# Long Term Ecological Population and Productivity Statistical Analysis

Daniel Crownover

2023-02-23

## Contents

Answer the following questions in 1 knitted PDF document. Your answers do not need to be in paragraphs (they can be bullets). Also turn in your .Rmd file.

## 0.1 Problem 1

a) Null and alternative hypotheses of your tests (2.5)

- H_0_1: The biomass between groups stayed the same between salt exposed and non salt exposed block groups
- H_a_1: The biomass between groups changed between salt exposed and non salt exposed block groops

b) Justification for your choice of test (2.5)

- ALL Figures "1.b"
- Our value is NOT statistically significant from shapiro wilks (p = .24) in figure 1.b Normality Assumption and Shaprio test. We fail/ do not have sufficient evidence to reject the null(that its normal) therefore, we assume normality. In the QQ plot(Figure 1.b.Normality and shapiro), as all the points fall approximately along the reference line, we can assume normality. In figure 1.b Homogeneity of Variance, the levine test showed, (p = .79) for block and (p = .69) for salt Non-statistically significant value, we fail to reject null for block groups between biomass – we assume their are equal variances. Using the Levene's test for biomass between salt concentration, is not significant. Therefore, we can assume the homogeneity of variances in the different groups in biomass between salt
- Knowing that there is homogeneity among the variances for each, or pretty much the variances are equal, we are justified in our use of one way anova.

c) A description of how you checked the assumptions of your statistical test (10) and d) Results of your statistical test, interpreting your test in 2-3 sentences that include the appropriate reporting of the statistics as well as an appropriate figure (10)

- Figure 1.c Anova, The salt effect has a significant impact on the response variable. The F-value of 17.709 with a very small p-value (8.08e-06) indicates that the mean differences among salt levels are unlikely due to random chance. The generalized Eta Squared (ges) of 0.855 suggests that about 85.5% of the total variance in the response variable can be attributed to the salt effect.
- Figure 1.d line plot shows the line plot which shows salts negative effects on biomass.

e) An interpretation of any necessary post-hoc tests (5)

- Salt 10 to Salt 20: The change in the response variable from Salt 10 to Salt 20 is estimated to be approximately -8.15 units. This suggests that increasing the salt level from 10 to 20 results in a decrease in the response variable by about 8.15 units.(p = **** highly significant)
- Salt 10 to Salt 25: The change in the response variable from Salt 10 to Salt 25 is estimated to be approximately -6.675 units. This indicates that increasing the salt level from 10 to 25 results in a decrease in the response variable by about 6.675 units. (p = **** highly significant)
- Salt 10 to Salt 30: The change in the response variable from Salt 10 to Salt 30 is estimated to be approximately -11.075 units. This suggests that increasing the salt level from 10 to 30 results in a decrease in the response variable by about 11.075 units.(p = **** highly significant)
- Salt 10 to Salt 35: The change in the response variable from Salt 10 to Salt 35 is estimated to be approximately -8.375 units. This indicates that increasing the salt level from 10 to 35 results in a decrease in the response variable by about 8.375 units. (p = **** highly significant)
- Salt 15 to Salt 20: The change in the response variable from Salt 15 to Salt 20 is estimated to be approximately -12.325 units. This suggests that increasing the salt level from 15 to 20 results in a decrease in the response variable by about 12.325 units. (p = **** highly significant)
- Salt 15 to Salt 25: The change in the response variable from Salt 15 to Salt 25 is estimated to be approximately -10.85 units. This indicates that increasing the salt level from 15 to 25 results in a decrease in the response variable by about 10.85 units. (p = **** highly significant)
- Salt 15 to Salt 30: The change in the response variable from Salt 15 to Salt 30 is estimated to be approximately -15.25 units. This suggests that increasing the salt level from 15 to 30 results in a decrease in the response variable by about 15.25 units. (p = **** highly significant)
- Salt 15 to Salt 35: The change in the response variable from Salt 15 to Salt 35 is estimated to be approximately -12.55 units. This indicates that increasing the salt level from 15 to 35 results in a decrease in the response variable by about 12.55 units. (p = **** highly significant)

kyle don't take points off for not reporting the p-value, you can just glace at them in the figure. I described each one and why since we are only looking at salt we don't care about block… although I added it in for good measure

## 0.2 Problem 2

a) Null and alternative hypotheses of your tests (2.5)

H_0: There is no significant effect of the dose of supplements or the type of supplement (Vitamin B and Zinc) on the thickness of pangolin scales

H_1: There is a significant effect of the dose of supplements or the type of supplement on the thickness of pangolin scales

b) Justification for your choice of test (2.5) ALL Figures labled “2.b” Figure 2.b Normality Assumption, shapiro test(p = .669) not significant, passes and assume normality. figure 2.b homogeneity of variance, the levene test shows (p = .14) which is not stat significant so we can assume normality of variances and are justified to continue with two way anova.

Figure 2.b and c

Interpretation: The interaction between dose and supplement type also shows a significant impact on scale thickness. The F-value of 4.107 with a small p-value (2.20e-02) suggests that the joint effect of dose and supplement type on scale thickness is significant. The generalized Eta Squared (ges) of 0.132 indicates that about 13.2% of the total variance in scale thickness can be attributed to the interaction between dose and supplement type. 77% of the variance in scale thickness can be attributed to dose and 22% can be attributed to dose.

c) A description of how you checked the assumptions of your statistical test (10)

- I also answered with some questions in “b” part of this question

I chose multiple comparison and simple main effects because if you have met the assumptions of the two-way ANOVA (e.g., homogeneity of variances), it is better to use the overall error term (from the two-way ANOVA) as input in the one-way ANOVA model. This will make it easier to detect any statistically significant differences if they exist (Keppel & Wickens, 2004; Maxwell & Delaney, 2004).

d) Results of your statistical test, interpreting your test in 2-3 sentences that include the appropriate reporting of the statistics as well as an appropriate figure (10)

Result of Anova Figure 2.d Anova - The two-way ANOVA conducted on pangolin scales thickness revealed significant main effects for both dose (F(2, 54) = 92.000, $p < .001$, $^2 = 0.773$) and supplement type (F(1, 54) = 15.572, p = .000231, $^2 = 0.224$). Additionally, a significant interaction effect was observed between dose and supplement type (F(2, 54) = 4.107, p = .022, $^2 = 0.132$), indicating that the relationship between dose and scales thickness varied depending on the supplement type administered. The asterisks (*) denote statistical significance at the $p < .05$ level. the output suggests that both the “dose” and “supp” factors have significant effects on pangolin scales thickness, and there is also a significant interaction effect between these two factors.

Figure 2.d Simple Main Effect - The effect of the Vitamin B supplement on scale thickness is highly significant. The F-value of 62.542 with an extremely small p-value (8.75e-15) suggests that the mean differences in scale thickness across different doses of the Vitamin B supplement are highly unlikely to occur due to random chance. The generalized Eta Squared (ges) of 0.698 indicates that about 69.8% of the total variance in scale thickness can be attributed to the effect of the Vitamin B supplement. - Similarly, the effect of the Zinc supplement on scale thickness is highly significant. The F-value of 33.565 with an extremely small p-value (3.36e-10) indicates that the mean differences in scale thickness across different doses of the Zinc supplement are highly unlikely to occur due to random chance. The generalized Eta Squared (ges) of 0.554 suggests that about 55.4% of the total variance in scale thickness can be attributed to the effect of the Zinc supplement.

2.d Multiple Comparison Test

- There is a significant difference in scale thickness between groups receiving 0.5 and 1 doses of the Vitamin B supplement. The estimated difference is 8.79 units, with a confidence interval between approximately 4.90 and 12.68 units. The p-value of 1.75e-05 indicates that this difference is highly unlikely to have occurred by chance alone
- there is a significant difference in scale thickness between groups receiving 0.5 and 2 doses of the Vitamin B supplement. The estimated difference is 18.16 units, with a confidence interval between approximately 14.27 and 22.05 units. The extremely small p-value of 1.66e-11 indicates the highly significant nature of this difference.

- There is a significant difference in scale thickness between groups receiving 1 and 2 doses of the Vitamin B supplement. The estimated difference is 9.37 units, with a confidence interval between approximately 5.48 and 13.26 units. The p-value of 6.61e-06 indicates high significance.
- There is a significant difference in scale thickness between groups receiving Dose 1 and Dose 2 of the Vitamin B supplement. The estimated difference is 9.37 units, with a confidence interval between approximately 5.48 and 13.26 units. The p-value of 6.61e-06 indicates that this difference is highly unlikely to have occurred by chance alone.
- There is a significant difference in scale thickness between groups receiving Dose 0.5 and Dose 1 of the Zinc supplement. The estimated difference is 9.47 units, with a confidence interval between approximately 5.31 and 13.63 units. The p-value of 1.58e-05 indicates high significance.

e) An interpretation of any necessary post-hoc tests (5)

- did not run post-hoc

## 0.3 Problem 3

(a) specify your hypotheses (5)

H_0: There is no difference in website loading times of the two providers H_a: There is a difference in the website loading times of the two providers

(b) check that the data meet the assumptions of the statistical test you plan to use (10)

- Figure 3.b Normality by Groups and Shapiro Wilks Both fail test shapiro wilks ($p = .015$) and ($p = .0019$) and pass equality of variance so we continue with parametric t test. The Levene's test is not significant ($p > 0.05$m, $p = 0.0719$). Therefore, we can assume the homogeneity of variances in the different groups. Continue with the parametric t test and we are assuming that the homogeneity of variance are equal.

