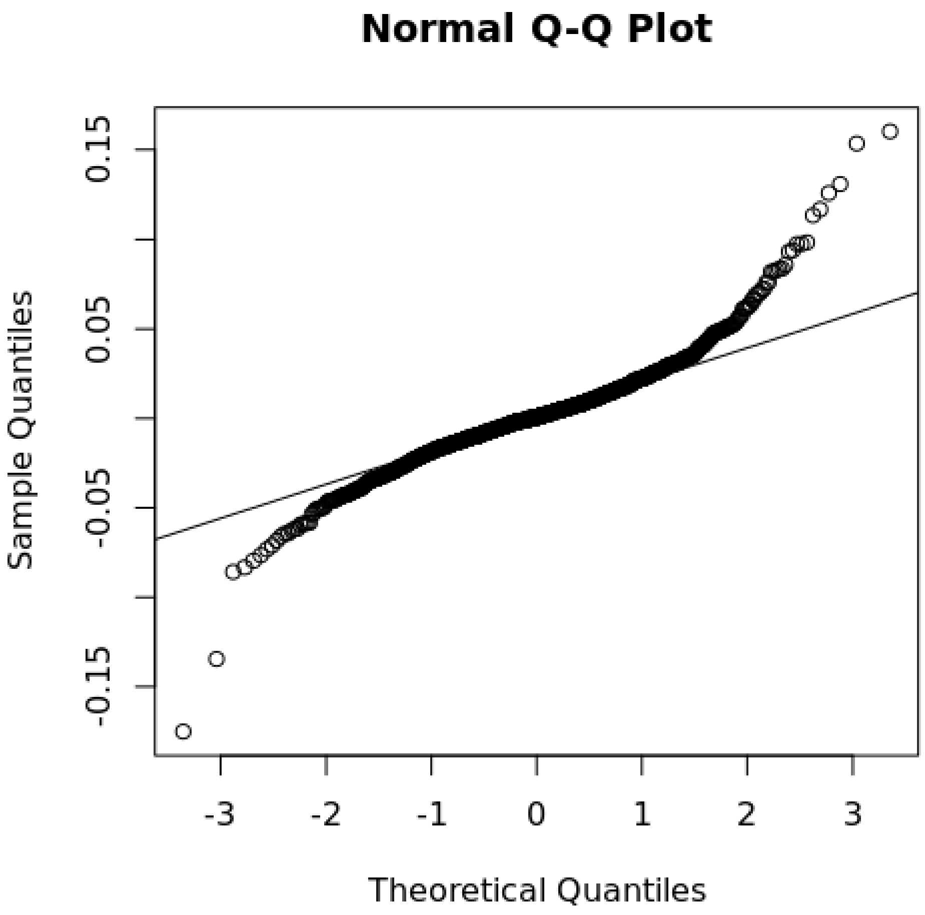
STATS 509 HW3

Author: Chongdan Pan(pandapcd)

Problem 1

(a)

```
df = read.csv("RecentFord.csv", header=TRUE)
ret = df$Adj.Close[2:nrow(df)] / df$Adj.Close[1:nrow(df)-1] - 1
qqnorm(ret)
qqline(ret)
```



I don't think the data follows a well normal distribution, because it has a much heavier tail than normal distribution.

