# Midterm ENEC 562

Nguyen Tien Anh Quach

2024-03-21

## Question 1

### Hypothesis

$H_0$: The average plant biomass among six different salt treatments have no difference, after accounting for the effect of geographic proximity.

$H_a$: At least one pair of salt treatments has a significantly different effect on the average plant biomass, after accounting for the effect of geographic proximity.

### Test justification

For this question, I am using the **blocked ANOVA** to test for the effects of salt treatment on plant biomass. This is because aside from the six different salt treatments, the geographic proximity may also affect the outcome and thus, needs to be accounted as a block effect.

### Assumptions of blocked ANOVA

a. There are two outliers reported, but none is extreme.

b. The plant biomass among six salt treaments are normally distributed, as Shapiro-Wilk test reported non-significant p-values for the 10 g/m$^2$ (0.92), 15 g/m$^2$ (0.08), 20 g/m$^2$ (0.3), 25 g/m$^2$ (0.29), 30 g/m$^2$ (0.85), and 35 g/m$^2$ (0.73). Normality of biomass within each treatment is also shown in Figure 1.

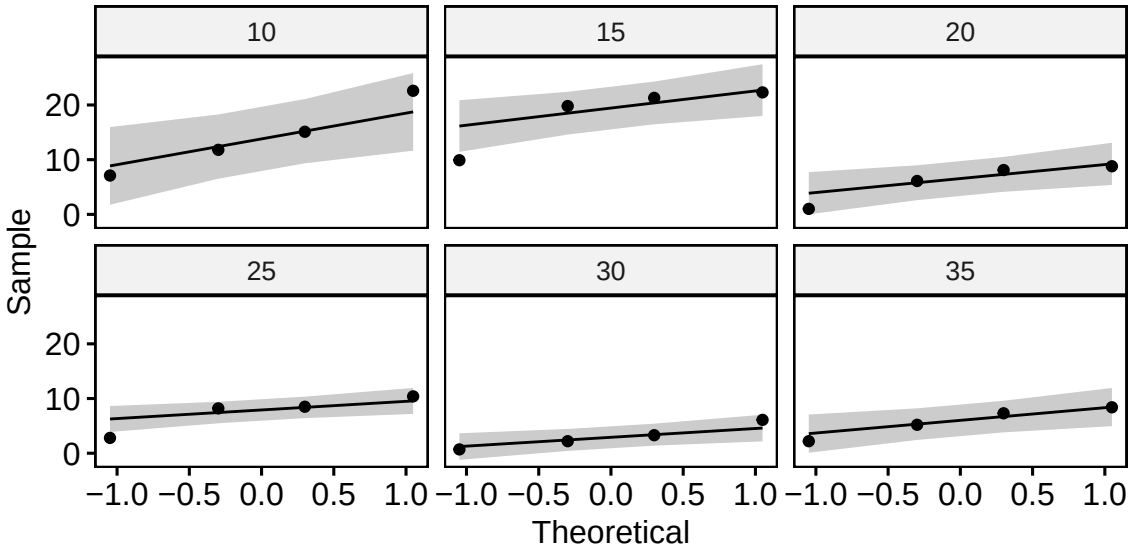



Figure 1. QQ plot of plant biomass collected from six levels of salt treatment.

c. Levene's test for equal variance returned the non-significant p-value of 0.69 across salt treatments and of 0.8 across blocks of geographic proximity. Therefore, the assumption of equal variance is not violated.

Figure 2 is to visualize how plant biomass varies among salt treatment groups:

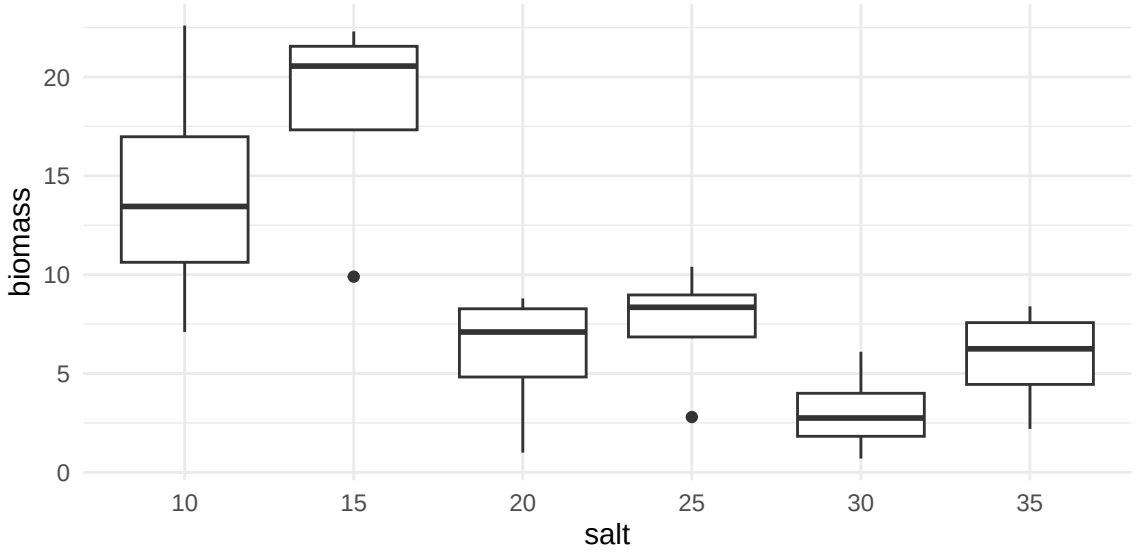



Figure 2. Box plot of plant biomass collected from six salt treatments.

## Blocked ANOVA test results

Results showed that salt treatment ($F(5,15) = 17.71$, $p = 8.08 \times 10^{-6}$) and geographic proximity ($F(3, 15) = 9.42$, $p = 9.6 \times 10^{-4}$) had significant effects on the average plant biomass.