Figure 3 Parametric t test ($p = .18$) gives non significant p value

(c) Test your hypotheses. If your data do not meet the assumptions, test your hypotheses both by transforming the data and by using a nonparametric approach (10) See all figures with 3.c

(d) Explain the results of your statistical tests. (10)

- See figure 3.d

## 0.4 Problem 4

(a) Hypothesize a relationship between your variables. Include the null and alternative hypothesis. (5)

Null hypothesis (H0): There is no relationship between GDP per capita (GDPpc) and the Yale Environmental Performance Index (EPI2018Score).

Alternative hypothesis (H1): There is a relationship between GDP per capita (GDPpc) and the Yale Environmental Performance Index (EPI2018Score).

(b) Calculate the correlation coefficient between the two variables. Interpret. (5)

Figure 4.b and c GGpairs and 4.b and c Coorelation

- .69 coefficient could be higher, more testing and transformation needed

(c) Develop and discuss the scatterplots between the two variables. Does the relationship look linear or linear-log (or something else)? Should you transform one or both variables? Include your scatterplots (original, transformed) and discussion.(10)

Linear Model 1 (Figure 4.c linear model 1 and 4.c plot linear model 1) show extensive skewness and kurtosis on GDPpc Figure 4.b and c Scatterplot, shows we only need to transform GDPpc because this has the largest kurtosis and skewness. 4.c Non-constant Error Variance Test Linear Model 1 [raw data with outliers] - P values (p = 4.6154e-12, 2.61e-17 and 2.83e-10) highly stat significant indicating heteroskedascitcity the results of both tests indicate that there is significant heteroscedasticity in the regression model, suggesting that the assumption of constant variance is violated. This implies that the variability of the residuals changes across different levels of GDP per capita, which could potentially affect the reliability of the regression model's estimates and inference. Transformed with figure 4.c log data

(d) Perform a simple linear regression analysis on the two variables (after transformation, if needed). Make sure that your x and y variables are correct. Write the model equation and interpret the coefficient for the slope (10)

Created simple linear model with log transformed data(Figure 4.d Linear Model Logged Data) Created equation, figure 4.d equation for slope of logged data

(e) Interpret p-value of the slope coefficient in terms of hypothesis above. (5)

Figure 4.e interpertation for equation above for every one percent change in GDP per capita the EPISCORE increases by 0.086 units.

(f) Interpret the R2 value (5) Shown in Figure Value is 0.6657, which means that approximately 66.57% of the variance in the Environmental Performance Index (EPI) can be explained by the variation in the log of GDP per capita (GDPpc).

## 0.5 Problem 5

(a) Null and alternative hypotheses of your model (5)

H_0: There is no significant effect of the predictor weather variables (temperature, solar radiation, wind speed) on ozone concentration or their corresponding interaction effects.

H_1: There is a significant effect of at least one weather variable on ozone concentration.

(b) Results of your statistical test, interpreting the fit of the selected model in 2-3 sentences that include the appropriate reporting of the statistics in a table (20) Figure 5.b first linear model and Figure 5.b VIF, our VIF all are under pass and pass test for multicollinearity, we will continue with cross validation Figure 5.b GVLMA assumptions fail

(c) An interpretation of the regression model coefficients (i.e., what do each of the main effect(s) and interaction effect(s) mean (10)

See figure 5.c exponentiated values of final linear model:

(d) A description of how you checked the assumptions of your statistical test (10)

Every time temperature increases by one unit ozone increases by 5.9%. Everytime wind increases by one unit ozone increases 5.64 %.

Interaction effects Temp:Wind: the relationship between temperature and ozone concentration is modified by wind speed. A negative coefficient (-0.1500873) indicates that the effect of temperature on ozone concentration decreases as wind speed increases, or vice versa. In other words, the impact of temperature on ozone concentration is tempered by wind speed. Every additional rise in unit temp decreases the positive/neg effect of wind on the dependent variable of ozone by -.15%

Interaction effect temp:rad: Similarly, this interaction effect suggests that the relationship between temperature and ozone concentration is influenced by radiation levels. A positive coefficient (0.0032171) indicates that the effect of temperature on ozone concentration increases with higher radiation levels, or vice versa. In other words, the impact of temperature on ozone concentration is enhanced by radiation levels. every additional in temp decreases the positive/neg effect of solar radiation on dependent variable of ozone by -0.00321%.

Steps for cross validation: From the cross validation and subsequent data wrangling we have determined several things: - Running the cross validation with the interactive effects on the logged ozone we were able to obtain more accurate results from the CV, with that and after removing outliers from that data frame, all assumptions passed in the GVLMA.

From these results we determine that the most accurate predictor variables are temp and wind with the most prominent interaction effects being temp:wind and temp:rad on the response variable of ozone.

(e) Graphs that depict (i) the relationship between the independent variable and dependent variable when a variable is only included as a main effect, (ii) the relationship between the interacting independent variables with the dependent variable (10)

Figure 5.e simple main effects Figure 5.e plot coefficients simple main effects and interaction effect plots Figure 5.e Interaction plots for interaction effects

## 0.6 Problem 6

a) What is the research question underpinning the study (10 points)

The main research question of the study is to identify the environmental factors that influence the distribution of coral reef fish species on a regional scale. The study aims to determine which remotely measured environmental variables are most influential in determining the distributions of coral reef fish species. This research is important for managing coral reef species effectively, as understanding the factors affecting their distribution can inform conservation and management efforts.

b) What type of model did the authors choose to use and why? (5 points)

The authors chose to use Generalized Linear Models (GLMs) for their species distribution models. GLMs are a statistical modeling framework that allows for the analysis of data with non-normal distributions, such as binomial data (presence/absence data in this case). There are several reasons why GLMs were chosen:

1.Binomial data: GLMs are well-suited for modeling binary response variables, which is appropriate for presence/absence data in this study.

2. Flexibility of GLMs offer flexibility in modeling complex ecological relationships realistically. They allow for the incorporation of multiple predictor variables to explain species distributions.

3. Strong statistical foundation: GLMs are based on a strong statistical foundation, making them reliable for analyzing ecological data.

4. Ease of interpretation: GLMs provide interpretable results, making it easier to understand the relationship between predictor variables and species distributions.

5. Transferability to GIS: Predictions resulting from GLMs can be easily transferred into Geographic Information Systems (GIS) for mapping, which is important for visualizing and analyzing spatial patterns of species distributions.

c) Describe the data that the authors used. What is the response variable? Explanatory variables? (10 points)

The authors used data collected from coral reef sites in Kimbe Bay, Papua New Guinea from the years 1994, 2001 and 2002

The response variable in their analysis is the presence or absence of each fish species at each surveyed site. This binary response variable indicates whether a particular species was observed at a given site during timed-swim visual observations.

The explanatory variables, also known as predictor variables, are the environmental factors that the authors hypothesized may influence the distribution of coral reef fish species. These variables were remotely measured and included:

Depth: The depth of the water at each surveyed site. Presence of a land-sea interface: This variable likely indicates the proximity of the reef site to the shoreline, potentially affecting factors such as water temperature and nutrient availability. Exposure: This variable refers to the degree of exposure of the reef site to prevailing wind directions. Distance to the nearest estuary: This variable measures the distance from each reef site to the nearest river estuary, which may influence water quality and sedimentation levels. These explanatory variables were chosen based on their potential ecological significance and their availability for remote measurement, which is essential for broad-scale conservation and management applications.

As they tested for these, they only found these important:

ocean charts, remote sensing and local expert knowledge

d) Interpret the equation and results from table 4 for Caesio lunaris. (10 points)

Here's the breakdown of the equation:

- ln(p/(1 - p)): This is the natural logarithm of the odds ratio of the probability of occurrence of Caesio lunaris. In logistic regression, the log odds of an event occurring are modeled as a linear combination of the predictor variables.

- -1.086: This is the intercept term of the equation. It represents the log odds of occurrence when all predictor variables are zero. In this case, it suggests the baseline log odds of occurrence for Caesio lunaris when Exposure and Depth are zero.

- 2.6 × Exposure: Exposure is one of the predictor variables. The coefficient 2.6 indicates the change in the log odds of occurrence of Caesio lunaris for a one-unit increase in Exposure, holding all other variables constant. A positive coefficient suggests that as Exposure increases, the log odds of occurrence of Caesio lunaris also increase.

- 0.003 × Depth: Depth is another predictor variable. The coefficient 0.003 indicates the change in the log odds of occurrence of Caesio lunaris for a one-unit increase in Depth, holding all other variables constant. Here, a positive coefficient suggests that as Depth increases, the log odds of occurrence of Caesio lunaris also increase, but to a lesser extent compared to Exposure.

In summary, the equation suggests that the probability of occurrence of Caesio lunaris is influenced by both Exposure and Depth. An increase in Exposure and Depth leads to an increase in the log odds of occurrence of Caesio lunaris, with Exposure having a stronger effect than Depth.

e) How did the authors compare models? (5 points)

The authors compared models using the Akaike Information Criterion (AIC). The AIC is a measure of the relative quality of a statistical model for a given set of data. It takes into account both the goodness of fit of the model and the complexity of the model, penalizing the latter to avoid overfitting.

