# Homework 1

PSTAT131-231

**Note:** If you are working in a group, please submit only one homework per group as a PDF on Gradescope and specify all group members at the time of submission (not just on the PDF).

---

## Predicting Algae Blooms

**Background** High concentrations of harmful algae in rivers can have serious impacts on both ecological communities and water quality. Monitoring and forecasting of algae blooms is essential to effective ecological stewardship.

**Objective** Your goal will be to model the relationship between algae levels and water chemistry and use the model to predict algal blooms.

**Data** Water samples were collected from European rivers at different times during an initial study period and again during a follow-up period. Multiple samples were collected at several physical locations during each sampling event. For each sample, concentrations of seven chemicals were measured along with the levels of seven harmful algae. Sampling conditions were also recorded: the season; the river size, and the river speed. In total, 200 sampling events were conducted in this initial study.

The data were aggregated by sampling event. The chemical profiles recorded for each sampling event comprise: maximum pH (`max_pH`); minimum oxygen (`min_O2`); mean chlorine (`Cl`); mean nitrate (`NO3`); mean ammonium (`NH4`); mean orthophosphate (`oPO4`); mean total phosphates (`PO4`); and mean chlorophyll (`Chla`). The algae measured are simply named `a1` through `a7`.

## Part 1. Exploration

It is wise to investigate some of the statistical properties of the data to get a better grasp of the problem. It is always a good idea to start an analysis with some kind of data exploration using descriptive statistics and visualizations. Begin by importing the training data as shown below, and then complete the following exploratory steps.

```
# import training data from initial study
load('data/algae.RData')
```

### Question 1 (a)

Make a table showing the number of observations in each season. Do the data seem balanced across seasons?

| season | n |
|---|---|
| autumn | 40 |
| spring | 53 |
| summer | 45 |
| winter | 62 |

**Answer** The data seems fairly balanced across seasons, although winter has much more observations in comparison to the other seasons.

### Question 1 (b)

Are there missing values? Make a table showing the proportion of missing values for each variable with missingness.

Table 2: Table continues below

| season | size | speed | max_pH | min_O2 | Cl | NO3 | NH4 | oPO4 | PO4 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0.005 | 0.01 | 0.05 | 0.01 | 0.01 | 0.01 | 0.01 |

| Chla | a1 | a2 | a3 | a4 | a5 | a6 | a7 |
|---|---|---|---|---|---|---|---|
| 0.06 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Answer** Yes, there are missing values in the following categories: max_pH, min_O2, Cl, NO3, NH4, oPO4, PO4, and Chla.

### Question 1 (c)

The simplest approach to handling missing values is to simply drop observations with missing values. Remove all observations with missing values and display the 50th observation in the resulting dataset (*hint*: use `na.omit()` and `slice()`). Describe a situation when this approach could be problematic.

```
## # A tibble: 1 x 18
##   season size  speed max_pH min_O2    Cl   NO3   NH4  oPO4   PO4  Chla    a1
##   <chr>  <chr> <chr>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 winter small high     7.9     11  6.17  1.17  18.3  7.75  11.8   0.5  81.9
## # ... with 6 more variables: a2 <dbl>, a3 <dbl>, a4 <dbl>, a5 <dbl>, a6 <dbl>,
## #   a7 <dbl>
```

**Answer** This approach could be problematic if our results are changed when we do not ignore the missing values. Moreover, we could have omitted important rows that might have included important information for the data analysis, which would skew the end conclusion.

### Question 1 (d)

Another approach is to fill in the missing values with the column mean or median. Fill in the missing values for the chemical variables using the median of the corresponding column. (*Hints*: use `mutate(across(...))` and see the documentation for `replace_na()`.) Store the result as `algae_imp`. Display the values of each chemical for observations 48 and 62 after performing the imputation. Use `algae_imp` for all subsequent questions.