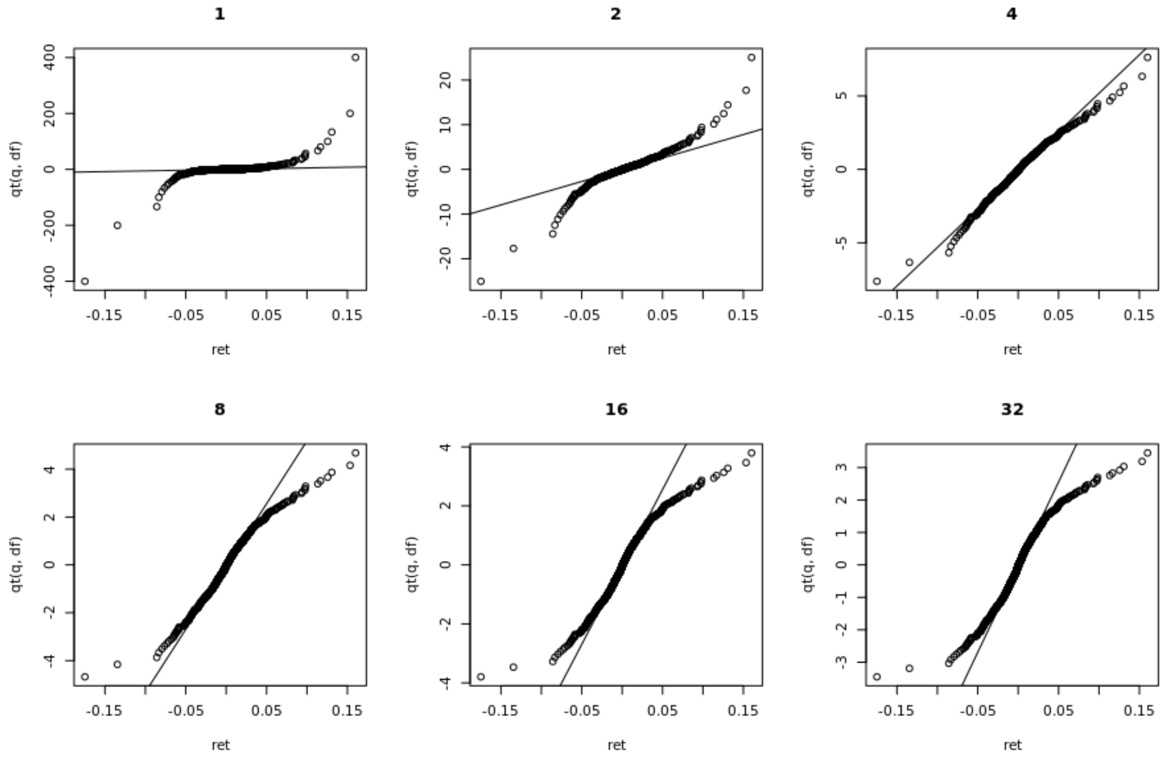
(b)

```
shapiro.test(ret)
# W = 0.95134, p-value < 2.2e-16
```

Since the P value is extremely small, I think the data doesn't follow a normal distribution

(c)

```
df = 1
par(mfrow=c(2,3))
q = (1:length(ret)) / (length(ret) + 1)
for(df in c(1, 2, 4, 8, 16, 32))
{
    qqplot(ret, qt(q, df), sub=title(df))
    qqline(ret, qt(q, df))
}
```



I set the degree of freedom to be 1, 2, 4, 8, 16, 32

I think the t-distribution with degree of freedom of 4 (button left) one works best

(d)

At first glance, the left tail and right tail are symmetric to each other, but in the plot with degree of freedom to be 2 (top middle), I think the right tail is more heavier than the left tail. When we set degree of freedom to be 4 (top right), we have a heavy right tail but a thin left tail due to the extreme value. Therefore, I conclude the extreme value may cause asymmetries.

Problem 2

(a)

$$E[\hat{f}_b(x)] = E[\frac{1}{100}\sum_{i=1}^{100} K_b(x-x_i)] = E[\int_{-\infty}^{\infty} K_b(x-x_i)f(x_i)\mathrm{d}x_i]$$

