PSTAT 231 Homework 1

Blaine Quackenbush

10/6/2021

```
# Taken from assignment
algae <- read_table(
  "algaeBloom.txt",
  col_names=c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4',
              'oPO4','PO4','Chla','a1','a2','a3','a4','a5','a6','a7'),
  na="XXXXXXX") # What does this do?
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   season = col_character(),
##   size = col_character(),
##   speed = col_character(),
##   mxPH = col_double(),
##   mnO2 = col_double(),
##   Cl = col_double(),
##   NO3 = col_double(),
##   NH4 = col_double(),
##   oPO4 = col_double(),
##   PO4 = col_double(),
##   Chla = col_double(),
##   a1 = col_double(),
##   a2 = col_double(),
##   a3 = col_double(),
##   a4 = col_double(),
##   a5 = col_double(),
##   a6 = col_double(),
##   a7 = col_double()
## )
```

```
glimpse(algae)
```

```
## Rows: 200
## Columns: 18
## $ season <chr> "winter", "spring", "autumn", "spring", "autumn", "winter", "su~
## $ size   <chr> "small", "small", "small", "small", "small", "small", "small", ~
## $ speed  <chr> "medium", "medium", "medium", "medium", "medium", "high", "high~
## $ mxPH   <dbl> 8.00, 8.35, 8.10, 8.07, 8.06, 8.25, 8.15, 8.05, 8.70, 7.93, 7.7~
## $ mnO2   <dbl> 9.8, 8.0, 11.4, 4.8, 9.0, 13.1, 10.3, 10.6, 3.4, 9.9, 10.2, 11.~
## $ Cl     <dbl> 60.80, 57.75, 40.02, 77.36, 55.35, 65.75, 73.25, 59.07, 21.95, ~
## $ NO3    <dbl> 6.238, 1.288, 5.330, 2.302, 10.416, 9.248, 1.535, 4.990, 0.886,~
```

```
## $ NH4    <dbl> 578.00, 370.00, 346.67, 98.18, 233.70, 430.00, 110.00, 205.67, ~
## $ oPO4   <dbl> 105.00, 428.75, 125.67, 61.18, 58.22, 18.25, 61.25, 44.67, 36.3~
## $ PO4    <dbl> 170.00, 558.75, 187.06, 138.70, 97.58, 56.67, 111.75, 77.43, 71~
## $ Chla   <dbl> 50.000, 1.300, 15.600, 1.400, 10.500, 28.400, 3.200, 6.900, 5.5~
## $ a1     <dbl> 0.0, 1.4, 3.3, 3.1, 9.2, 15.1, 2.4, 18.2, 25.4, 17.0, 16.6, 32.~
## $ a2     <dbl> 0.0, 7.6, 53.6, 41.0, 2.9, 14.6, 1.2, 1.6, 5.4, 0.0, 0.0, 0.0, ~
## $ a3     <dbl> 0.0, 4.8, 1.9, 18.9, 7.5, 1.4, 3.2, 0.0, 2.5, 0.0, 0.0, 0.0, 2.~
## $ a4     <dbl> 0.0, 1.9, 0.0, 0.0, 0.0, 0.0, 3.9, 0.0, 0.0, 2.9, 0.0, 0.0, 0.0~
## $ a5     <dbl> 34.2, 6.7, 0.0, 1.4, 7.5, 22.5, 5.8, 5.5, 0.0, 0.0, 1.2, 0.0, 1~
## $ a6     <dbl> 8.3, 0.0, 0.0, 0.0, 4.1, 12.6, 6.8, 8.7, 0.0, 0.0, 0.0, 0.0, 0.~
## $ a7     <dbl> 0.0, 2.1, 9.7, 1.4, 1.0, 2.9, 0.0, 0.0, 0.0, 1.7, 6.0, 1.5, 2.1~
```

1. Descriptive Summary Statistics. Given the lack of further information on the problem domain, it is wise to investigate some of the statistical properties of the data, so as to get a better grasp of the problem. It is always a good idea to start our analysis with some kind of exploratory data analysis. A first idea of the statistical properties of the data can be obtained through a summary of its descriptive statistics.