In their study, the authors developed logistic regression models for presence/absence data of 227 fish species. They then tested the efficiency of these models by comparing their AIC values to those of hypothetical species with random distributions. By comparing the AIC values of real species with the AIC values of hypothetical species distributions, they were able to identify efficient models.

Efficiency in this context refers to how well the model describes the observed data while avoiding overfitting. Models with lower AIC values are considered to be more efficient, indicating a better balance between goodness of fit and model complexity.

The authors found efficient models for 118 species, most of which were highly habitat-specific. This comparison allowed them to identify the most effective models for predicting the distribution of coral reef fish species based on the selected predictor variables.

f) Discuss two of the limitations the authors identify in their analysis? Discuss one potential limitation they do not bring up. (10 points)

Two limitations identified by the authors in their analysis are:

1. Data Uncertainty: The authors acknowledge that the accuracy of the environmental and habitat data used in their study may vary. While they used local-scale data layers for Kimbe Bay from various sources, including printed marine charts, expert consultation, GIS analysis, and multivariate bioregionalization of coral data, they did not explicitly consider the uncertainty associated with these data layers. For example, exposure data were classified based on expert opinion rather than directly measured wave energy, which could introduce inaccuracies. Similarly, bathymetry data were obtained from remote imagery with coarse resolution, which may not accurately represent the underwater terrain. The authors note that future studies may require in-situ measurements to obtain more accurate exposure and bathymetry data.
2. Model Generalization: Another limitation mentioned by the authors is the challenge of applying their species distribution models to other regions or scales. The environmental and habitat data used in their study were specific to Kimbe Bay, and the models may not generalize well to other locations with different environmental conditions or spatial scales. For example, while the models were effective for predicting the distribution of coral reef fish species in Kimbe Bay, they may not be applicable to regions with different reef structures or anthropogenic impacts. This limitation highlights the need for caution when extrapolating the results of localized studies to broader spatial scales or different geographic regions.

One potential limitation that the authors do not explicitly address in their analysis is the assumption of spatial independence in the presence/absence data used to fit the logistic regression models. Spatial autocorrelation, where the presence or absence of a species at one site may be correlated with nearby sites due to shared environmental characteristics or dispersal processes, could violate this assumption. Ignoring spatial autocorrelation may lead to biased parameter estimates and inflated type I error rates. Therefore, future studies could benefit from incorporating spatial modeling techniques to account for potential spatial autocorrelation in the data.

#Problem 1 - 30

| salt | variable | n | mean | sd |
|---|---|---|---|---|
| 10 | biomass | 4 | 14.150 | 6.520 |
| 15 | biomass | 4 | 18.325 | 5.710 |
| 20 | biomass | 4 | 6.000 | 3.524 |
| 25 | biomass | 4 | 7.475 | 3.265 |
| 30 | biomass | 4 | 3.075 | 2.281 |
| 35 | biomass | 4 | 5.775 | 2.728 |

| salt | obs | block | biomass | is.outlier | is.extreme |
|---|---|---|---|---|---|
| 15 | 20 | 4 | 9.9 | TRUE | FALSE |
| 25 | 22 | 4 | 2.8 | TRUE | FALSE |

### 0.7 1.a H0 State H_0 and H_a

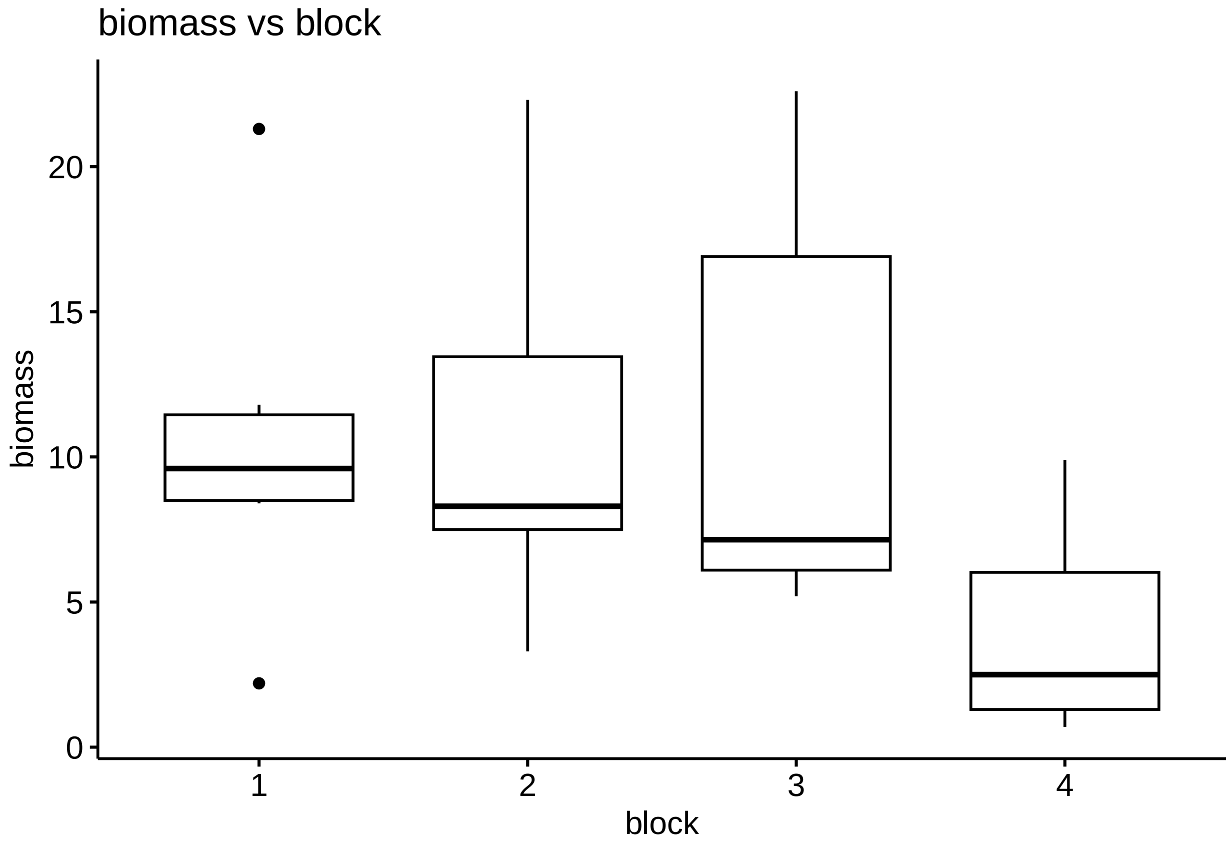
H_0_1: The biomass between groups stayed the same between salt exposed and non salt exposed block grouops

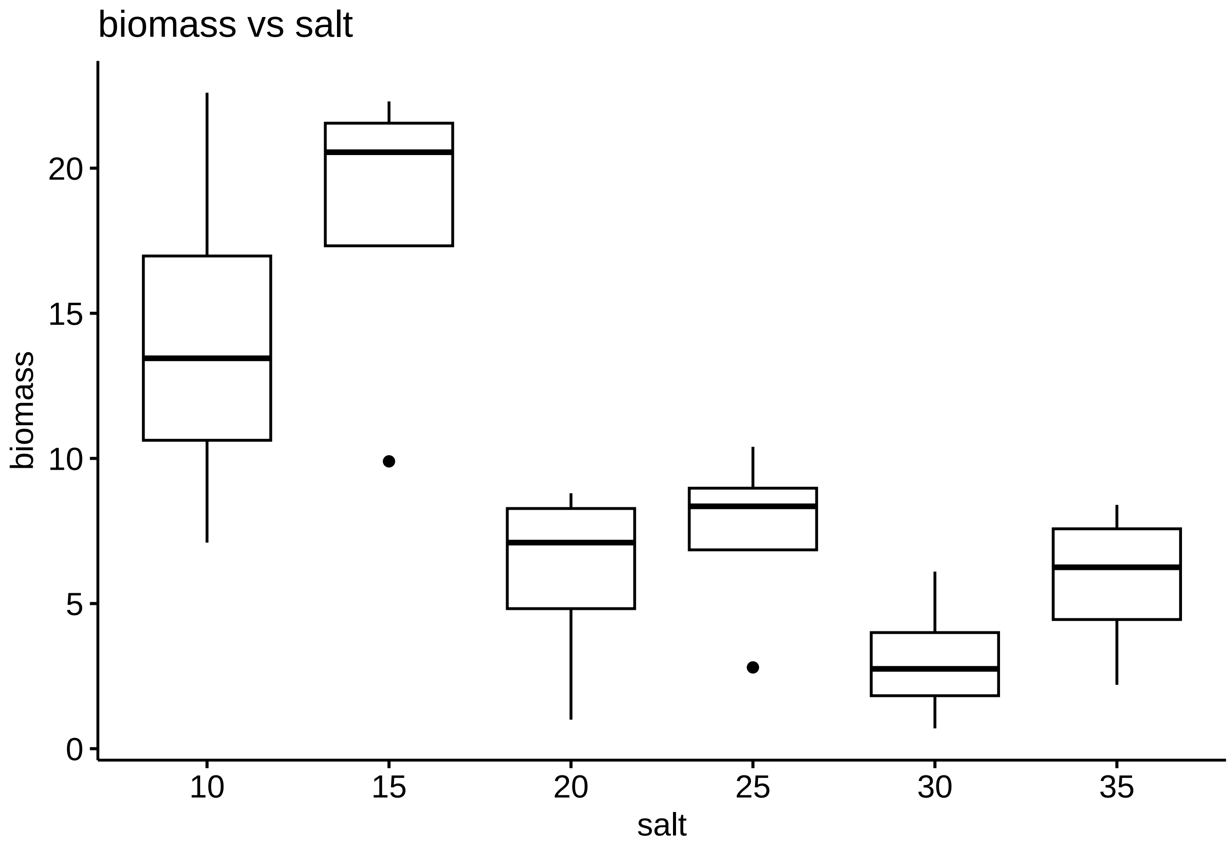
H_a_1: The biomass between groups changed between salt exposed and non salt exposed block groups