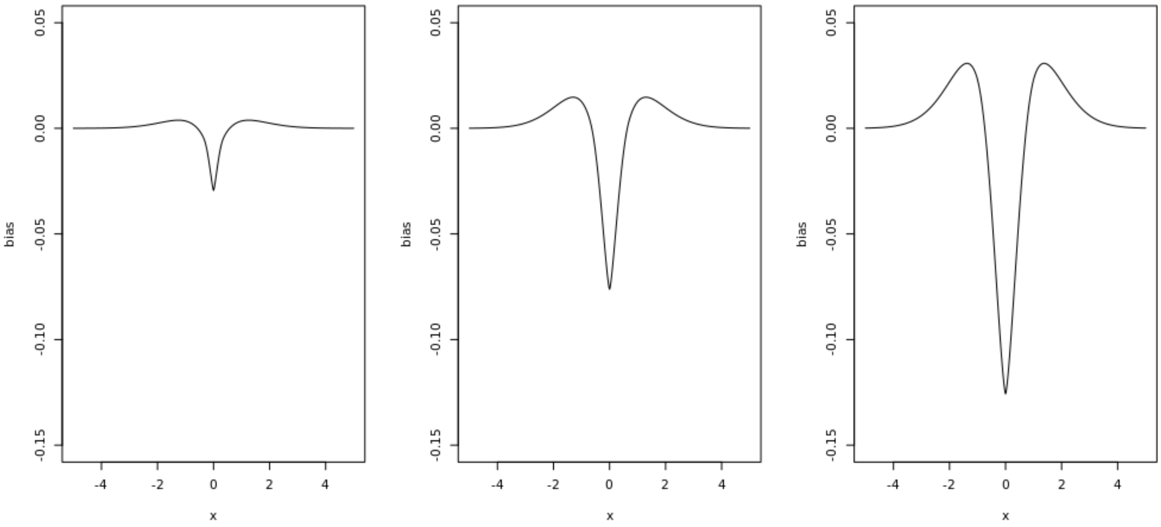
Based on our setting of K_b and $f(x)$

$$E[\hat{f}_b(x)] = \frac{1}{w}\int_{x-w/2}^{x+w/2} f(x_i)\mathrm{d}x_i = \frac{1}{w}(F_{ged,1.5}(x+w/2) - F_{ged,1.5}(x-w/2))$$

(b)

The bias should be
$E[\hat{f}_b(x)] - f(x) = \frac{1}{w}[F_{ged,1.5}(x+w/2) - F_{ged,1.5}(x-w/2)] - f_{ged,1.5}(x)$

```
par(mfrow=c(1,3))
x = seq(-5, 5, 0.01)
for(b in c(0.2, 0.4, 0.6))
{
    w = b * 3.464
    bias = (pged(x+w/2, 0, 1, 1.5) - pged(x-w/2, 0, 1, 1.5)) / w - dged(x, 0, 1,
1.5)
    plot(x, bias, type="l", ylim=c(-0.15, 0.05), sub=title(b))
}
```



(c)

$$Var[\hat{f}_b(x)] = \frac{1}{100}Var[K_b(x-x_i)] = \frac{1}{100}[E[(K_b(x-x_i))^2] - (E[K_b(x-x_i)])^2]$$

For $E[K_b(x-x_i)]^2$, from (a) we have $\frac{1}{w^2}[F_{ged,1.5}(x+w/2) - F_{ged,1.5}(x-w/2)]^2$