(a) Count the number of observations in each size using summarise() in dplyr

```
#Count number of observations by size
algae %>%
  group_by(size) %>%
  summarise(n())
```

```
## # A tibble: 3 x 2
##   size   `n()`
##   <chr>  <int>
## 1 large     45
## 2 medium    84
## 3 small     71
```

(b) Are there missing values? Calculate the mean and variance of each chemical (ignore a_1 through a_7). What do you notice about the magnitude of the two quantities for different chemicals?

Yes, there are missing values as we can see below:

```
#Retrieve missing values, summarise by column
algae %>% select(everything()) %>% summarise_all(funs(sum(is.na(.))))
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
## # A tibble: 1 x 18
##   season  size speed  mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla    a1    a2
##    <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
## 1      0     0     0     1     2    10     2     2     2     2    12     0     0
## # ... with 5 more variables: a3 <int>, a4 <int>, a5 <int>, a6 <int>, a7 <int>
```

Also, below are the means and variances of each column below:

```
#Retrieve mean of all numerical columns except for a1 through a7
algae %>% select(mxPH:Chla) %>% summarise_if(is.numeric, mean, na.rm = TRUE)
```

```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  8.01  9.12  43.6  3.28  501.  73.6  138.  14.0
```

```
#Retrieve mean of all numerical columns except for a1 through a7
algae %>% select(mxPH:Chla) %>% summarise_if(is.numeric, var, na.rm = TRUE)
```

```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3      NH4  oPO4    PO4  Chla
##   <dbl> <dbl> <dbl> <dbl>    <dbl> <dbl>  <dbl> <dbl>
## 1 0.358  5.72 2193.  14.3 3851585. 8306. 16639.  420.
```

Upon inspection of the means, we can see that the means of chloride, orthophosphate, ammonium, and phosphate are the highest, with the latter two being significantly higher than the others. That being said, chloride, ammonium, orthophosphate, and phosphate also have significantly higher variances than any of the other chemicals. Thus, it could be the case that there are a few bodies of water that are skewing both the means and variances for these chemicals.

(c) Compute the median and MAD of each chemical and compare the two sets of quantities. What do you notice?

```
#Retrieve median of all numerical columns except for a1 through a7
algae %>% select(mxPH:Chla) %>% summarise_if(is.numeric, median, na.rm = TRUE)
```

```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  8.06   9.8  32.7  2.68  103.  40.2  103.  5.48
```

```
#Retrieve MAD of all numerical columns except for a1 through a7
algae %>% select(mxPH:Chla) %>% summarise_if(is.numeric, mad, na.rm = TRUE)
```

```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.504  2.05  33.2  2.17  112.  44.0  122.  6.67
```

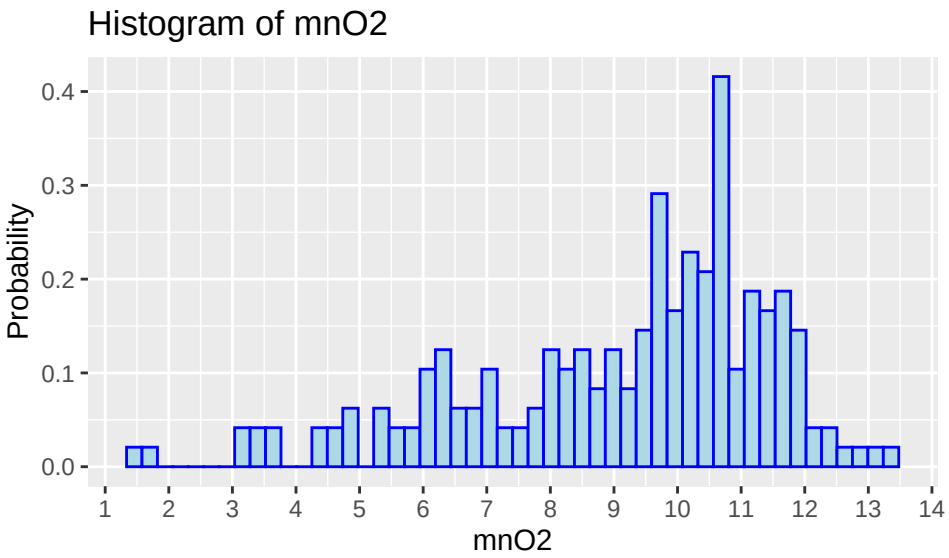
All of the values above are much smaller than and seem more reasonable for the four chemicals mentioned above. The other chemicals haven't seen too much of a difference between mean/variance versus median/MAD. This shows that there must be some outliers for chloride, ammonium, orthophosphate, and phosphate that are significantly increasing their means and variances.