### 0.8 1.b Get summary statistics and visualize

### 0.9 1.b Identify outliers in each block

failed the outlier test, if other test are stat significant, take outlier out ## 1.b Boxplot

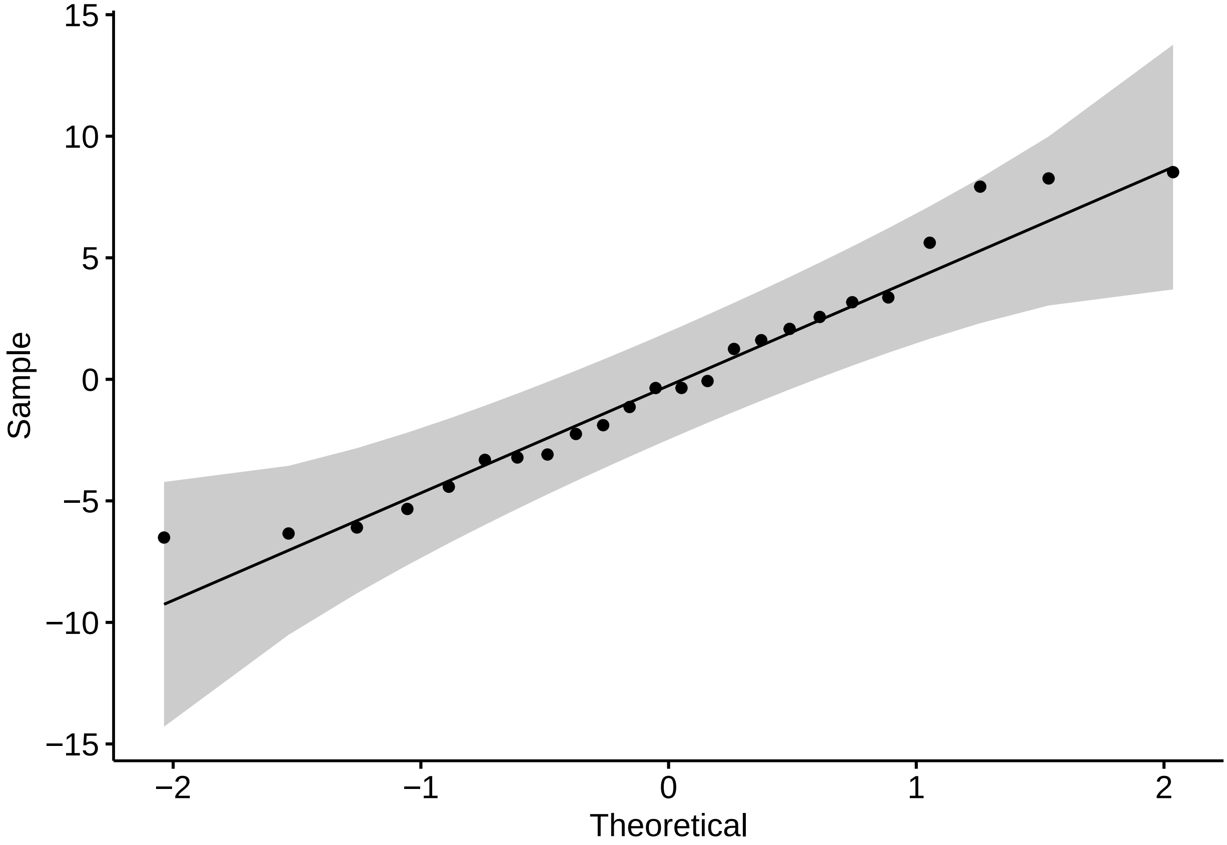



biomass vs salt
biomass
0
5
10
15
20
salt
10
15
20
25
30
35

| variable | statistic | p.value |
|---|---|---|
| residuals(model2) | 0.9478092 | 0.2427801 |

| df1 | df2 | statistic | p |
|---|---|---|---|
| 3 | 20 | 0.3397915 | 0.7967769 |

### 0.10 1.b Normality Assumption and Shapriro Test



answer to 1b.

Our value is NOT statistically significant from shapiro wilks (p = .24), we fail/ do not have sufficient evidence to reject the null(that its normal) therefore, we assume normality. In the QQ plot, as all the points fall approximately along the reference line, we can assume normality.

### 0.11 1.b Homogneity of variances

### 0.12 1.c Compute One Way ANVOA

ANOVA Table (type II tests)

| df1 | df2 | statistic | p |
|---|---|---|---|
| 5 | 18 | 0.6080903 | 0.6948354 |

| term | group1 | group2 | null.value | estimate | conf.low | conf.high | p.adj | p.adj.signif |
|---|---|---|---:|---:|---:|---:|---:|---|
| salt | 10 | 15 | 0 | 4.1750000 | -2.195427 | 10.5454267 | 3.24e-01 | ns |
| salt | 10 | 20 | 0 | -8.1500000 | -14.520427 | -1.7795733 | 8.94e-03 | ** |
| salt | 10 | 25 | 0 | -6.6750000 | -13.045427 | -0.3045733 | 3.74e-02 | * |
| salt | 10 | 30 | 0 | -11.0750000 | -17.445427 | -4.7045733 | 5.44e-04 | *** |
| salt | 10 | 35 | 0 | -8.3750000 | -14.745427 | -2.0045733 | 7.17e-03 | ** |
| salt | 15 | 20 | 0 | -12.3250000 | -18.695427 | -5.9545733 | 1.76e-04 | *** |
| salt | 15 | 25 | 0 | -10.8500000 | -17.220427 | -4.4795733 | 6.70e-04 | *** |
| salt | 15 | 30 | 0 | -15.2500000 | -21.620427 | -8.8795733 | 1.51e-05 | **** |
| salt | 15 | 35 | 0 | -12.5500000 | -18.920427 | -6.1795733 | 1.44e-04 | *** |
| salt | 20 | 25 | 0 | 1.4750000 | -4.895427 | 7.8454267 | 9.72e-01 | ns |
| salt | 20 | 30 | 0 | -2.9250000 | -9.295427 | 3.4454267 | 6.74e-01 | ns |
| salt | 20 | 35 | 0 | -0.2250000 | -6.595427 | 6.1454267 | 1.00e+00 | ns |
| salt | 25 | 30 | 0 | -4.4000000 | -10.770427 | 1.9704267 | 2.75e-01 | ns |
| salt | 25 | 35 | 0 | -1.7000000 | -8.070427 | 4.6704267 | 9.49e-01 | ns |
| salt | 30 | 35 | 0 | 2.7000000 | -3.670427 | 9.0704267 | 7.39e-01 | ns |
| block | 1 | 2 | 0 | 0.2833333 | -4.330839 | 4.8975057 | 9.98e-01 | ns |
| block | 1 | 3 | 0 | 0.8500000 | -3.764172 | 5.4641724 | 9.50e-01 | ns |
| block | 1 | 4 | 0 | -6.5333333 | -11.147506 | -1.9191609 | 4.85e-03 | ** |
| block | 2 | 3 | 0 | 0.5666667 | -4.047506 | 5.1808391 | 9.84e-01 | ns |
| block | 2 | 4 | 0 | -6.8166667 | -11.430839 | -2.2024943 | 3.43e-03 | ** |
| block | 3 | 4 | 0 | -7.3833333 | -11.997506 | -2.7691609 | 1.72e-03 | ** |

Effect SSn SSd DFn DFd F p p<.05 ges 1 salt 680.813 115.337 5 15 17.709 8.08e-06 * 0.855 2 block 217.183 115.337 3 15 9.415 9.60e-04 * 0.653 Df Sum Sq Mean Sq F value Pr(>F)
salt 1 424.1 424.1 15.84 0.000634 *** Residuals 22 589.2 26.8
— Signif. codes: 0 ‘***’ 0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