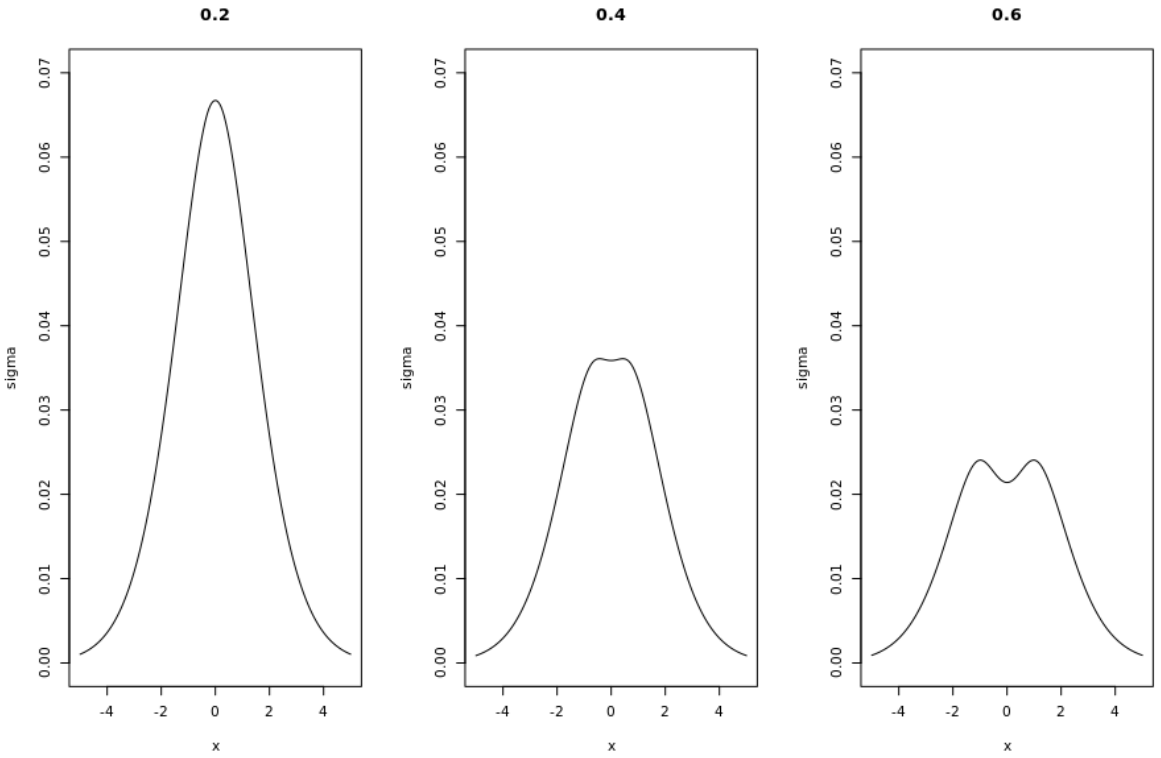
For $E[(K_b(x-x_i))^2]$, we have

$$\int_{-\infty}^{\infty}[K_b(x-x_i)]^2 f(x_i)\mathrm{d}x_i = \frac{1}{w^2}(F_{ged,1.5}(x+w/2) - F_{ged,1.5}(x-w/2))$$