## Post-hoc test

Tukey HSD test showed that the average plant biomass were different among:

a. 10 vs 20 $g/m^2$ ($p = 0.00894$)

b. 10 vs 25 $g/m^2$ ($p = 0.0374$)

c. 10 vs 30 $g/m^2$ ($p = 5.44 \times 10^{-4}$)

d. 10 vs 35 $g/m^2$ ($p = 0.00717$)

e. 15 vs 20 $g/m^2$ ($p = 1.76 \times 10^{-4}$)

f. 15 vs 25 $g/m^2$ ($p = 6.7 \times 10^{-4}$)

g. 15 vs 30 $g/m^2$ ($p = 1.51 \times 10^{-5}$)

h. 15 vs 35 $g/m^2$ ($p = 1.44 \times 10^{-4}$)

The difference in average plant biomass among salt treatments and blocks can also be visualized in Figure 3:

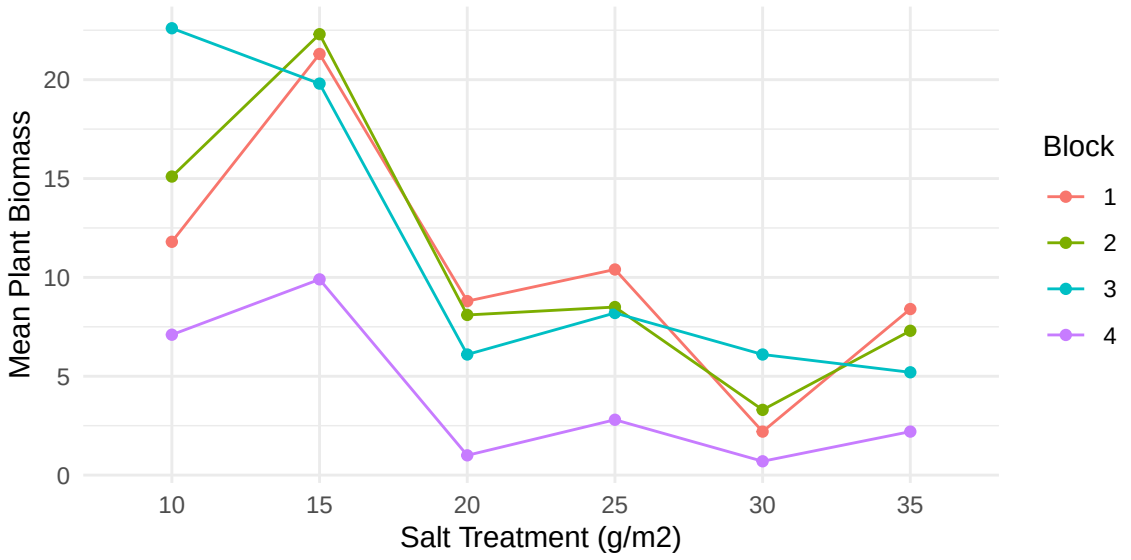



Figure 3. Line graph of average plant biomass among salt treatments and blocks.

# Question 2

## Hypothesis

$H_0$: The average scale thickness among different supplement and doses have no difference AND there is no interaction effect between supplement and dose.

$H_a$: At least one pair of supplement and dose treatments has a significantly different effect on the average scale thickness AND there is an interaction effect between supplement and dose on the average scale thickness.

## Test justification

Scale thickness is a quantitative variable and tested on two different categorical variables, supplement and dose. As I am trying to test for the significant difference in average scale thickness in different groups of supplements and doses, a **two-way ANOVA** works best.

## Assumptions of two-way ANOVA

a. There are two outliers reported, but none is extreme.

b. I built a linear model of the scale thickness, the supplement treatment, and the dose treatment. I then tested the normality of the model residuals. As shown in the QQ plot (Fig. 4), all the points fall approximately along the reference line. In addition, Shapiro-Wilk test yielded a non-significant p-value of 0.50. Therefore, normality assumption is met!

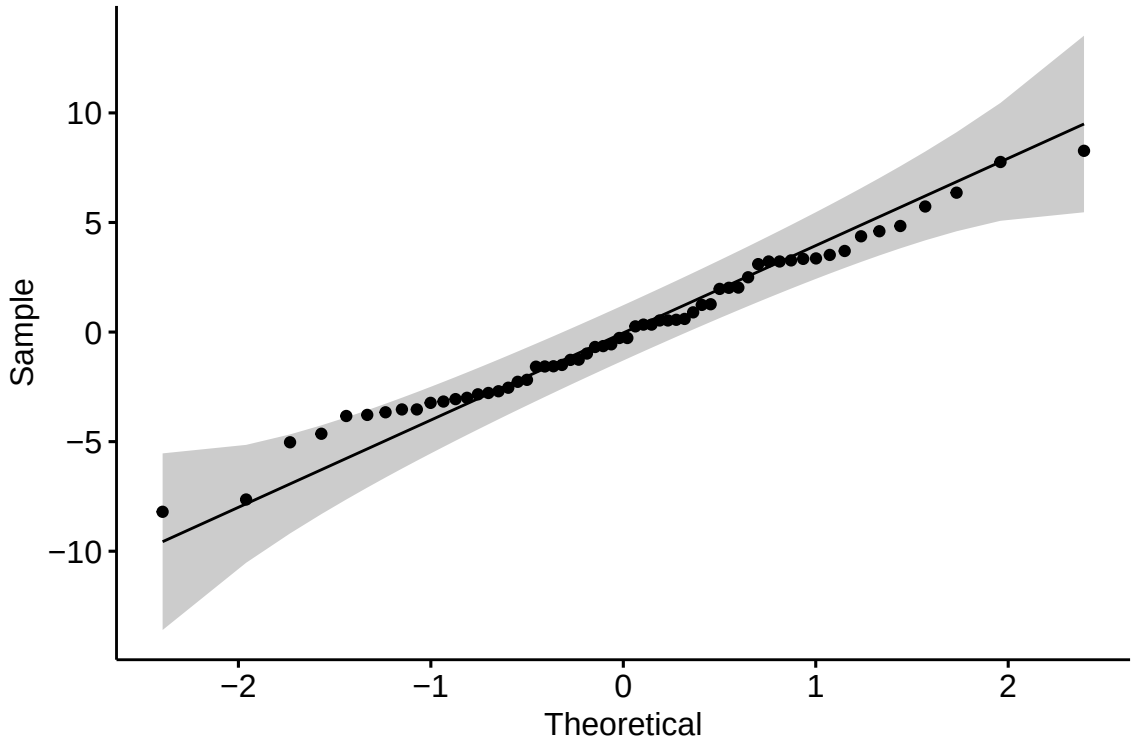



Figure 4. QQ plot of linear model residuals (thick  dose * supp).

c. Levene's test for equal variance returned the non-significant p-value of 0.15. Therefore, the assumption of equal variance is not violated.

## Two-way ANOVA test results

Results showed that supplement treatment (F(1, 54) = 15.57, p = $2.31 \times 10^{-4}$) and dose treatment (F(2, 54) = 92, p = $4.05 \times 10^{-18}$) had significant effects on the average pangolin scale thickness. In addition, there was a statistically significant interaction effect between the supplement and dose treatments on the average scale thickness (F(2, 54) = 4.11, p = 0.022).

## Post-hoc test

As the interaction effect is significant, I am going to determine the simple main effect of each treatment and conduct multiple pairwise comparisons.

### Simple main effect

The simple main effect of "dose" on scale thickness was statistically significant at $\alpha$ of 0.025 for VitB (F(2, 54) = 62.54, p < 0.001) and Zinc (F(2, 54) = 33.56, p < 0.001) supplements.