### 0.13 1 c d and e Post Hoc Test:

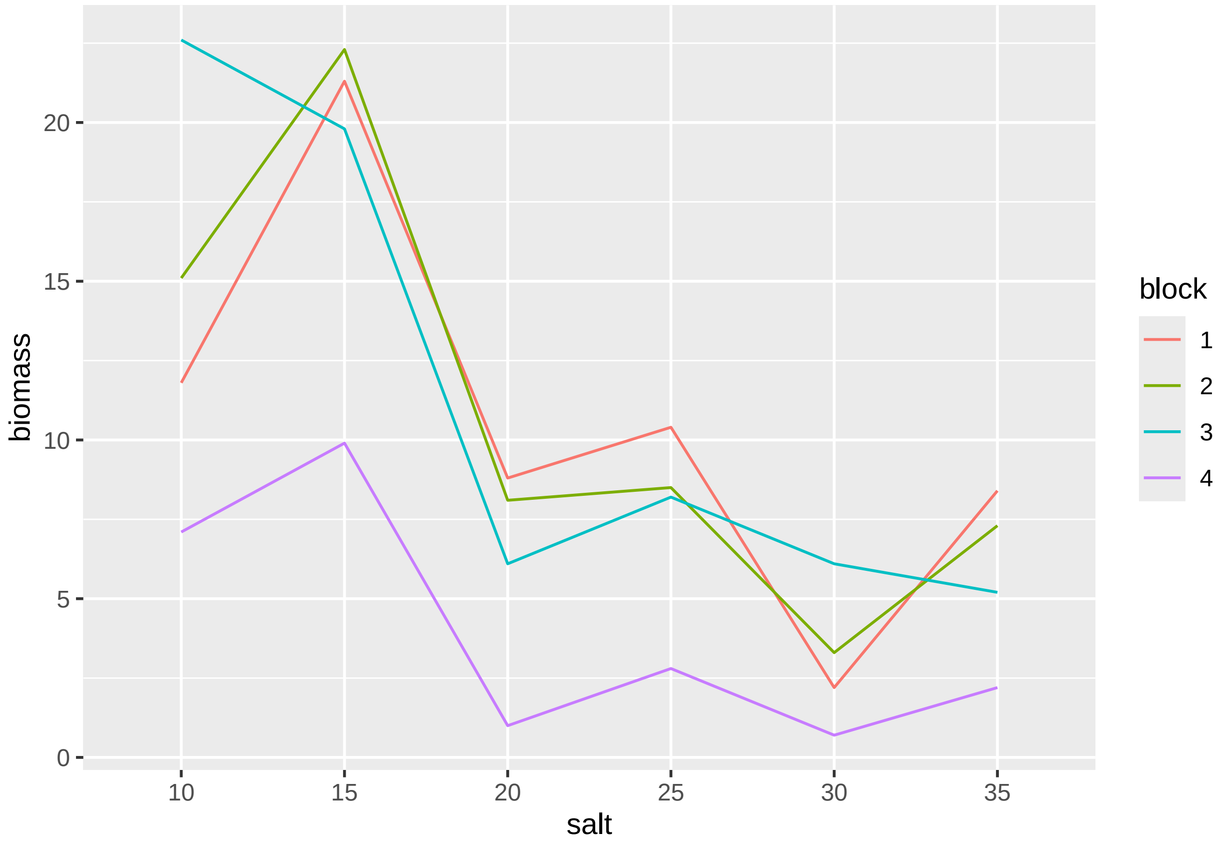
# 1 A tibble: 21 x 9

term group1 group2 null.value estimate conf.low conf.high p.adj * 1 salt 10 15 0 4.18 -2.20 10.5 0.324
2 salt 10 20 0 -8.15 -14.5 -1.78 0.00894
3 salt 10 25 0 -6.68 -13.0 -0.305 0.0374
4 salt 10 30 0 -11.1 -17.4 -4.70 0.000544 5 salt 10 35 0 -8.38 -14.7 -2.00 0.00717
6 salt 15 20 0 -12.3 -18.7 -5.95 0.000176 7 salt 15 25 0 -10.9 -17.2 -4.48 0.00067
8 salt 15 30 0 -15.3 -21.6 -8.88 0.0000151 9 salt 15 35 0 -12.6 -18.9 -6.18 0.000144 10 salt 20 25 0 1.47 -4.90 7.85 0.972
# i 11 more rows # i 1 more variable: p.adj.signif

| supp | dose | variable | n | mean | sd |
|---|---|---|---|---|---|
| VitB | 0.5 | thick | 10 | 7.98 | 2.747 |
| Zinc | 0.5 | thick | 10 | 13.23 | 4.460 |
| VitB | 1 | thick | 10 | 16.77 | 2.515 |
| Zinc | 1 | thick | 10 | 22.70 | 3.911 |
| VitB | 2 | thick | 10 | 26.14 | 4.798 |
| Zinc | 2 | thick | 10 | 26.06 | 2.655 |

| supp | dose | X | thick | is.outlier | is.extreme |
|---|---|---|---|---|---|
| VitB | 1 | 15 | 22.5 | TRUE | FALSE |
| Zinc | 2 | 56 | 30.9 | TRUE | FALSE |

## 1.1 1.d line plot



## 1.2 Problem 2

## 1.3 2.b Summary Statistics

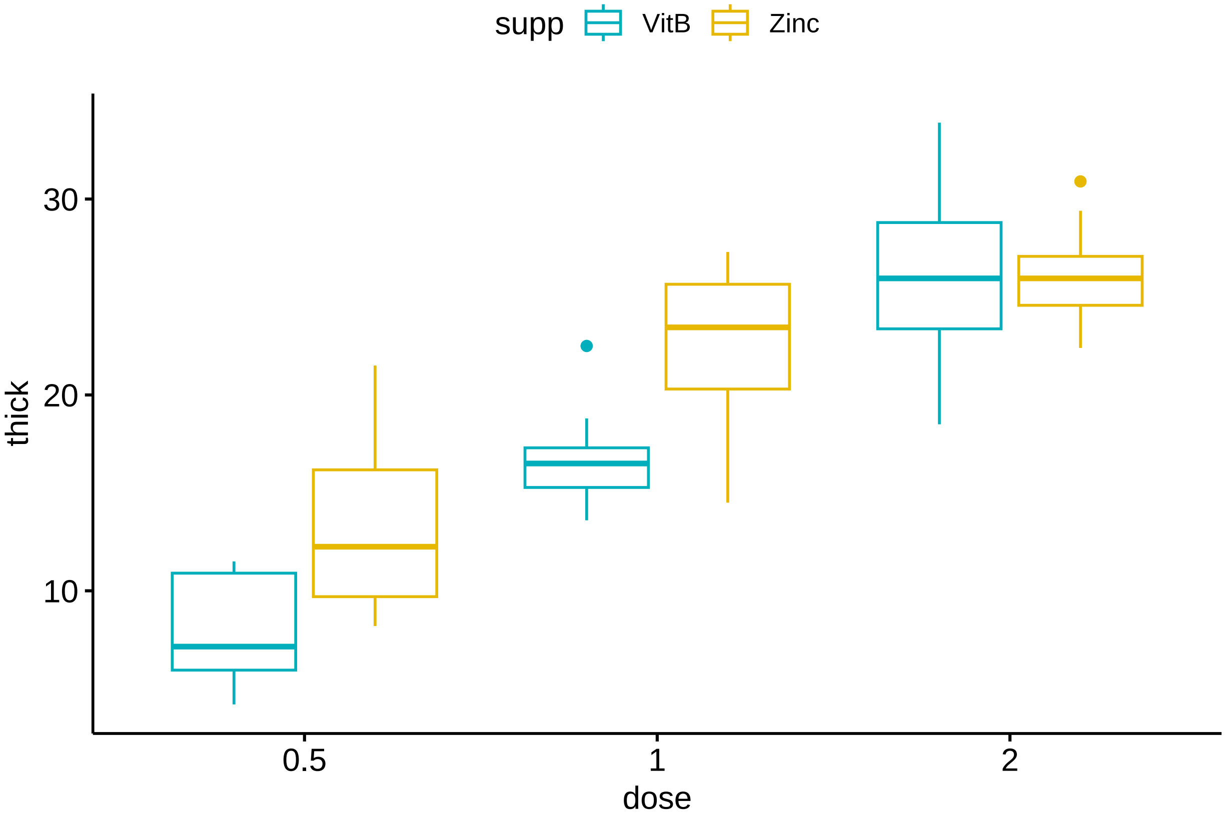
## 1.4 2.b Identify outliers in each block

failed the outlier test, if other test are stat significant, take outlier out ## 2 boxplot

