lab-02-additional-exercises

Exercise 1

1) Variance:

a) We know that k-nearest regression is defined as:

$$[\hat{f}(x) = \frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} y_i]$$

b) To find the variance:

$$Var(\hat{f}(x)) = Var\frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} y_i$$

=

$$Var(\hat{f}(x)) = \frac{1}{k^2} Var(\sum_{x_i \in \mathcal{N}(x)} y_i)$$

=

$$Var(\hat{f}(x)) = \frac{1}{k^2} \sum_{x_i \in \mathcal{N}(x)} Var(y_i)$$

=

$$Var(\hat{f}(x)) = \frac{1}{k^2} \sum_{x_i \in \mathcal{N}(x)} (\sigma^2)$$

=

$$Var(\hat{f}(x)) = \frac{1}{k^2} k(\sigma^2)$$

=

$$Var(\hat{f}(x)) = \frac{k}{k^2} (\sigma^2)$$

=

$$Var(\hat{f}(x)) = \frac{\sigma^2}{k}$$

2) Bias

a) We know that k-nearest regression is defined as:

$$[\hat{f}(x) = \frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} y_i]$$

b)

$$[Bias(\hat{f}(x))]^2 = [f(x) - E[\hat{f}(x)]]^2$$

=

$$[Bias(\hat{f}(x))]^2 = [f(x) - E[\frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} y_i]^2$$

=

$$[Bias(\hat{f}(x))]^2 = [f(x) - [\frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} E(y_i)]^2$$

=

$$[Bias(\hat{f}(x))]^2 = [f(x) - [\frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} f(x_i)]^2$$

2) Bias

a) We know that k-nearest regression is defined as:

$$[\hat{f}(x) = \frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} y_i]$$

b)

$$[Bias(\hat{f}(x))]^2 = [f(x) - E[\hat{f}(x)]]^2$$

=

$$[Bias(\hat{f}(x))]^2 = [f(x) - E[\frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} y_i]^2$$

=

$$[Bias(\hat{f}(x))]^2 = [f(x) - [\frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} E(y_i)]^2$$

=

$$[Bias(\hat{f}(x))]^2 = [f(x) - [\frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} f(x_i)]^2$$

c) Decomposition of MSE:

$$MSE = Var(f(x)) + Bias(\hat{f}(x))^2 + Var(\hat{f}x)$$

=

$$MSE = Var(f(x)) + [f(x) - [\frac{1}{k} \sum_{x_i \in \mathcal{N}(x)} f(x_i)]^2 + \frac{\sigma^2}{k}$$