### Pairwise comparison

Tukey HSD test showed that the average scale thickness were different among:

VitB Supplement:

a. 0.5 vs 1 mg (p = $1.75 \times 10^{-5}$)

b. 0.5 vs 2 mg (p = $1.66 \times 10^{-11}$)

c. 1 vs 2 mg ($p = 6.61 \times 10^{-6}$)

Zinc Supplement:

a. 0.5 vs 1 mg ($p = 1.58 \times 10^{-5}$)

b. 0.5 vs 2 mg ($p = 9.39 \times 10^{-8}$)

The difference in scale thickness among groups of supplement and dose treatments can be observed in Figure 5 below:

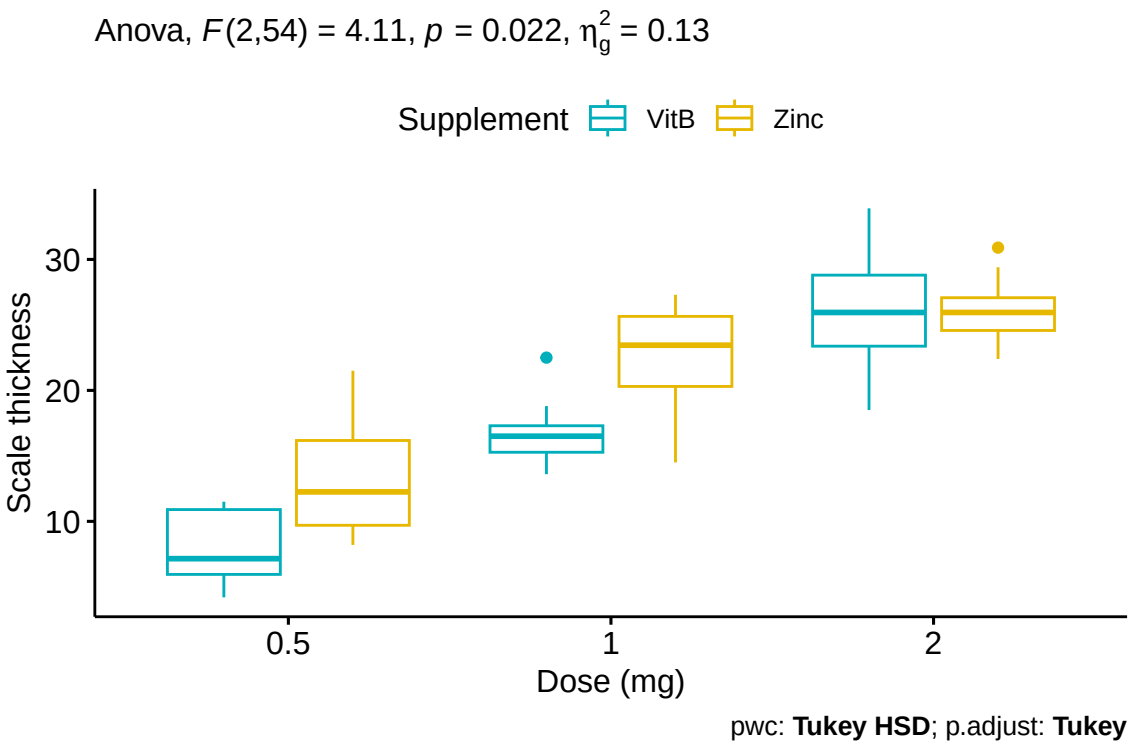



Figure 5. Boxplot of pangolin scale thickness among groups of supplement and dose treatment.

# Question 3

## Hypothesis

$H_0$: The average loading time is the same between the two providers.

$H_a$: The average loading time is different between the two providers.

## Test justification

As I am comparing the mean loading time between the two independent internet providers, it is best to use **two-sample t-test**. If data do not meet the assumptions of the test, I will attempt to transform the data and redo the test. In addition, I will conduct a **Wilcoxon rank sum test** as a non-parametric test.

## Assumptions of two-sample t-test

a. There is one outlier for data from each internet provider, Turbo Net and Speed Web. However, both are not extreme outliers.

b. Despite looking quite normal (Fig. 6), the Shapiro-Wilk test showed that loading times from both internet providers are non-normal.

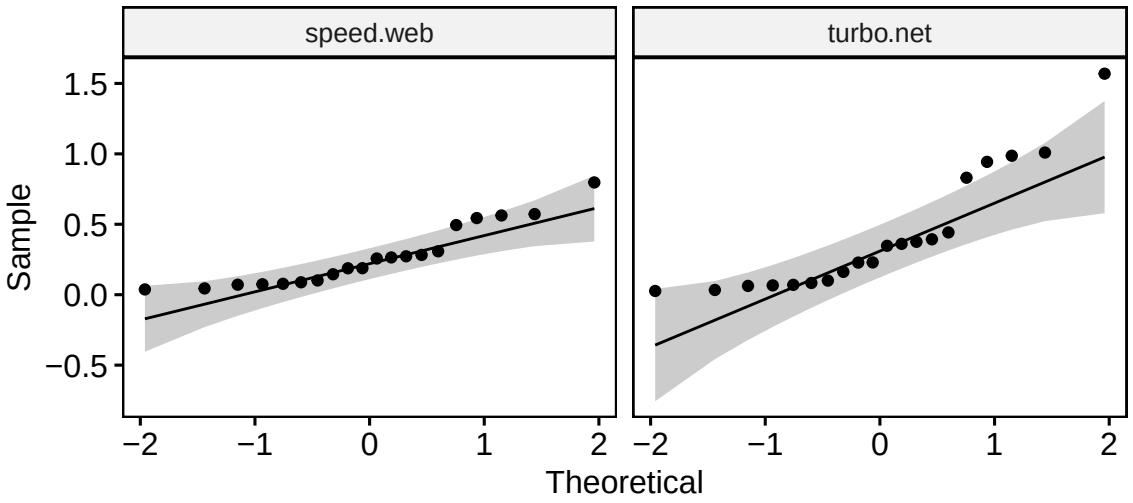



Figure 6. QQ plot of loading times from two internet providers, Turbo Net and Speed Web.

## Data transformation

Now I am attempting to log transform the loading times and re-checking the assumptions.

## Assumptions of two-sample t-test

a. There is no outlier reported this time.

b. Log-transformed loading times look more normal as most points fall approximately on the line. In addition, Shapiro-Wilk test yielded non-significant p-values (Turbo net: $p = 0.40$; Speed Web: $p = 0.40$).