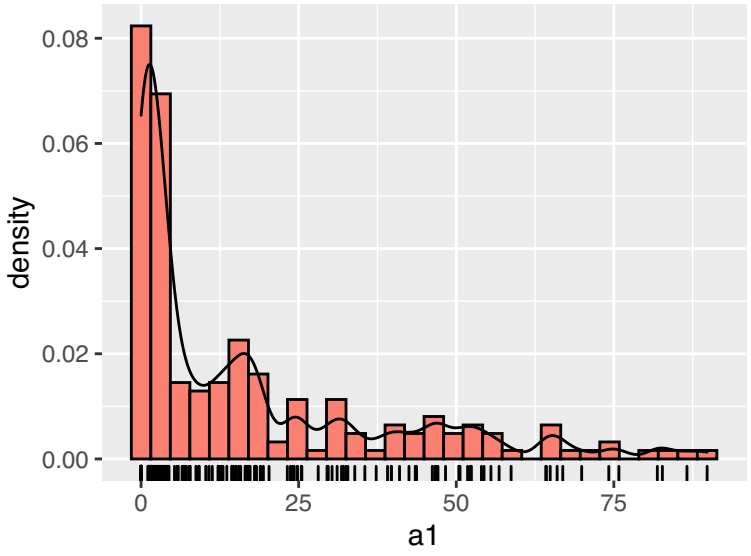
Table 4: Table continues below

| season | size | speed | max_pH | min_O2 | Cl | NO3 | NH4 | oPO4 |
|---|---|---|---|---|---|---|---|---|
| winter | small | low | 8.06 | 12.6 | 9 | 0.23 | 10 | 5 |
| summer | small | medium | 6.4 | 9.8 | 32.73 | 2.675 | 103.2 | 40.15 |

| PO4 | Chla | a1 | a2 | a3 | a4 | a5 | a6 | a7 |
|---|---|---|---|---|---|---|---|---|
| 6 | 1.1 | 35.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 5.475 | 19.4 | 0 | 0 | 2 | 0 | 3.9 | 1.7 |

### Question 1 (e)

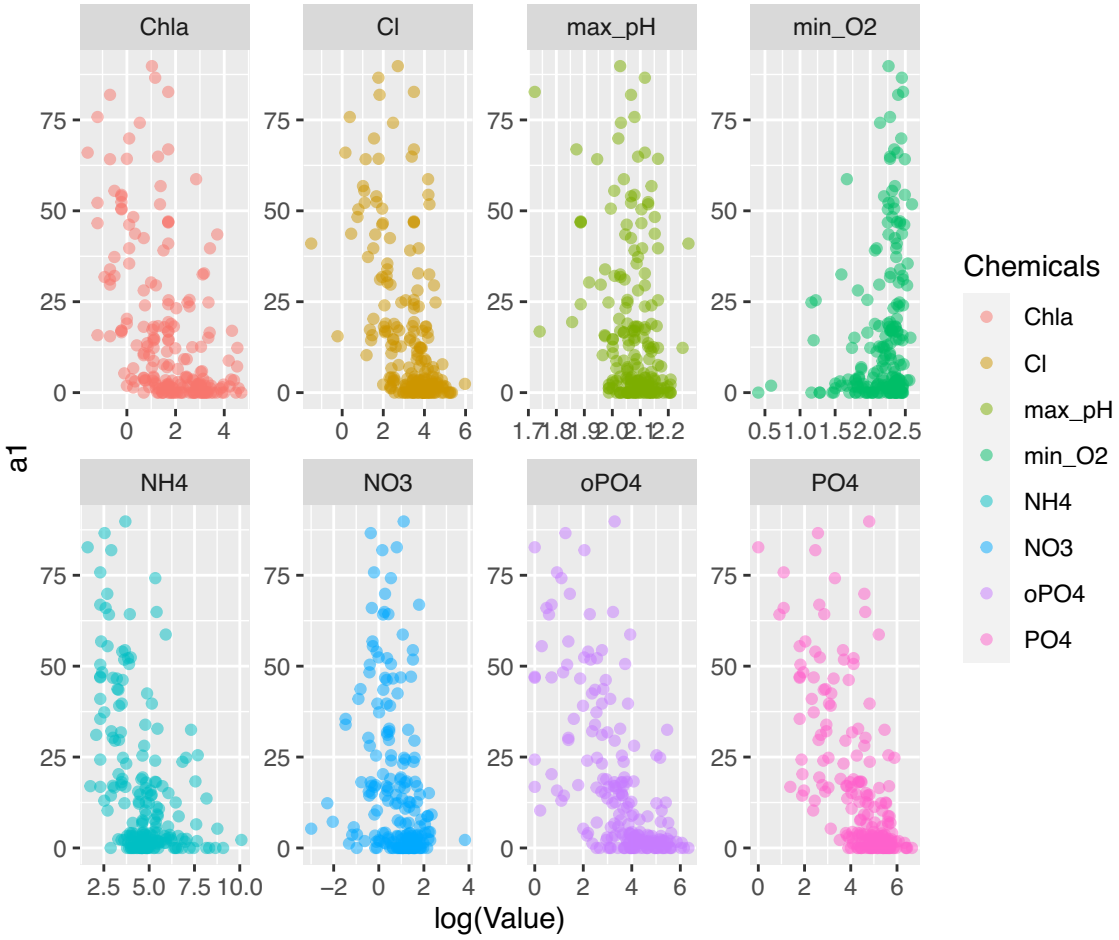
We're going to focus on algae 1. Construct a histogram of algae 1 levels with a superimposed kernel density estimate, a rug plot underneath the horizontal axis, and appropriate axis labels. (Hint: ensure the histogram is on the density scale, not the count scale, and use `geom_density()` and `geom_rug()`; careful about your choice of smoothing parameter (`bw = ...`) in the kernel density estimate). Is the distribution skewed?




