      y_ep_5[j] = sum((3/4)*(1-(abs(x[new_i[j]]-x[-new_i])/(b/10))^2) * y[-new_i])/
    }
    mse_ep_5[b] = mse_ep_5[b] + mean((y[new_i] - y_ep_5)^2)
  }
}
mse_ep_5 = mse_ep_5 / 5
```