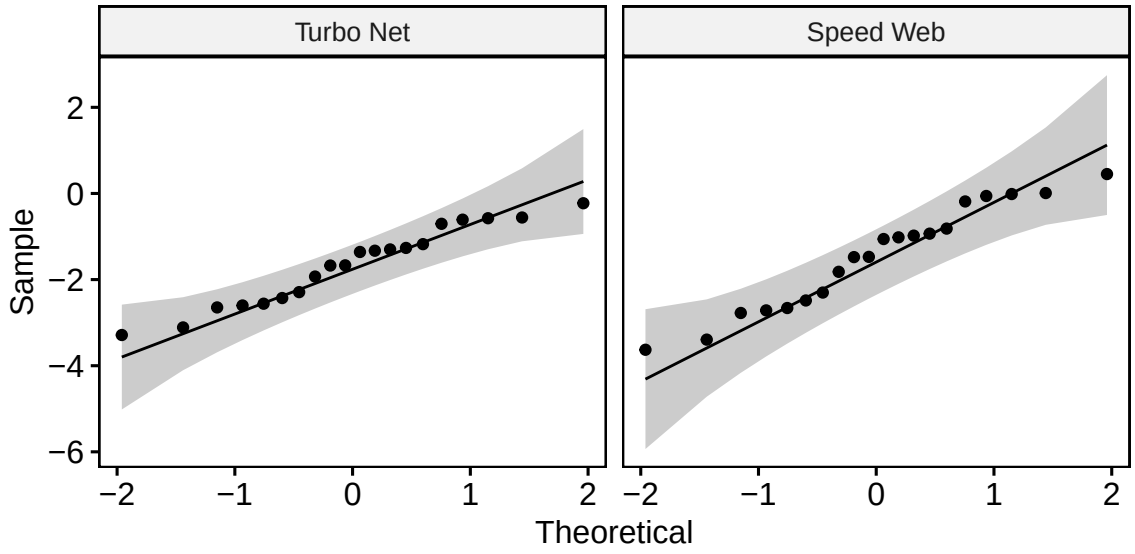



Figure 7. QQ plot of log-transformed loading times from two internet providers, Turbo Net and Speed Web.

The p-value of the Levene's test is non-significant (p = 0.18), suggesting that there is no significant difference between the variances of the two internet providers' loading times.

### Two-sample t-test result

Results from two-sample t-test showed that the log loading times of two internet providers are not significantly different (t (38) = -0.59; p = 0.56) with negligible effect size (d = -0.19). Therefore, I failed to reject the null hypothesis!

The distribution of loading times can be seen below in Figure 8.

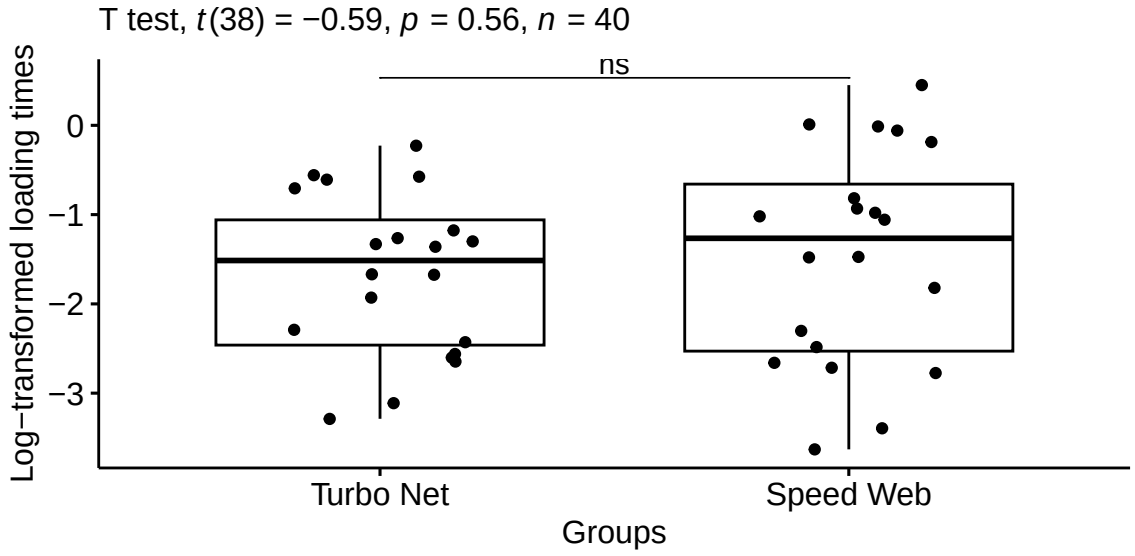



Figure 8. Boxplot of log-transformed loading times from two internet providers, Turbo Net and Speed Web and the denoted t-test results.

### Wilcoxon rank sum test

Now I am going to conduct the Wilcoxin rank sum test on the original dataset.

Similar to the two-sample t-test, the Wilcoxon rank sum test yielded non-significant p-value of 0.602 and small effect size (0.085). Therefore, I failed to reject the null hypothesis!

### Conclusion

Overall, the two different tests (two-sample t-test on log-transformed time and Wilcoxon rank sum test) both showed that the average loading times are not significantly different.

## Question 4

### Hypothesis

$H_0$: There is no significant linear relationship between mean GDP per capita and mean EPI.

$H_a$: There is a significant linear relationship between mean GDP per capita and mean EPI.

### Correlation between the two variables

The correlation between GDP per capita and EPI is 0.7. This means that GDP per capita and EPI are positively and significantly ($p < 0.001$) correlated. The higher the GDP per capita, the higher the EPI.

### Scatterplot of the two variables

The relationship between GDP per capita and EPI does not look linear (Fig. 9). It looks like the high GDP per capita data points are strongly influencing the relationship. A log-transformation of the GDP per capita may help.

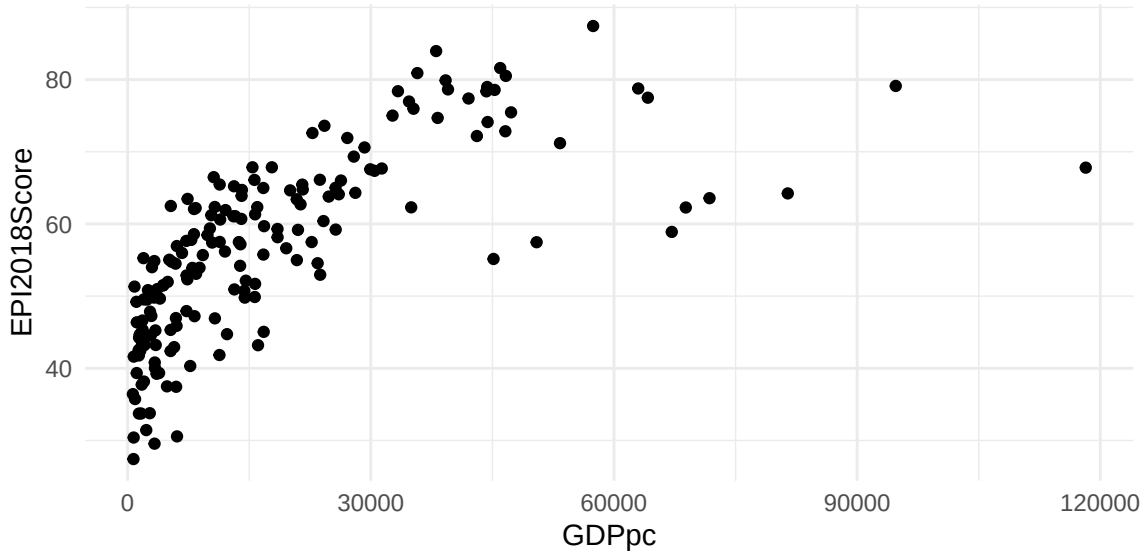



Figure 9. Scatterplot of GDP per capita and EPI.