2. Data Visualization. Most of the time, the information in the data set is also well captured graphically. Histogram, scatter plot, boxplot, Q-Q plot are frequently used tools for data visualization. Use ggplot for all of these visualizations.

(a) Produce a histogram of mnO2 with the title 'Histogram of mnO2' based on algae data set. Use an appropriate argument to show the probability instead of the frequency as the vertical axis. (Hint: look at the examples in the help file geom_histogram()) Is the distribution skewed?

```
ggplot(algae, aes(x = mnO2)) + #Specify mnO2 column should be x axis
  #Plot histogram, change count to density
  geom_histogram(aes(y = ..density..), color = 'blue', fill = 'lightblue', bins = 50) +
  labs(title = 'Histogram of mnO2', x = 'mnO2', y = 'Probability')+ #Label axes and title
  scale_x_continuous(breaks = seq(1, 14, 1)) #Increase ticks on x axis
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



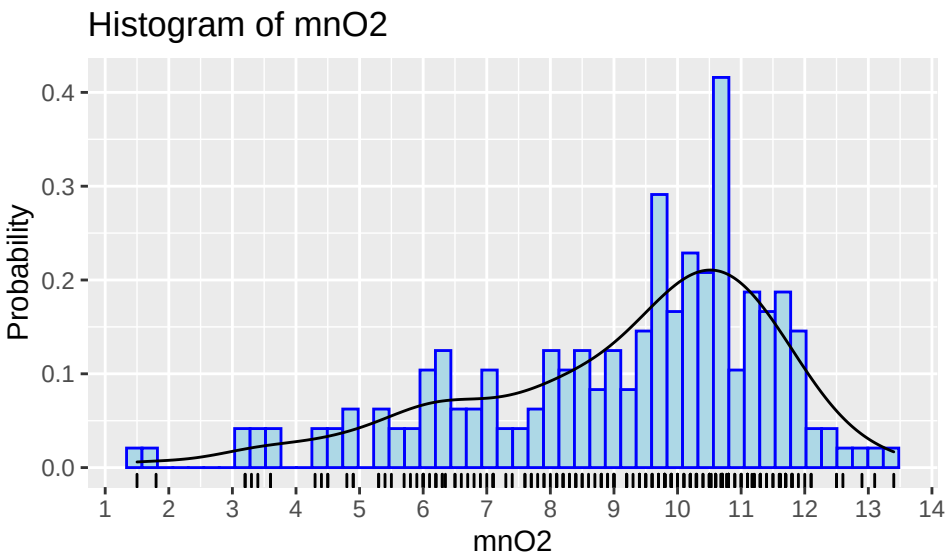
The distribution is certainly skewed, as the range of mnO2 values ranges between ~1.5 to ~13.5, but the median and means are 9.12 and 9.8, respectively. Also, we can see that the probability density is much higher for values between 8 and 12.5 than values below 8 and above 12.5.

(b) Add a density curve using geom_density() and rug plots using geom_rug() to above histogram.