Exercise 2

```
library(tidyverse)
```

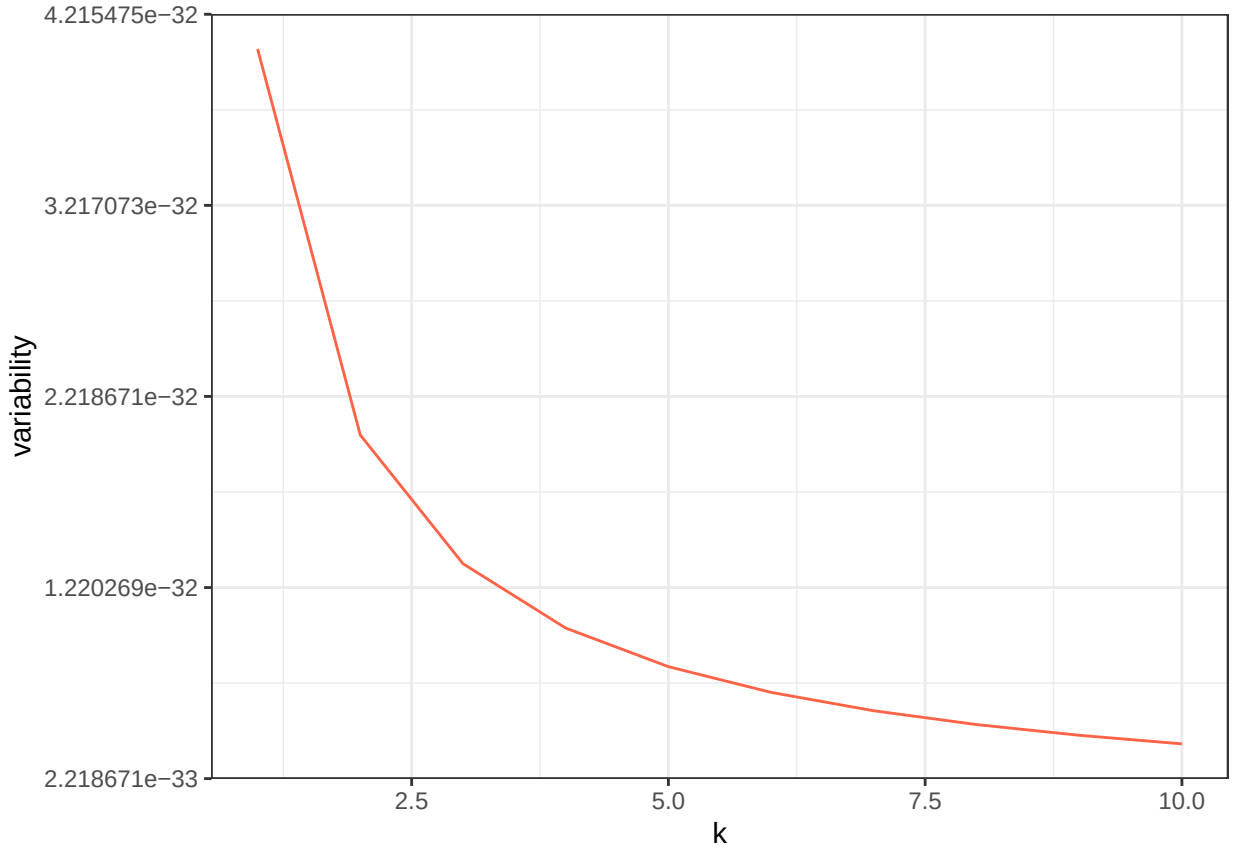
```
## -- Attaching packages --------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

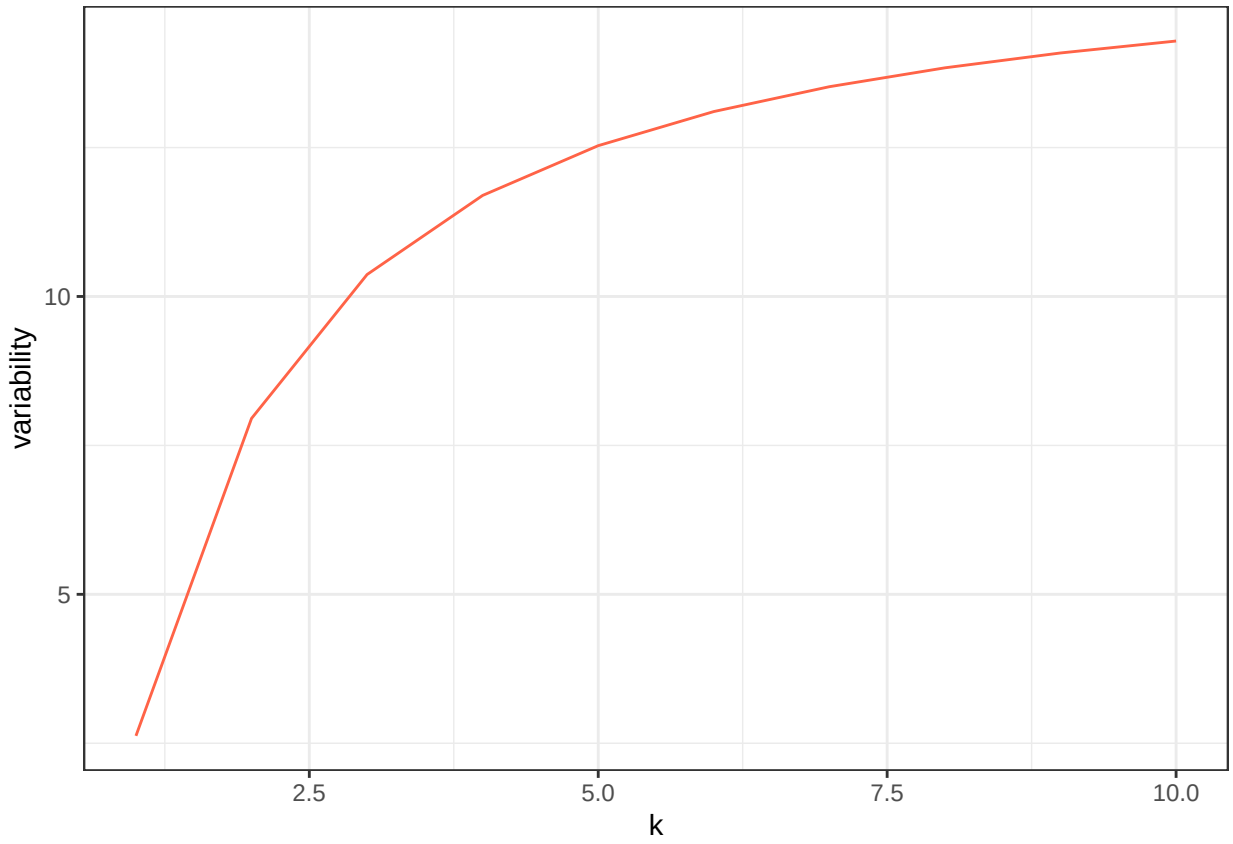
```
## -- Conflicts ------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
x <- c(1:3, 5:12)
y <- c(-7.1, -7.1, .5, -3.6, -2, -1.7, -4, -.2, -1.2, -1.2, -3.5)
y_mean = mean(y)

std_dev <- function(y) {(sum(y-y_mean)^2)/11}
std_dev <- std_dev(y)