#### Transform GDPpc

I log-transformed GDP per capita and the relationship between the two variables looks so much more linear now (Fig. 10), as the high GDP per capita values now become less of an outlier.

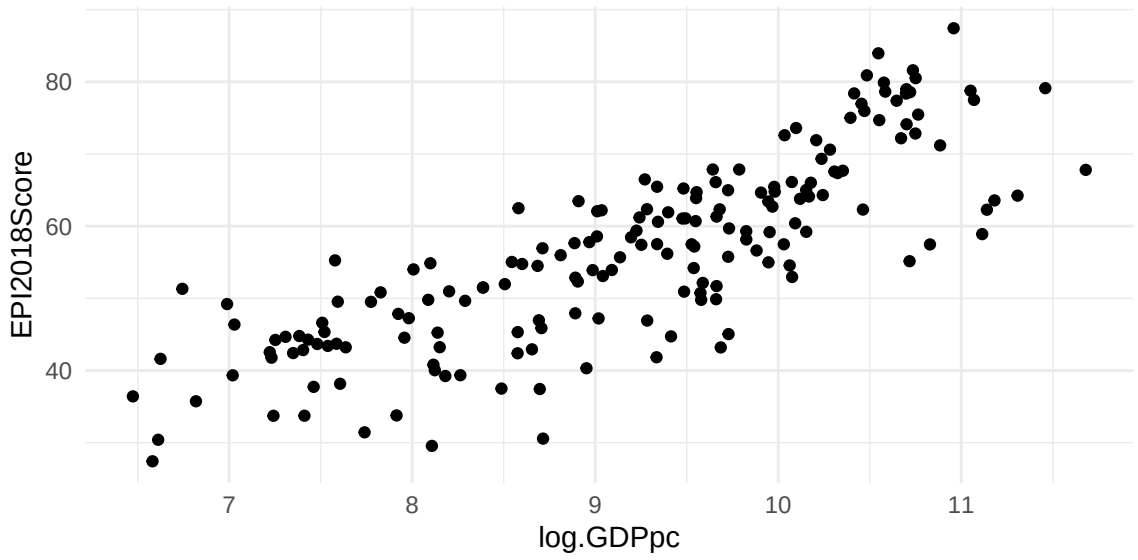



Figure 10. Scatterplot of log-transformed GDP per capita and EPI.

## Revised hypothesis

$H_0$: There is no significant linear relationship between mean log-transformed GDP per capita and mean EPI.

$H_a$: There is a significant linear relationship between mean log-transformed GDP per capita and mean EPI.

## Linear regression

Mean log-transformed GDP per capita has a significantly positive linear relationship with mean EPI (Fig. 11). A 1 unit increase in mean log-transformed GDP per capita resulting in a 8.68 unit increase in mean EPI ($R^2 = 0.66$, Adj. $R^2 = 0.66$, $F(1,178) = 354.5$, $p < 0.001$). Alternatively, as the coefficient of GDPpc is 8.68, for every 1% increase in GDP per capita, the EPI increases by about 0.08. The regression equation is :

$$\hat{y_i} = -23.28 + 8.68 \times log.GDPpci$$

One thing to notice is that the intercept of the linear model is -23.28, which means that when the mean log-transformed GPD per capita is 0, the mean EPI is -23.28. That is not realistic!

## Interpretation

### p-value

The null hypothesis is rejected. The p-value of the linear model is significant ($p < 0.001$). However, this does not mean that the relationship between mean GPD per capita and mean EPI is significant. The p-value only indicates that mean log-transformed GPD per capita and mean EPI have a statistically significant relationship.

### R-squared value

The adj R-squared value of 0.66 means that approximately 66% of the variability in the EPI score can be explained by the log-transformed GDP per capita.

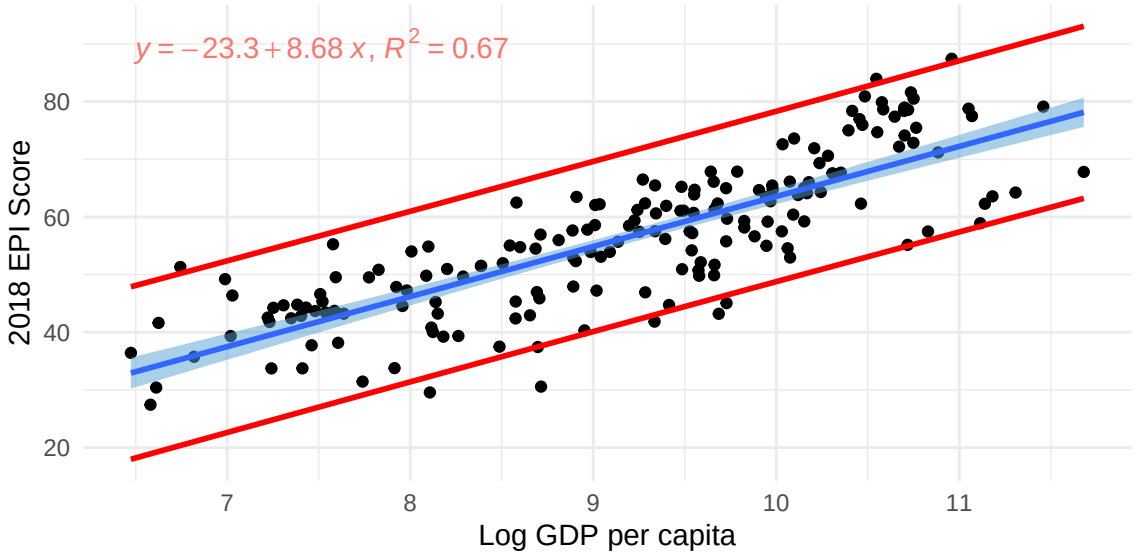



Figure 11. Scatterplot of log-transformed GDP per capita and EPI with confidence (blue) and prediction (red) intervals. Linear equation and R-squared value are denoted.

# Question 5

## Log-transformation justification

I went through the entire question using raw ozone concentration. However, the model's assumption of the normal distribution of error failed every time. Therefore, I decided to log transformed the ozone concentration (from ozone to log.ozone) and the normality assumption was barely met. I am not showing the entire process in the exam because it may be too repetitive to run the model, test the assumptions, and rerun the model again.

## Hypothesis

$H_0$: There is no significant linear relationship between mean log-transformed ozone concentration and mean solar radiation, wind speed, and temperature.