```
ggplot(algae, aes(x = mnO2)) + #Specify mnO2 column as x axis
  #Plot histogram, change count to density
  geom_histogram(aes(y = ..density..), color = 'blue', fill = 'lightblue', bins = 50) +
  labs(title = 'Histogram of mnO2', x = 'mnO2', y = 'Probability') + #Labels
  geom_density(aes(y = ..density..)) + #Add density curve
  geom_rug() + #Add rug plot
  scale_x_continuous(breaks = seq(1, 14, 1)) #Increase tick marks on x axis
```

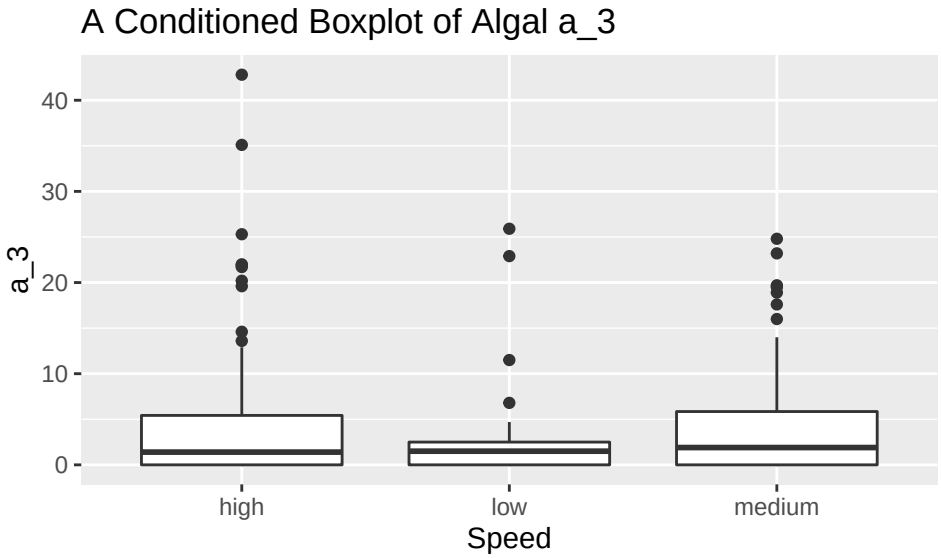
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```



(c) Create a boxplot with the title 'A conditioned Boxplot of Algal a_3' for a_3 grouped by speed. What do you notice?

```
ggplot(algae, aes(x = speed)) + #Specify boxes by speed
  geom_boxplot(aes(y = a3)) + #Use a3 column
  labs(title = 'A Conditioned Boxplot of Algal a_3', x = 'Speed', y = 'a_3') #Labels
```



The majority of the rivers seem to have lower amounts of algal a_3, although it seems that rivers that have a faster flow can have higher concentrations of a_3 than slow moving rivers.

3. Dealing with missing values.

(a) How many observations contain missing values? How many missing values are there in each variable?

```
sum(is.na(algae))
```

```
## [1] 33
```

```
#Summarize number of missing values by column
algae %>% select(everything()) %>% summarise_all(funs(sum(is.na(.))))
```

```
## # A tibble: 1 x 18
##   season  size speed  mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla    a1    a2
##    <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
## 1      0     0     0     1     2    10     2     2     2     2    12     0     0
## # ... with 5 more variables: a3 <int>, a4 <int>, a5 <int>, a6 <int>, a7 <int>
```

As we can see above, there are 33 total missing values, and they are categozied by variable as well.

(b) Removing observations with missing values: use filter() function in dplyr package to observations with any missing value, and save the resulting dataset without missing values as algae.del. Report how many observations are in algae.del.

```
algae.del <- algae %>% filter(complete.cases(.)) #New data frame with no missing values
algae.del %>% summarise(., n()) #Summarise new data frame
```