**Answer** From the histogram, we can see that the distribution skewed to the right.

**Question 1 (f)**

Since you're going to predict algae 1 levels using the chemical variables, it is helpful to check this relationship visually to see what kind of model structure is sensible. Construct a 2x4 panel of scatterplots of algae 1 levels against the log of each chemical variable (hint: apply the log transformation within the `aes()` arguments; there is no need to mutate the dataset). Do any of the relationships look approximately linear on the transformed scale?




**Answer** When we have a free scale, none of the relationships look approximately linear on the transformed scale.

## Part 2. Prediction

Here you'll fit a linear model to the data with missing values imputed using the median and estimate the prediction error.

### Question 2 (a)

Log transform all chemical variables via $\log(x + 1)$, store the result, and select just the seasonal and (transformed) chemical variables and algae 1 levels. Save this as a new tibble `algae_trans`. Then split `algae_trans` into 80% training and 20% test partitions. Please be sure to use the rng seed provided in the code chunk. (*Hint*: `resample_partition()`.)

```
# for reproducibility
set.seed(32921)

# select variables for regression and log transform
algae_trans <- algae_imp %>%
    mutate(across(max_pH:Chla, ~log(.x + 1))) %>%
    select(season,max_pH:Chla, a1, size, speed)

# training and test partitions
algae_trans_part <- resample_partition(data = algae_trans, p = c(test = 0.2, train = 0.8))
```