### 1.5 Plotting the Data

## 2 A tibble: 6 x 6

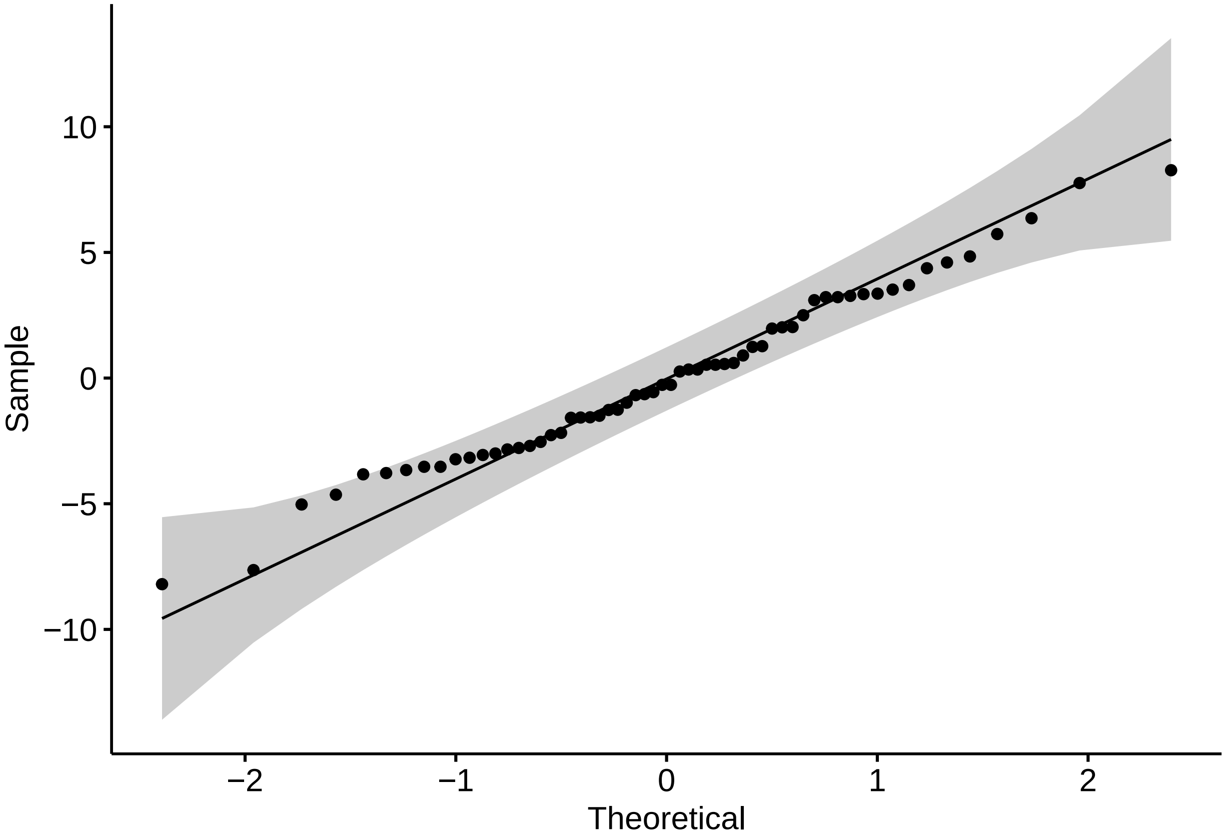
supp dose variable n mean sd 1 VitB 0.5 thick 10 7.98 2.75 2 VitB 1 thick 10 16.8 2.52 3 VitB 2 thick 10 26.1 4.80 4 Zinc 0.5 thick 10 13.2 4.46 5 Zinc 1 thick 10 22.7 3.91 6 Zinc 2 thick 10 26.1 2.66



| variable | statistic | p.value |
|---|---|---|
| residuals(model3) | 0.9849884 | 0.6694242 |

| df1 | df2 | statistic | p |
|---|---|---|---|
| 5 | 54 | 1.708578 | 0.1483606 |

## 2.1 2.b Normality Assumption



| Effect | DFn | DFd | F | p | p<.05 | ges |
|---|---|---|---|---|---|---|
| dose | 2 | 54 | 92.000 | 0.000000 | * | 0.773 |
| supp | 1 | 54 | 15.572 | 0.000231 | * | 0.224 |
| dose:supp | 2 | 54 | 4.107 | 0.022000 | * | 0.132 |

| supp | Effect | DFn | DFd | F | p | p<.05 | ges |
|---|---|---|---|---|---|---|---|
| VitB | dose | 2 | 54 | 62.542 | 0 | * | 0.698 |
| Zinc | dose | 2 | 54 | 33.565 | 0 | * | 0.554 |

| supp | term | group1 | group2 | null.value | estimate | conf.low | conf.high | p.adj | p.adj.signif |
|---|---|---|---|---|---|---|---|---|---|
| VitB | dose | 0.5 | 1 | 0 | 8.79 | 4.9017650 | 12.678235 | 1.75e-05 | **** |
| VitB | dose | 0.5 | 2 | 0 | 18.16 | 14.2717650 | 22.048235 | 0.00e+00 | **** |
| VitB | dose | 1 | 2 | 0 | 9.37 | 5.4817650 | 13.258235 | 6.60e-06 | **** |
| Zinc | dose | 0.5 | 1 | 0 | 9.47 | 5.3096046 | 13.630395 | 1.58e-05 | **** |
| Zinc | dose | 0.5 | 2 | 0 | 12.83 | 8.6696046 | 16.990395 | 1.00e-07 | **** |
| Zinc | dose | 1 | 2 | 0 | 3.36 | -0.8003954 | 7.520395 | 1.31e-01 | ns |

## 2.2 2.b Homogneity of Variance

## 2.3 2.d Compute two way anova

## 2.4 2.d Simple Main Effect

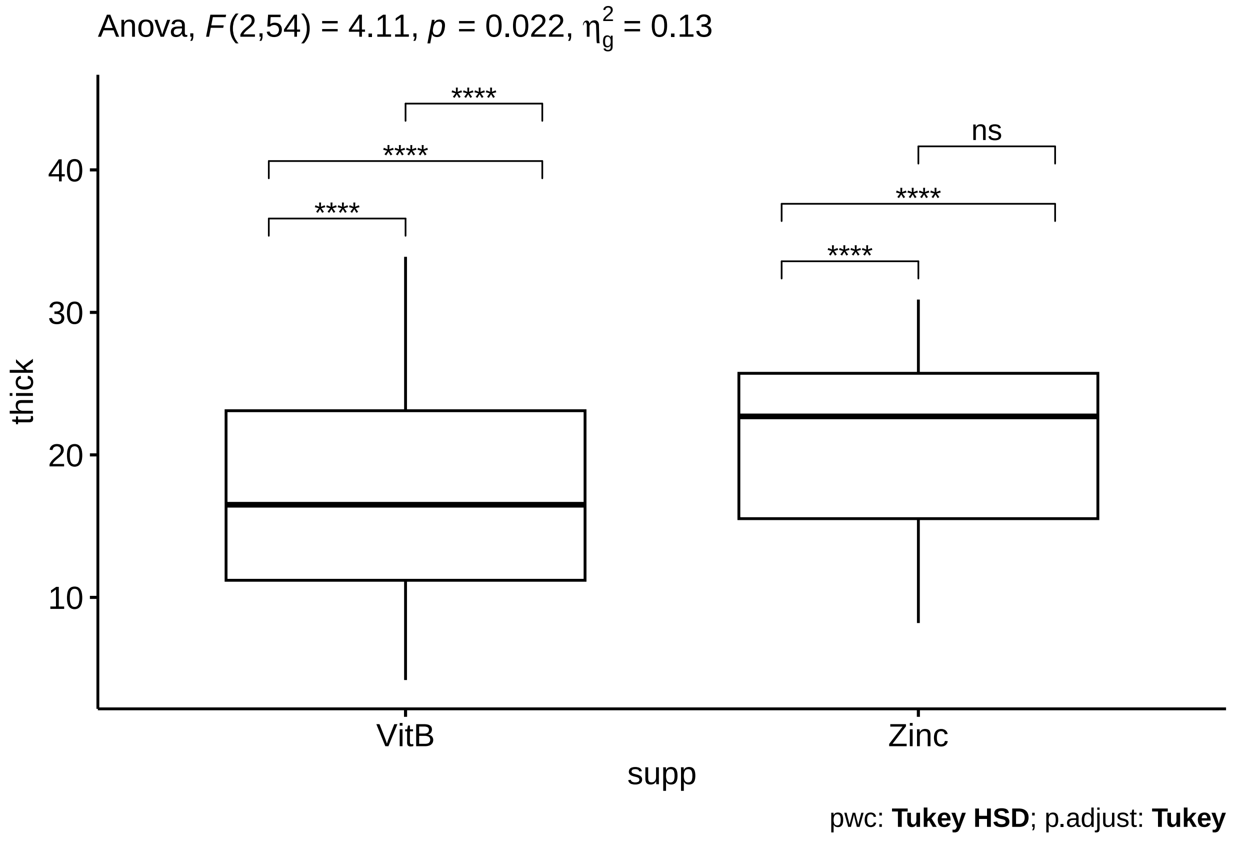
## 2.5 2.d Multiple Comparison Test

Interpretation: There is a significant difference in scale thickness between groups receiving 0.5 and 1 doses of the Vitamin B supplement. The estimated difference is 8.79 units, with a confidence interval between approximately 4.90 and 12.68 units. The p-value of 1.75e-05 indicates that this difference is highly unlikely to have occurred by chance alone Interpretation: Similarly, there is a significant difference in scale thickness between groups receiving 0.5 and 2 doses of the Vitamin B supplement. The estimated difference is 18.16 units, with a confidence interval between approximately 14.27 and 22.05 units. The extremely small p-value of 1.66e-11 indicates the highly significant nature of this difference. nterpretation: There is a significant difference in scale thickness between groups receiving 1 and 2 doses of the Vitamin B supplement. The estimated difference is 9.37 units, with a confidence interval between approximately 5.48 and 13.26 units. The p-value of 6.61e-06 indicates high significance. Iterpretation: There is a significant difference in scale thickness between groups receiving Dose 1 and Dose 2 of the Vitamin B supplement. The estimated difference is 9.37 units, with a confidence interval between approximately 5.48 and 13.26 units. The p-value of 6.61e-06 indicates that this difference is highly unlikely to have occurred by chance alone. Interpretation: There is a significant difference in scale thickness between groups receiving Dose 0.5 and Dose 1 of the Zinc supplement. The estimated difference is 9.47 units, with a confidence interval between approximately 5.31 and 13.63 units. The p-value of 1.58e-05 indicates high significance.