```
## # A tibble: 1 x 1
##   `n()`
##   <int>
## 1   184
```

As we can see, there are 184 observations without missing data.

4. In lecture we present the bias-variance tradeoff that takes the form

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \mathrm{Var}(\hat{f}(x_0)) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \mathrm{Var}(\epsilon)$$

where the underlying model $Y = f(X) + \epsilon$ satisfies: (1) ϵ is a zero-mean random noise, and X is non-random (all randomness in Y comes from ϵ); (2) (x_0, y_0) is a test observation, independent of the training set, and drawn from the same model; (3) $\hat{f}(\cdot)$ is the estimate of f obtained from the training set.

(a) Which of the term(s) in the bias-variance tradeoff above represent the reducible error? Which of the term(s) represent the irreducible error?

Since we have defined the following:

$$\mathrm{Bias}(\hat{f}(x_0)) := \mathbb{E}\left[\hat{f}(x_0)\right] - f(x_0),$$

and

$$\mathrm{Var}(\hat{f}(x_0)) := \mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}\hat{f}(x_0)\right)^2\right],$$

as well as our assumption that the only randomness from Y comes from ϵ, we can see that the reducible error arises from the bias and variance of $\hat{f}(x_0)$ and the irreducible error comes from the variance of ϵ.

(b) Use the bias-variance tradeoff above to show that the expected test error is always at least as large as the irreducible error.

Using that $y_0 = f(x_0) + \epsilon$, we can expand the expected test error:

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \mathbb{E}\left[(f(x_0))^2) - 2f(x_0)\hat{f}(x_0) + (\hat{f}(x_0))^2 + 2\epsilon f(x_0) - 2\epsilon\hat{f}(x_0) + \epsilon^2\right]$$

$$= \mathbb{E}\left[(f(x_0) - \hat{f}(x_0))^2\right] + 2\mathbb{E}[\epsilon(f(x_0) - \hat{f}(x_0))] + \mathbb{E}[\epsilon^2].$$

Clearly, we can minimize this error in the case that $f(x_0) = \hat{f}(x_0)$, and so we have

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] \geq \mathbb{E}\left[(f(x_0) - f(x_0))^2\right] + 2\mathbb{E}[\epsilon(f(x_0) - f(x_0))] + \mathbb{E}[\epsilon^2] = \mathbb{E}[\epsilon^2]$$

and so our lower bound for the expected test error will always be $\mathbb{E}[\epsilon^2] = \text{Var}(\epsilon)$, or the irreducible error.

5. Prove the bias-variance tradeoff where $\text{Bias}(\hat{f}(x_0)) = \mathbb{E}[\hat{f}(x_0)] - f(x_0)$.

Proof. Expanding the expected test error, where we assume $y_0 = f(x_0) + \epsilon$, we get

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \mathbb{E}\left[(f(x_0))^2) - 2f(x_0)\hat{f}(x_0) + (\hat{f}(x_0))^2 + 2\epsilon f(x_0) - 2\epsilon\hat{f}(x_0) + \epsilon^2\right]$$

$$= \mathbb{E}\left[(f(x_0) - \hat{f}(x_0))^2\right] + 2\mathbb{E}[\epsilon(f(x_0) - \hat{f}(x_0))] + \mathbb{E}[\epsilon^2]$$

Above, we have the first piece of the equation shown, and this is $\mathbb{E}[\epsilon^2] = \text{Var}(\epsilon)$, since we assume the mean of ϵ is zero. Note that by our assumptions on ϵ, we have that ϵ is independent of $f(x_0) - \hat{f}(x_0)$, and so we have

$$\mathbb{E}[\epsilon(f(x_0) - \hat{f}(x_0))] = \mathbb{E}[f(x_0) - \hat{f}(x_0)]\mathbb{E}[\epsilon] = 0$$

since $\mathbb{E}[\epsilon] = 0$, again by our assumptions on ϵ. Now we have

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \mathbb{E}\left[(f(x_0) - \hat{f}(x_0))^2\right] + \text{Var}(\epsilon).$$

Now let us expand the first term on the right hand side above:

$$\mathbb{E}\left[(f(x_0) - \hat{f}(x_0))^2\right] = \mathbb{E}\left[((f(x_0) - \mathbb{E}[\hat{f}(x_0)]) + (\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2\right]$$

$$= \mathbb{E}\left[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2 + 2(f(x_0) - \mathbb{E}[\hat{f}(x_0)])(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)) + (\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2\right]$$

$$= \mathbb{E}\left[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2\right] + \mathbb{E}\left[2(f(x_0) - \mathbb{E}[\hat{f}(x_0)])(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))\right] + \mathbb{E}\left[(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2\right].$$