### Question 2 (b)

Regress algae 1 levels on the remaining variables using the training partition. Display a table of the coefficient estimates. Compute the square root of the training MSE. How well does the model seem to fit, and how do you know?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | 69.25 | 56.02 | 1.236 | 0.2183 |
| **seasonspring** | -0.9517 | 4.431 | -0.2148 | 0.8302 |
| **seasonsummer** | 0.5182 | 4.151 | 0.1249 | 0.9008 |
| **seasonwinter** | 1.941 | 4.058 | 0.4783 | 0.6331 |
| **max_pH** | -1.525 | 25.38 | -0.06012 | 0.9521 |
| **min_O2** | -3.346 | 5.927 | -0.5645 | 0.5733 |
| **Cl** | -2.734 | 1.993 | -1.372 | 0.1723 |
| **NO3** | -0.04581 | 3.153 | -0.01453 | 0.9884 |
| **NH4** | -1.313 | 1.211 | -1.085 | 0.2799 |
| **oPO4** | -4.97 | 2.608 | -1.906 | 0.05865 |
| **PO4** | -2.148 | 3.032 | -0.7083 | 0.4799 |
| **Chla** | -2.823 | 1.579 | -1.788 | 0.0759 |
| **sizemedium** | 4.988 | 3.816 | 1.307 | 0.1932 |
| **sizesmall** | 8.589 | 4.243 | 2.024 | 0.04475 |
| **speedlow** | 6.345 | 4.779 | 1.328 | 0.1864 |
| **speedmedium** | 1.142 | 3.341 | 0.3418 | 0.733 |

Table 7: Fitting linear model: a1 ~ .

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 161 | 16.15 | 0.4746 | 0.4203 |

```
## [1] 15.33034
```

**Answer** The $R^2$ value is a statistical measure of how close the data are to the fitted regression line. Therefore, to see if our model is a good fit, we can analyze out $R^2$ value. In summary output, we can see that the $R^2$ value is 0.4242. Therefore, we would say that 42.42% of the variability in the outcome data cannot be explained by the model. In my opinion, I don't think this model fits the data well.

**Question 2 (c)**

Now estimate the prediction error for this model by computing test RMSE. How accurate does this suggest to you that the predictions are? (*Hint:* is the prediction RMSE much smaller than the raw variation in the test data?)

```
## [1] 14.91353
```

```
## [1] 22.15102
```

**Answer** We can see that the prediction RMSE is about 14.9, while the raw variation is about 22.2. Since the prediction RMSE is much smaller than the raw variation, we can say that the model does a good job predicting the data, so the predictions are fairly accurate.

**Question 2(d)**

How, if at all, would you suggest obtaining improved predictions? Answer in 1-2 sentences.

**Answer** My suggestion to obtaining improved predictions would be to try out different models such as spline and polynomial, instead of defaulting to the linear model. We have seen that polynomial models do well in this case because of their flexibility, so we could use that to our advantage.