Then for the standard deviation, we have

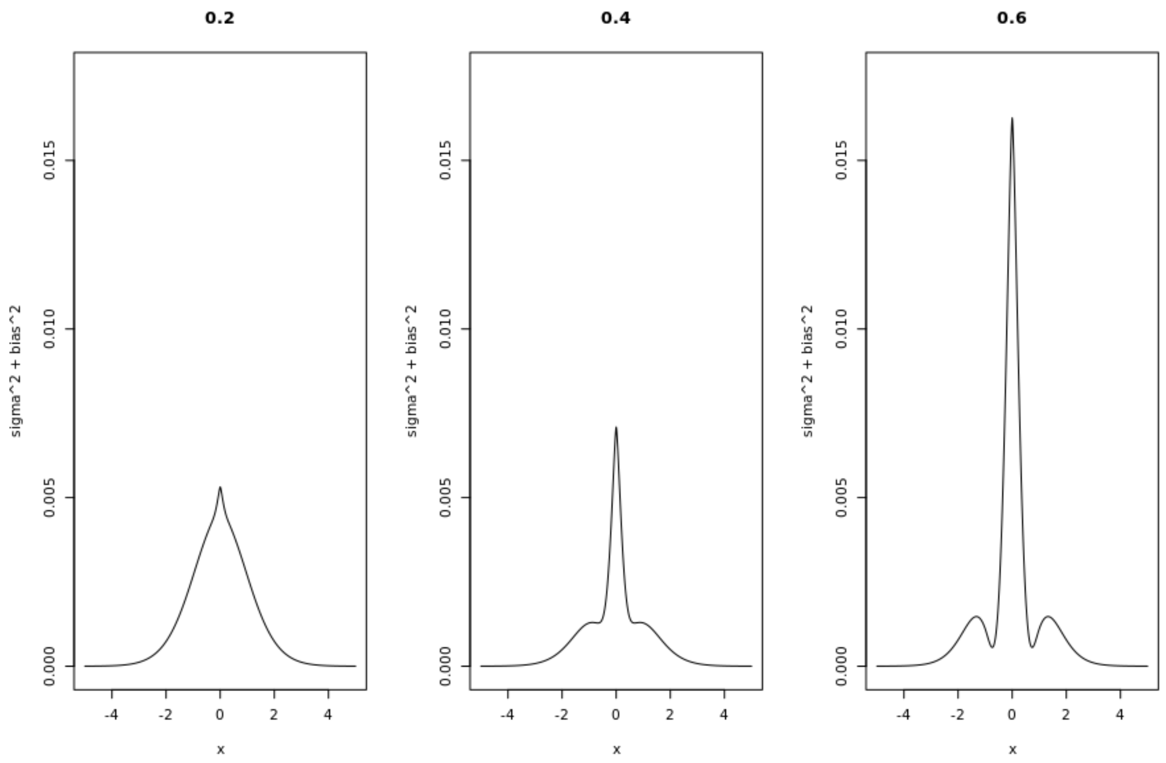
$$\sigma = \frac{1}{10w}\sqrt{[F_{ged,1.5}(x+w/2) - F_{ged,1.5}(x-w/2)] - [F_{ged,1.5}(x+w/2) - F_{ged,1.5}(x-w/2)]^2}$$

```
par(mfrow=c(1,3))
x = seq(-5, 5, 0.01)
for(b in c(0.2, 0.4, 0.6))
{
    w = b * 3.464
    lambda = 1
    sigma = (pged(x+w/2, 0, lambda, 1.5) - pged(x-w/2, 0, lambda, 1.5) -
(pged(x+w/2, 0, 1, 1.5) - pged(x-w/2, 0, 1, 1.5))^2) ^ 0.5 / (10 * w)
    plot(x, sigma, type="l", ylim=c(0, 0.07), sub=title(b))
}
```



(d)