$H_a$: There is a significant linear relationship between mean log-transformed ozone concentration and (1) at least one independent variable, mean solar radiation, wind speed, and temperature and (2) at least an interaction effects among the independent variables.

## Model selection

Model training codes can be found in "Q5 Model Training.R". This is to reduce the knitting time.

From the results of repeated 3-fold cross validation with 100 repeats, the best model was:

$$\text{log.ozone} \sim \text{rad} + \text{temp} + \text{wind} + \text{wind*temp}$$

This model was reported to have the lowest RMSE (0.52) and highest R-squared value (0.66).

## Run multiple linear regression

| Coefficient | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -2.5845520 | 1.1936395 | -2.165270 | 0.0326096 |
| rad | 0.0025939 | 0.0005483 | 4.731089 | 0.0000069 |
| temp | 0.0782341 | 0.0145784 | 5.366434 | 0.0000005 |
| wind | 0.1661571 | 0.1052917 | 1.578063 | 0.1175308 |
| temp:wind | -0.0029299 | 0.0013399 | -2.186651 | 0.0309666 |

The result showed that the model was statistically significant (Adjusted $R^2 = 0.67$, $F(4, 106) = 45.28$, $p < 0.001$).The regression equation is :

$$\hat{y_i} = -2.58 + 0.002 \times radi + 0.078 \times tempi + 0.16 \times windi - 0.003 \times temp * windi$$

For this model, temperature ($p < 0.001$, df = 106), radiation ($p< 0.001$, df = 106), and the interaction effect between temp and wind ($p = 0.031$) have statistically significant coefficients.

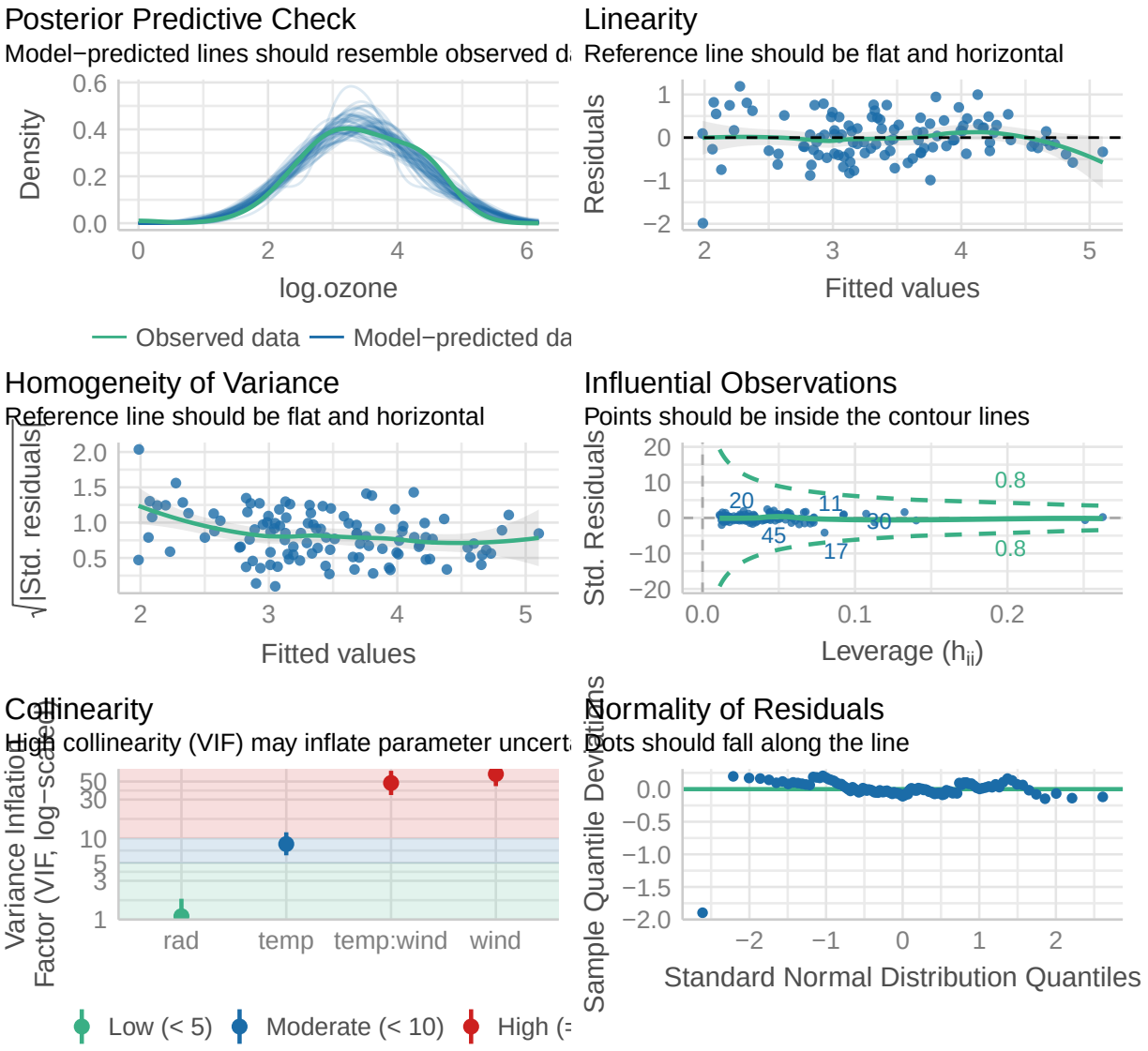
A 1 unit increase in mean radiation results in a 0.002 unit increase in mean log-transformed ozone concentration.

A 1 unit increase in mean temperature results in a 0.078 unit increase in mean log-transformed ozone concentration.

A 1 unit increase in mean wind speed results in a 0.16 unit increase in mean log-transformed ozone concentration. However, the effect of wind on log-transformed ozone concentration was not significant.

A 1 unit increase in mean temperature, holding wind speed constant, results in a 0.003 unit decrease in mean log-transformed ozone concentration.

## Check model assumptions




VIFs are extremely high due to interaction terms. Therefore, we can safely ignore them!Mean residuals are approximately 0. Errors are uncorrelated as they are randomly scattered around $\epsilon = 0$. These assumptions are met!

Results from GVLMA and Model Performance Check showed that (1) independent and dependent variables have no linear correlations and (2) the assumptions of residuals' normality and constant variance were not met.

In addition, there are many influential points: 11, 17, 18, 20, 30, 45, 77, 85. I am removing them and rerunning the model.