## Part 3. Theory

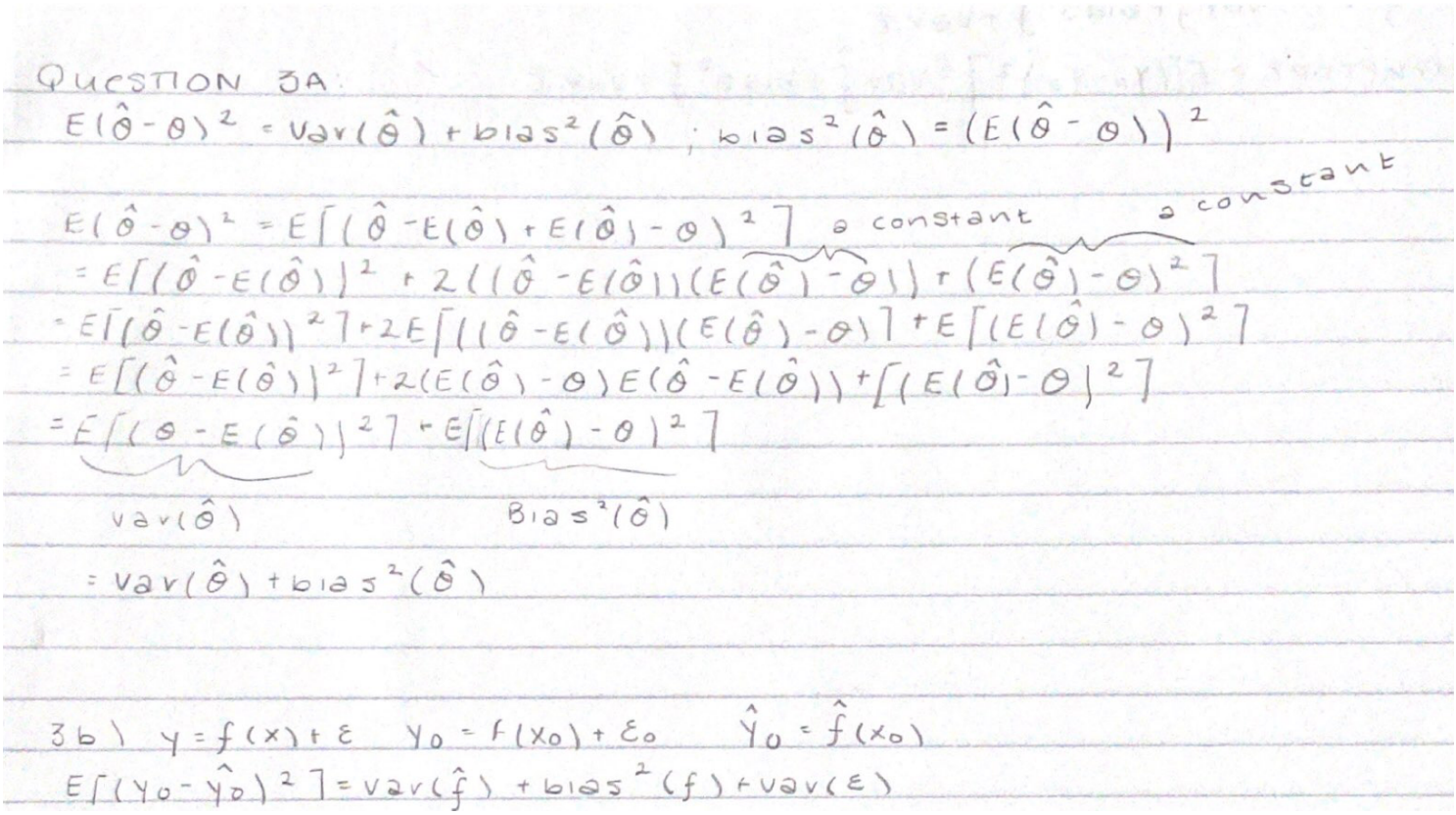
For this part, you do not have to type your answers. Working them out on pen and paper and scanning the pages is perfectly acceptable.

### Question 3 (a)

Show that for an arbitrary estimator $\hat{\theta}$,

$$E\left(\hat{\theta}-\theta\right)^2 = \text{var}\left(\hat{\theta}\right) + \text{bias}^2\left(\hat{\theta}\right)$$

where $\text{bias}^2(\hat{\theta}) = \left(E\left(\hat{\theta}-\theta\right)\right)^2$.



QUESTION 3A:

$E(\hat{\theta}-\theta)^2 = \text{Var}(\hat{\theta}) + \text{bias}^2(\hat{\theta})$ ; $\text{bias}^2(\hat{\theta}) = (E(\hat{\theta}-\theta))^2$

$E(\hat{\theta}-\theta)^2 = E[(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta)^2]$ — $E(\hat{\theta})$ a constant, $\theta$ a constant

$= E[(\hat{\theta}-E(\hat{\theta}))^2 + 2((\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta)) + (E(\hat{\theta})-\theta)^2]$

$= E[(\hat{\theta}-E(\hat{\theta}))^2] + 2E[((\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta)] + E[(E(\hat{\theta})-\theta)^2]$

$= E[(\hat{\theta}-E(\hat{\theta}))^2] + 2(E(\hat{\theta})-\theta)E(\hat{\theta}-E(\hat{\theta})) + [(E(\hat{\theta})-\theta)^2]$