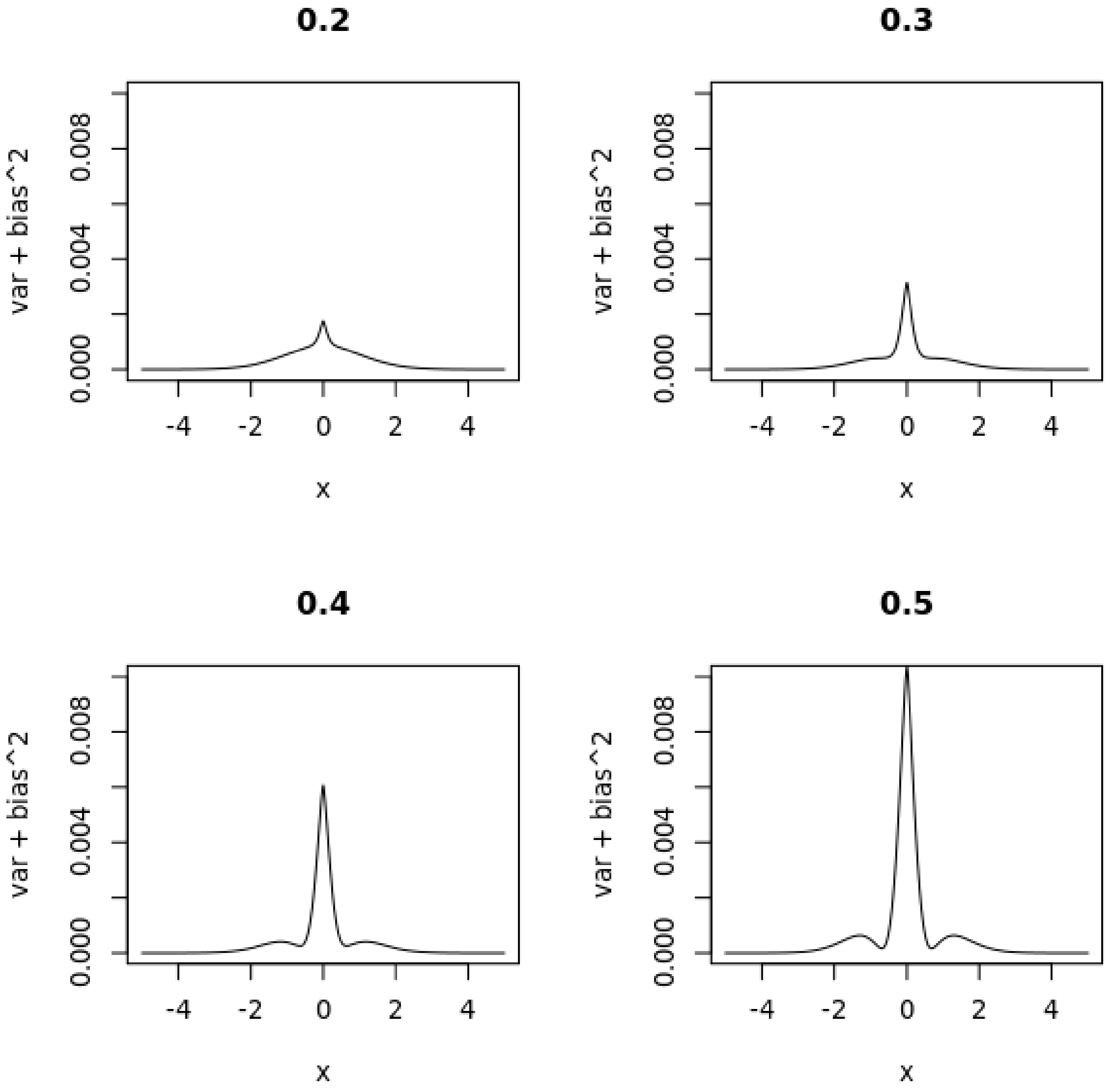
```
par(mfrow=c(1,3))
x = seq(-5, 5, 0.01)
for(b in c(0.2, 0.4, 0.6))
{
    w = b * 3.464
    lambda = 1
    sigma = (pged(x+w/2, 0, lambda, 1.5) - pged(x-w/2, 0, lambda, 1.5) -
(pged(x+w/2, 0, 1, 1.5) - pged(x-w/2, 0, 1, 1.5))^2) ^ 0.5 / (10 * w)
    bias = (pged(x+w/2, 0, 1, 1.5) - pged(x-w/2, 0, 1, 1.5)) / w - dged(x, 0, 1,
1.5)
    plot(x, sigma ^ 2 + bias ^ 2, type="l", ylim=c(0, 0.0175), sub=title(b))
}
```



Based on the plot, I think $b = 0.4$, since it gives us a lower maximum MSE than 0.6 as well as higher kurtosis than $b = 0.2$

(e)

```
par(mfrow=c(2,2))
x = seq(-5, 5, 0.01)
n = 500
for(b in c(0.2, 0.3, 0.4, 0.5))
{
    w = b * 3.464
    lambda = 1
    var = (pged(x+w/2, 0, lambda, 1.5) - pged(x-w/2, 0, lambda, 1.5) -
(pged(x+w/2, 0, 1, 1.5) - pged(x-w/2, 0, 1, 1.5))^2) / (n * w ^ 2)
    bias = (pged(x+w/2, 0, 1, 1.5) - pged(x-w/2, 0, 1, 1.5)) / w - dged(x, 0, 1,
1.5)
    plot(x, var + bias ^ 2, type="l", ylim=c(0, 0.01), sub=title(b))
}
```



Yes, because when the MSE will be much higher if we keep setting $b = 4$. Based on the new plots when we have 500 samples, I will decrease the bandwidth to be 0.3 to have lower MSE as well as kurtosis. Since, we have more samples, there is less probability of overfitting, which means we can use a lower bandwidth.