## Rerun model

| Coefficient | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -1.8975649 | 1.0613575 | -1.787866 | 0.0768884 |
| rad | 0.0021503 | 0.0004819 | 4.461708 | 0.0000217 |
| temp | 0.0709259 | 0.0128817 | 5.505954 | 0.0000003 |
| wind | 0.0952244 | 0.0931297 | 1.022492 | 0.3090657 |
| temp:wind | -0.0020422 | 0.0011822 | -1.727396 | 0.0872470 |

The result showed that the model was statistically significant (Adjusted $R^2 = 0.72$, $F(4, 98) = 67.22$, $p < 0.001$).The regression equation is :

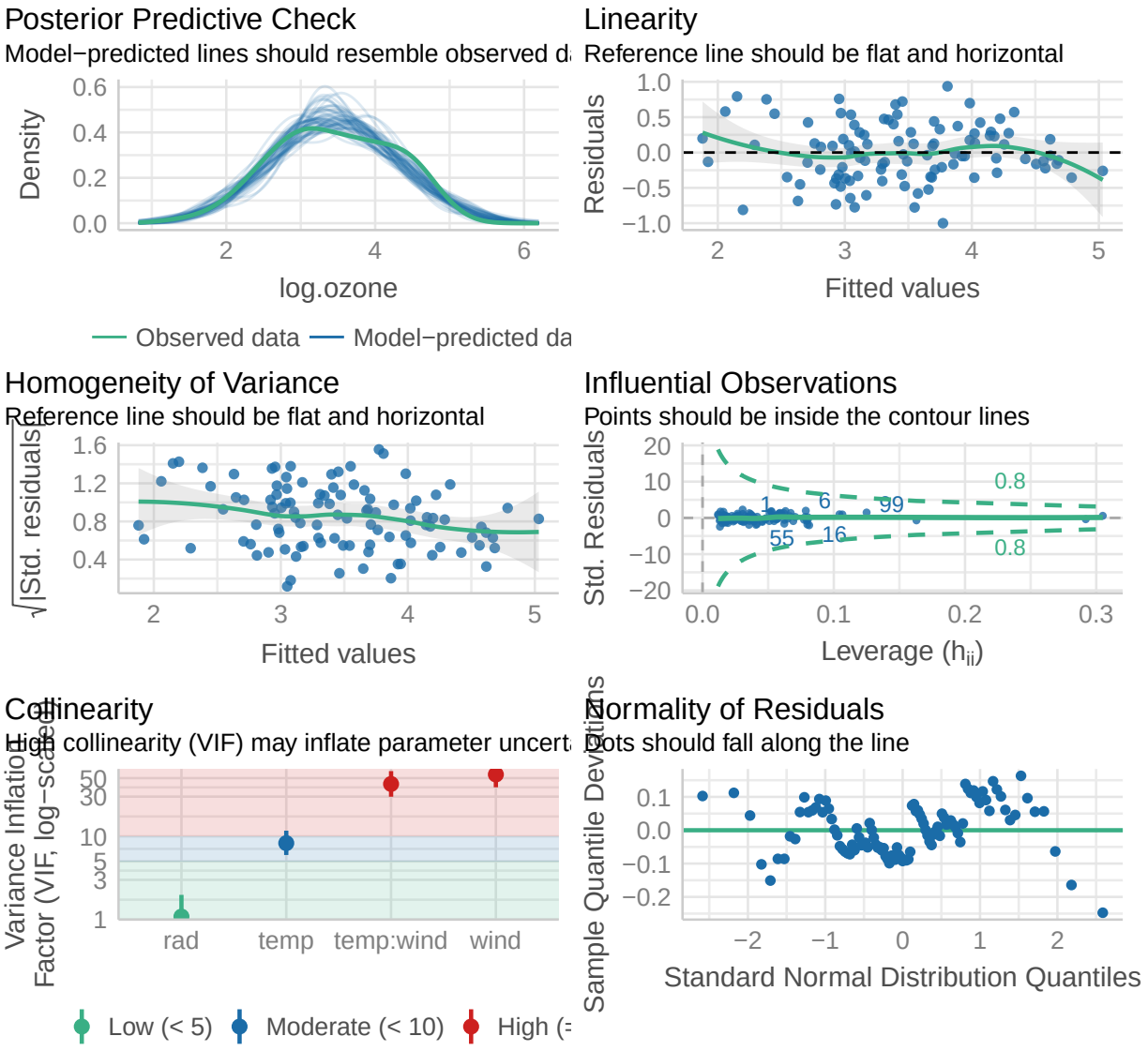
$$\hat{y_i} = -1.89 + 0.002 \times radi + 0.07 \times tempi + 0.095 \times windi - 0.002 \times temp * windi$$

For this model, only temperature ($p < 0.001$, $df = 98$) and radiation ($p< 0.001$, $df = 98$) had significant coefficients. Wind ($p = 0.30$) and the interaction effect between temp and wind ($p = 0.087$) became non-significant in this model.

A 1 unit increase in mean radiation results in a 0.002 unit increase in mean log-transformed ozone concentration.

A 1 unit increase in mean temperature results in a 0.07 unit increase in mean log-transformed ozone concentration.