| term | group1 | group2 | null.value | estimate | conf.low | conf.high | p.adj | p.adj.signif |
|---|---|---|---:|---:|---:|---:|---:|---|
| supp | VitB | Zinc | 0 | 3.700 | 1.8201723 | 5.579828 | 2.31e-04 | *** |
| dose | 0.5 | 1 | 0 | 9.130 | 6.3624875 | 11.897512 | 0.00e+00 | **** |
| dose | 0.5 | 2 | 0 | 15.495 | 12.7274875 | 18.262512 | 0.00e+00 | **** |
| dose | 1 | 2 | 0 | 6.365 | 3.5974875 | 9.132513 | 2.70e-06 | **** |
| supp:dose | VitB:0.5 | Zinc:0.5 | 0 | 5.250 | 0.4518762 | 10.048124 | 2.43e-02 | * |
| supp:dose | VitB:0.5 | VitB:1 | 0 | 8.790 | 3.9918762 | 13.588124 | 2.10e-05 | **** |
| supp:dose | VitB:0.5 | Zinc:1 | 0 | 14.720 | 9.9218762 | 19.518124 | 0.00e+00 | **** |
| supp:dose | VitB:0.5 | VitB:2 | 0 | 18.160 | 13.3618762 | 22.958124 | 0.00e+00 | **** |
| supp:dose | VitB:0.5 | Zinc:2 | 0 | 18.080 | 13.2818762 | 22.878124 | 0.00e+00 | **** |
| supp:dose | Zinc:0.5 | VitB:1 | 0 | 3.540 | -1.2581238 | 8.338124 | 2.64e-01 | ns |
| supp:dose | Zinc:0.5 | Zinc:1 | 0 | 9.470 | 4.6718762 | 14.268124 | 4.60e-06 | **** |
| supp:dose | Zinc:0.5 | VitB:2 | 0 | 12.910 | 8.1118762 | 17.708124 | 0.00e+00 | **** |
| supp:dose | Zinc:0.5 | Zinc:2 | 0 | 12.830 | 8.0318762 | 17.628124 | 0.00e+00 | **** |
| supp:dose | VitB:1 | Zinc:1 | 0 | 5.930 | 1.1318762 | 10.728124 | 7.39e-03 | ** |
| supp:dose | VitB:1 | VitB:2 | 0 | 9.370 | 4.5718762 | 14.168124 | 5.80e-06 | **** |
| supp:dose | VitB:1 | Zinc:2 | 0 | 9.290 | 4.4918762 | 14.088124 | 6.90e-06 | **** |
| supp:dose | Zinc:1 | VitB:2 | 0 | 3.440 | -1.3581238 | 8.238124 | 2.94e-01 | ns |
| supp:dose | Zinc:1 | Zinc:2 | 0 | 3.360 | -1.4381238 | 8.158124 | 3.19e-01 | ns |
| supp:dose | VitB:2 | Zinc:2 | 0 | -0.080 | -4.8781238 | 4.718124 | 1.00e+00 | ns |

## 2.6 2 tukey post hoc test

## 2.7 2 Visualization: box plots with p-values



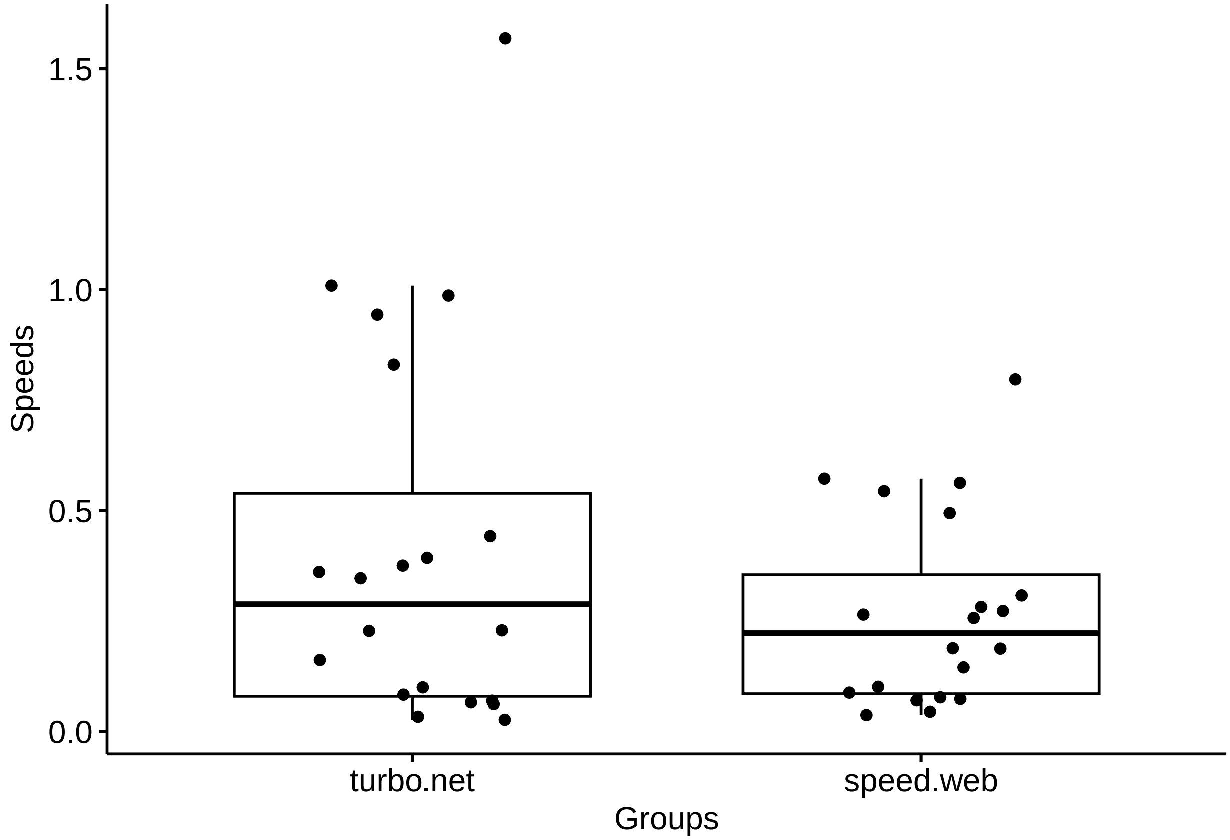
| group | variable | n | mean | sd |
|---|---|---|---|---|
| speed.web | speed | 20 | 0.268 | 0.217 |
| turbo.net | speed | 20 | 0.416 | 0.428 |

| group | id | speed | is.outlier | is.extreme |
|---|---|---|---|---|
| speed.web | 31 | 0.7968549 | TRUE | FALSE |
| turbo.net | 11 | 1.5691851 | TRUE | FALSE |

### 2.8 Problem 3

### 2.9 3 Summary Stats

### 2.10 3 Boxplot Data



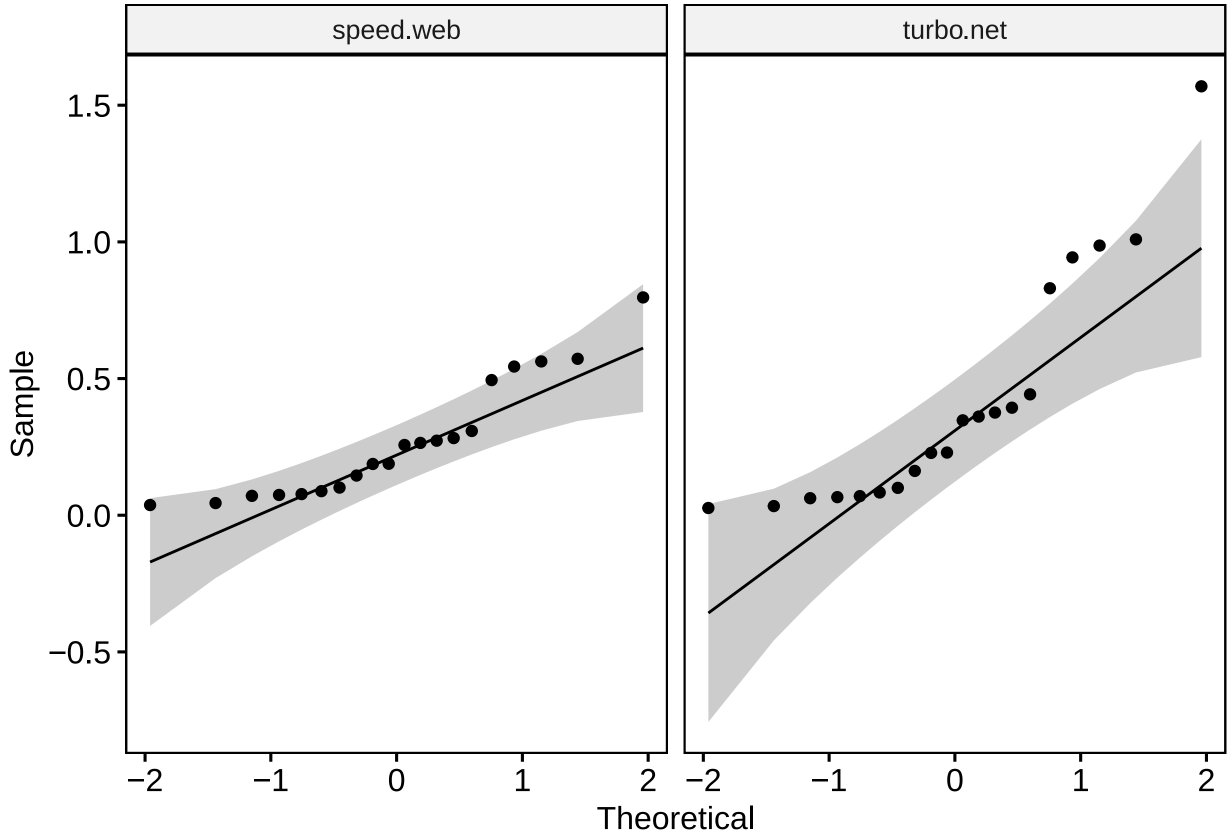
### 2.11 3 Identify outliers by groups

There were no extreme outliers.