$$= \mathbb{E}\left[(\mathbb{E}[\hat{f}(x_0)] - (f(x_0))^2\right] + \mathbb{E}\left[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])^2\right] + \mathbb{E}\left[2(f(x_0) - \mathbb{E}[\hat{f}(x_0)])(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))\right].$$

Now, we can note that that $\text{Bias}(\hat{f}(x_0)) = \mathbb{E}\left[\hat{f}(x_0)\right] - f(x_0)$ is a constant, and so taking the expectation of this value does not change. This gives us

$$= \left[\text{Bias}(\hat{f}(x_0))\right]^2 + \text{Var}(\hat{f}(x_0) + \mathbb{E}\left[2(f(x_0) - \mathbb{E}[\hat{f}(x_0)])(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))\right]$$

Now let us turn our focus to the term $\mathbb{E}\left[2(f(x_0) - \mathbb{E}[\hat{f}(x_0)])(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))\right]$. Note that $f(x_0) - \mathbb{E}(\hat{f}(x_0))$ is simply a number, and so we have

$$\mathbb{E}\left[2(f(x_0) - \mathbb{E}[\hat{f}(x_0)])(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))\right] = 2(f(x_0) - \mathbb{E}(\hat{f}(x_0)))\mathbb{E}\left[\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right]$$

$$= 2(f(x_0) - \mathbb{E}(\hat{f}(x_0))) \left(\mathbb{E}[\hat{f}(x_0)] - \mathbb{E}[\mathbb{E}[\hat{f}(x_0)]]\right)$$
$$= 2(f(x_0) - \mathbb{E}(\hat{f}(x_0))) \left(\mathbb{E}[\hat{f}(x_0)] - \mathbb{E}[\hat{f}(x_0)]\right) = 0.$$

Therefore, putting this all together, we have

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \text{Var}(\hat{f}(x_0)) + \left[\text{Bias}(\hat{f}(x_0))\right]^2 + \text{Var}(\epsilon).$$

This concludes our proof.

6. Distance metrics are a very important concept used in KNN. Here x, y are p-dimensional vectors. Show that the follwoing measures are distance metrics by showing the above properties hold:

(a) $d(x, y) = \|x - y\|_2$ For all three cases let $x = (x_1, \ldots, x_p), y = (y_1, \ldots y_p), z = (z_1, \ldots, z_p)$.

Positivity: Note that we have $\|x - y\|_2 = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$. For each $1 \leq i \leq p$, we have that $(x_i - y_i)^2 \geq 0$ and will be identically 0 if and only if $x_i = y_i$. Therefore, we must have that $\|x - y\| \geq 0$ since it is the square root of a sum of nonnegative terms, and can only be identically zero if $x_i = y_i$ for all i, namely $x = y$.

Symmetry: Note that $(x_i - y_i)^2 = (y_i - x_i)^2$ for each $1 \leq i \leq p$, and so $\sum_{i=1}^p (x_i - y_i)^2 = \sum_{i=1}^p (y_i - x_i)^2$, implying that $\|x - y\| = \|y - x\|$.

Triangle inequality: For each i let $u_i = x_i - y_i$ and $v_i = y_i - z_i$. Then letting $u = (u_1, \ldots, u_p), v = (v_1, \ldots, v_p)$ we have

$$\|x - z\|^2 = \|u + v\|^2 = \sum_{i=1}^p (u_i + v_i)^2$$
$$= \sum_{i=1}^p u_i^2 + \sum_{i=1}^p v_i^2 + 2\sum_{i=1}^p u_i v_i$$

Let $(u, v) = \sum_{i=1}^p u_i v_i$ denote the standard inner product. Then we have $\|u + v\| = \|u\|^2 + 2(u, v) + \|v\|^2$. Finally, using the Cauchy-Schwarz inequality $(u, v) \leq \|u\|\|v\|$, we have

$$\|x - z\|^2 = \|u + v\|^2 = \|u\| + 2(u, v) + \|v\|^2 \leq \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 \leq (\|u\| + \|v\|)^2$$

Taking the square root of both sides and using that $u = x - y, v = y - z$ we have

$$\|x - z\| \leq \|x - y\| + \|y - z\|,$$

or $d(x, z) \leq d(x, y) + d(y, z)$.

(b) $d(x, y) = \|x - y\|_\infty$. For all three cases let $x = (x_1, \ldots, x_p), y = (y_1, \ldots y_p), z = (z_1, \ldots, z_p)$.

Positivity: Note that $\|x - y\| = \max_{1 \leq i \leq p} |x_i - y_i|$. Thus, we must have by definition of absolute value that $d(x, y) \geq 0$ and further, $d(x, y) = 0$ if and only if $|x_i - y_i| = 0$ for all i, in which case $x = y$.

Symmetry: This follows from the property of absolute value, namely that $|x_i - y_i| = |y_i - x_i|$ for any real numbers x_i, y_i and so $d(x, y) = d(y, x)$.

Triangle inequality: WThis follows from the classic triangle inequality $|x_i - z_i| \leq |x_i - y_i| + |y_i - z_i|$. Namely for each $1 \leq i \leq p$ we have

$$|x_i - z_i| \leq |x_i - y_i| + |y_i - z_i| \leq \max_{1 \leq i \leq p} |x_i - y_i| + \max_{1 \leq i \leq p} |y_i - z_i| = d(x, y) + d(y, z).$$

Thereofore, $d(x, z) \leq d(x, y) + d(y, z)$.