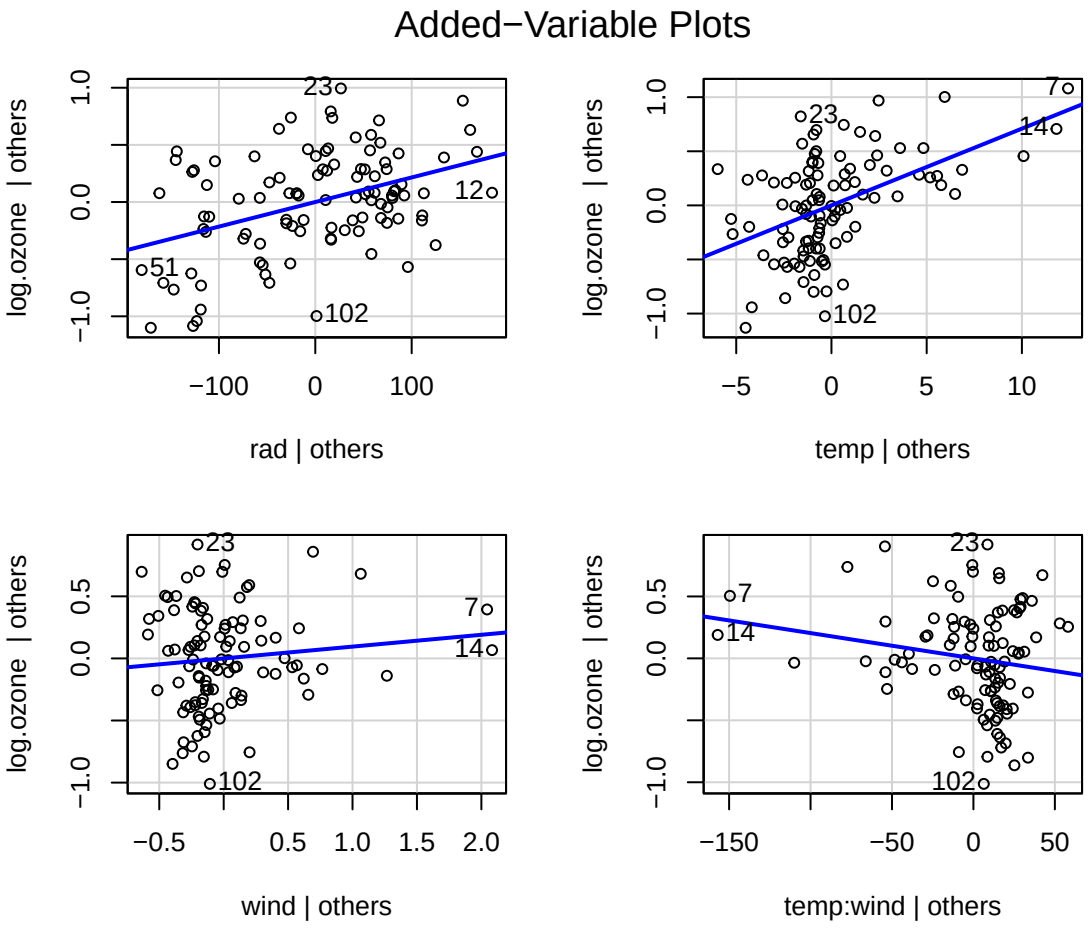
## Check model assumptions




Mean residuals are approximately 0. Errors are uncorrelated as they are randomly scattered around $\epsilon = 0$. These assumptions are met! Also, results from GVLMA and Model Performance Check showed that all assumptions are now met (normality, constant variance)

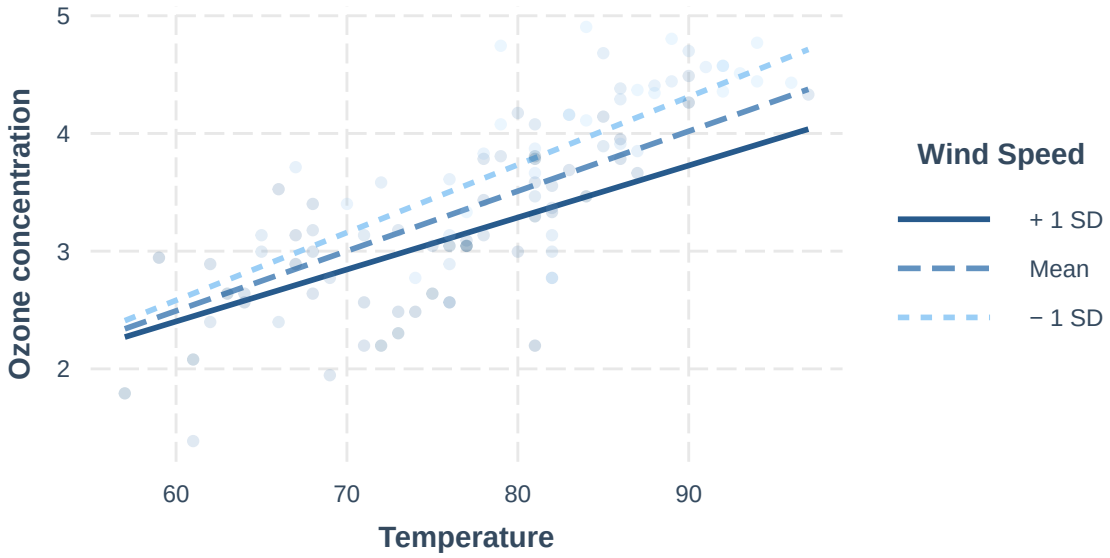
## Plots

Below are the added variable plots that show the relationship between the independent variable and dependent variable when a variable is only included as a main effect:




The relationships match with what was shown from the LM results.

And although non-significant, below is the plot showing the interactions between temp and wind:



# Question 6

## Research question

What were the environmental factors that influence the distributions of coral reef fish specieson a sub-regional scale?

## Model type

The authors chose to use the GLM (General Linear Model). Specifically, they used binary logistic regression due to the binomial nature of the dependent variable (presence or absence of fish species).

## Data and variables

### Data

The authors used fish data from three different sources that are, in total, available at 105 sites. They also collected physical data from remote sensing, community analysis, ocean charts, and local and expert knowledge. The authors then identified a subset of uncorrelated predictor variables as proxies for several environmental variables.

### Variables

The response variable is the presence or absence of fish species.

The predictor variables are:

1. Reef class: categorical, 5 classes
2. Bioregions: categorical, 9 classes
3. Exposure: ordinal, 4 classes / CHOSEN!
4. Presence of land-water interface: ordinal, 2 classes / CHOSEN!
5. Mean depth at 500 m proximity: numerical / CHOSEN!
6. Mean depth at 1000 m proximity: numerical
7. Distance to nearest estuary: numerical / CHOSEN!
8. Distance to nearest land: numerical

## Table 4 results

The regression equation for *Caesio lunaris* was:

$$\ln(\text{p/1-p}) = -1.086 + 2.6 * \text{Exposure} + 0.003 * \text{Depth}$$

Table 4 in general presented information about whether fishes' habitat preferences depend on any physical characteristics. For *Caesio lunaris*, its habitat preference depends on the exposure variable and depth variable.

Holding all other variables constant:

For every one unit increase in exposure, the log odds of detection/presence increase by 2.6. In other words, as the odds ratio is 13.46, for every increase of 1 unit in exposure, the chance to find a *Caesio lunaris* goes up by about 1200%.

For every one unit increase in depth, the log odds of detection/presence increase by 0.003. In other words, as the odds ratio is 1.003, for every increase of 1 unit in depth, the chance to find a *Caesio lunaris* goes up by about 0.3%.

Table 4 also has information about the efficiency of a model for each species based on its AIC scores.

## Model comparison

The authors compare models based on a reference model and its AIC (95% CI). Whenever the AIC of a model of a fish species falls inside the 95% CI of the reference AIC, that means the distribution of that fish species did not differ significantly from a random distribution. Ecologically speaking, the fish species did not have any preference for a habitat and usually, they are the ones that are abundant and common at many sites.

## Limitations

1. Data collected came from many different sources, thus contain uncertainties (i.e., different levels of accuracy). Other free datasets can come from NOAA and NASA, but the resolution is too coarse to be applicable to the study area.

2. The response of fish species to the environmental predictors may not be linear. Therefore, using GLM may not be the best method to predict how fish species are distributed based on physical characteristics. The authors gave an example about the suspended solids, how they change non-linearly from estuary to outer reef, and how fish response may also not be linear.

3. First, I want to mention the spatial autocorrelation. Fish can swim to places as they are a mobile organism. In addition, closer sites may represent similar habitat features. Second, there are so many other environmental and chemical variables that may affect the distribution of fish (e.g., pH, DO, etc.). Having only 4 variables at the end that are only about depth and distance is underestimating other confounding factors. Lastly, I want to mention that using stepwise selection to choose model may not be the best option. Authors can use cross validation to select the best model, which have many advantages over stepwise selection (e.g., training-test data sets, RMSE, no AIC).