var_fun <- function(k, x, y) {
  v_k <- rep(NA, length(k))
  for (i in 1:length(k)) {
    v_k[i] <- std_dev/k[i]
  }
  v_k
```

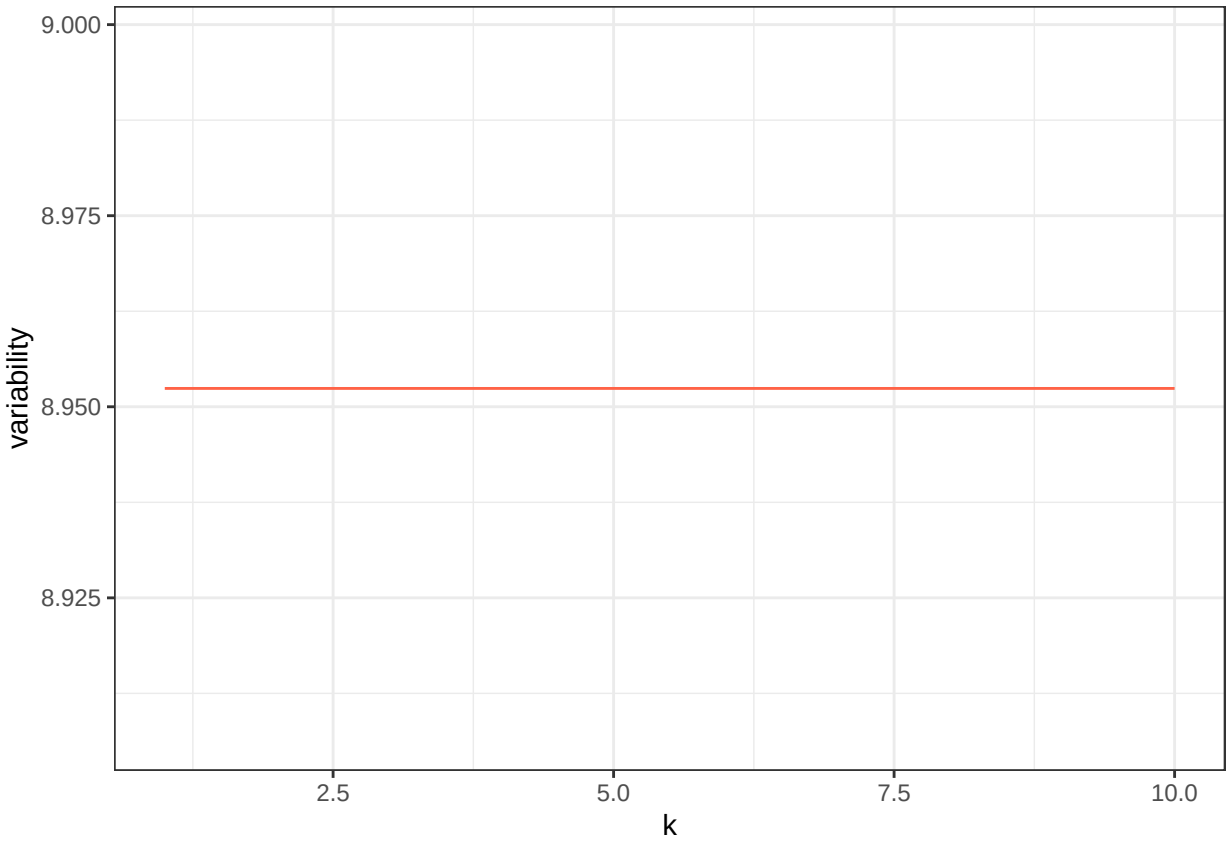
```r
}

f <- function(x) {
  f = -9.3 + 2.6 * x - 0.3 * x^2 + .01 * x^3
}
bias_fun <- function(k, x, y) {
  b_k <- rep(NA, length(k))
  for (i in 1:length(k)) {
    b_k[i] <- (-4.02 - (1/k[i]*-2.4))^2
  }
  b_k
}


MSE_fun <- function(k, x, y) {
  m_k <- rep(NA, length(k))
  for (i in 1:length(k)) {
    m_k[i] <- 7.9524 + 1 + std_dev/k[i]
  }
  m_k
}

k <- 1:10
b_k <- bias_fun(k, x, y)

k <- 1:10
m_k <- MSE_fun(k, x, y)



k <- 1:10
v_k <- var_fun(k, x, y)

df <- tibble(k = k, v_k = v_k)
ggplot(df, aes(x = k, y = v_k)) + geom_line (col = "tomato") +
  theme_bw() +
  ylab("variability")
```




```
df <- tibble(k = k, b_k = b_k)
ggplot(df, aes(x = k, y = b_k)) + geom_line (col = "tomato") +
  theme_bw() +
  ylab("variability")
```




```
df <- tibble(k = k, m_k = m_k)
ggplot(df, aes(x = k, y = m_k)) + geom_line (col = "tomato") +
  theme_bw() +
  ylab("variability")
```


9.000
8.975
8.950
8.925
variability
2.5
5.0
7.5
10.0
k