$= E[(\theta-E(\hat{\theta}))^2] + E[(E(\hat{\theta})-\theta)^2]$

$\text{Var}(\hat{\theta})$ $\quad$ $\text{Bias}^2(\hat{\theta})$

$= \text{Var}(\hat{\theta}) + \text{bias}^2(\hat{\theta})$

3b) $y = f(x) + \varepsilon$ $\quad y_0 = f(x_0) + \varepsilon_0$ $\quad \hat{y}_0 = \hat{f}(x_0)$

$E[(y_0-\hat{y}_0)^2] = \text{Var}(\hat{f}) + \text{bias}^2(f) + \text{Var}(\varepsilon)$

**Question 3 (b)**

Let

$$\mathbf{y} = f(\mathbf{X}) + \epsilon$$

where $\mathbf{X}$ is fixed and $f$ is an arbitrary function. Let $\hat{f}$ be an estimator of $f$ based on $\mathbf{y}, \mathbf{X}$. Let $y_0 = f(\mathbf{x}_0) + \epsilon_0$ and assume that $\epsilon_0 \perp y_j$ for every $y_j \in \mathbf{y}$ and $\mathbb{E}\epsilon_0 = 0$. Finally, define $\hat{y}_0 = \hat{f}(\mathbf{x}_0)$. Show that

$$\mathbb{E}\left(y_0 - \hat{y}_0\right)^2 = \text{var}\hat{f} + \text{bias}^2\hat{f} + \text{var}\epsilon$$

3b) $y = f(x) + \varepsilon \quad y_0 = f(x_0) + \varepsilon_0 \quad \hat{y}_0 = \hat{f}(x_0)$

$E[(y_0 - \hat{y}_0)^2] = \text{var}(\hat{f}) + \text{bias}^2(f) + \text{var}(\varepsilon)$

$E[(y_0 - \hat{y}_0)^2] = E[(f(x_0) + \varepsilon_0 - \hat{f}(x_0))^2] = (f(x_0) + \varepsilon_0 - \hat{f}(x_0))(f(x_0) + \varepsilon_0 - \hat{f}(x_0))$

$E[f(x_0)^2 + 2\varepsilon_0 f(x_0) - 2f(x_0)\hat{f}(x_0) - 2\hat{f}(x_0)\varepsilon_0 + \varepsilon_0^2 + \hat{f}(x_0)^2$

$= E[(\hat{f}(x_0) - f(x_0))^2 + \varepsilon_0^2 - 2\hat{f}(x_0)\varepsilon_0 + 2f(x_0)\varepsilon_0]$

$= E[(\hat{f}(x_0) - f(x_0))^2] + E[\varepsilon_0^2] + 2E[\varepsilon_0(f(x_0) - \hat{f}(x_0))]$

$E[(\hat{f}(x_0) - f(x_0))^2] + E[\varepsilon_0^2] + 2E[\varepsilon_0] \cdot E[(f(x_0) - \hat{f}(x_0))]$

$\text{var}\,\varepsilon \qquad = 0$

$= E[(\hat{f}(x_0) - f(x_0))^2] + \text{var}\,\varepsilon$

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

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

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

constant term $+ (E[\hat{f}(x_0)] - f(x_0))^2$

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

$E[(X - E[X])^2] + E[(E(\hat{f}(x_0)) - f(x_0))^2]$

$E[X] - E[X] = 0 \qquad \text{CONSTANT}$

$\Rightarrow \text{var}\hat{f} + 0 + (E[\hat{f}(x_0)] - f(x_0))^2 = \text{var}\hat{f} + \text{bias}^2 f$

$\text{bias}^2 f$

plugging in $\text{var}\hat{f} + \text{bias}^2\hat{f}$ into $E[(\hat{f}(x_0) - f(x_0))^2] + \text{var}\,\varepsilon$

we get $\rightarrow \text{var}\,\hat{f} + \text{bias}^2\hat{f} + \text{var}\,\varepsilon$

THEREFORE $\rightarrow E[(y_0 - \hat{y}_0)^2] = \text{var}\,\hat{f} + \text{bias}^2 f + \text{var}\,\varepsilon$ ✓