| group | variable | statistic | p |
|---|---|---|---|
| speed.web | speed | 0.8762551 | 0.0151610 |
| turbo.net | speed | 0.8227558 | 0.0019263 |

| df1 | df2 | statistic | p |
|---|---|---|---|
| 1 | 38 | 3.427251 | 0.0719122 |

### 2.12 3.b Normality by Groups and Shapiro Wilks



failed both test(p = .015) for speedweb,(p=0.0019) for turbonet - need to log

### 2.13 3.b Equality of Variances

- pass so we continue with parametric (p = 0.07)

| estimate | estimate1 | estimate2 | .y. | group1 | group2 | n1 | n2 | statistic | p | df | conf.lo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| -0.1474817 | 0.2684941 | 0.4159759 | speed | speed.web | turbo.net | 20 | 20 | -1.375206 | 0.18 | 28.14879 | -0.367107 |

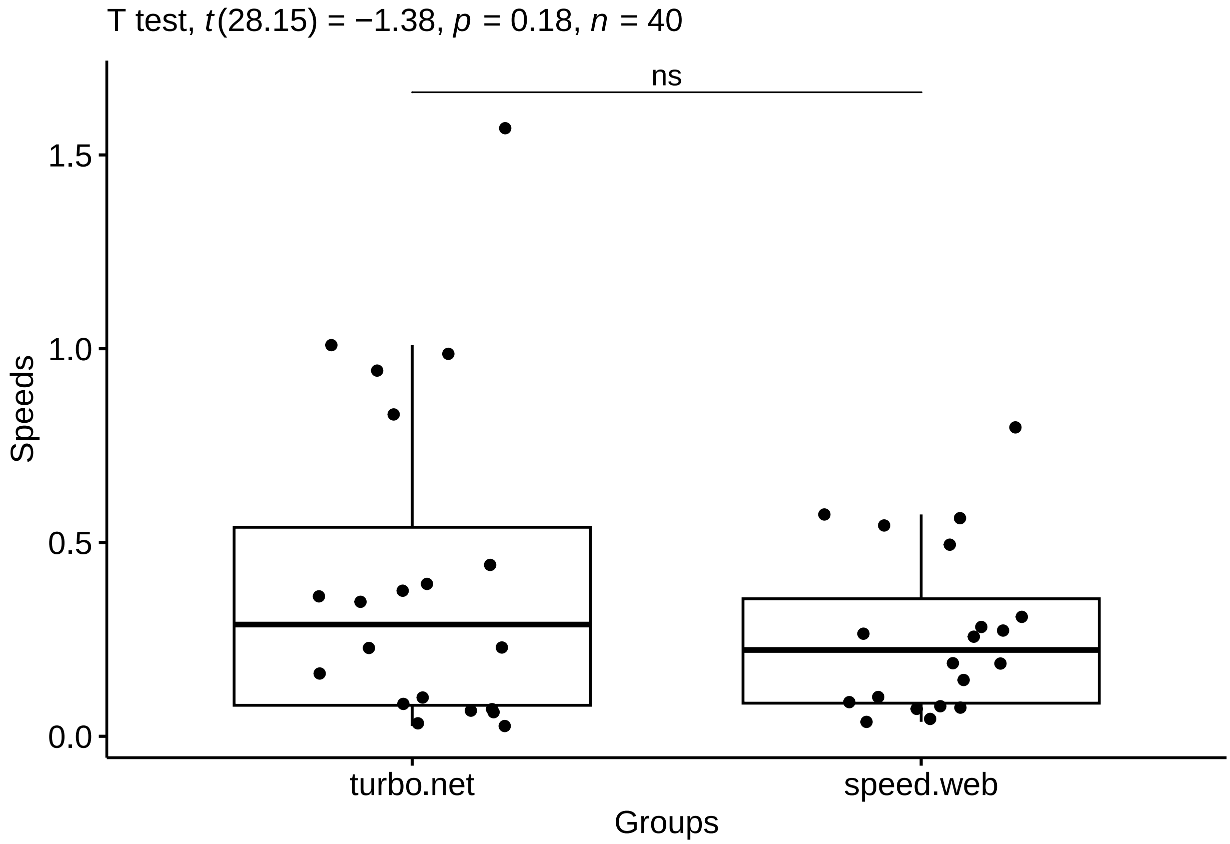
| .y. | group1 | group2 | effsize | n1 | n2 | magnitude |
|---|---|---|---|---|---|---|
| speed | speed.web | turbo.net | -0.4348784 | 20 | 20 | small |

| df1 | df2 | statistic | p |
|---|---|---|---|
| 1 | 38 | 1.831772 | 0.1839149 |

## 2.14 3 Parametric t test regular data

## 2.15 3 Effect Size

## 2.16 3 Report results parametric t test



## 2.17 3.c Log Transforming Data (1)

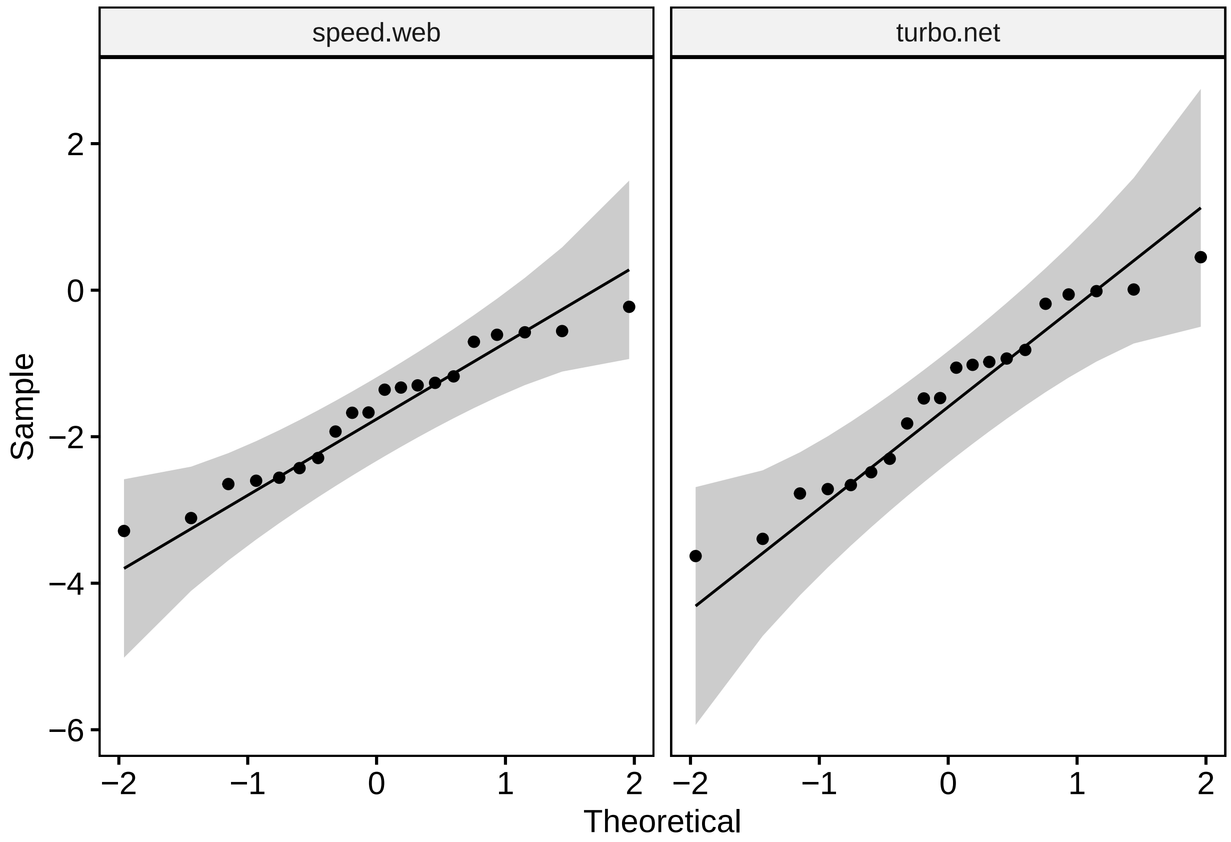
## 2.18 3.c Levine test log data

pass(p = .18)

| group | variable | statistic | p |
|---|---|---|---|
| speed.web | log.speed1 | 0.9523678 | 0.4044680 |
| turbo.net | log.speed1 | 0.9523678 | 0.4044681 |

| estimate | estimate1 | estimate2 | .y. | group1 | group2 | n1 | n2 | statistic | p | df | con |
|---|---|---|---|---|---|---|---|---|---|---|---|
| -0.1983263 | -1.665021 | -1.466695 | log.speed1 | speed.web | turbo.net | 20 | 20 | -0.5860788 | 0.562 | 35.23739 | -0.8 |

## 2.19 3.c Shapiro Test Logged



pass(p>.05) speed web(p=.4), turbonet(p-.4) pass and pass

variances are equal we conduct parametric t-test

paired t test

## 2.20 3.c Parametric t test logged

## 2.21 Confidence Interval

## 2.22 3.c Effect Size logged

pass effect size is negligible ## 3.c Non-parametric Logged wilcoxin test

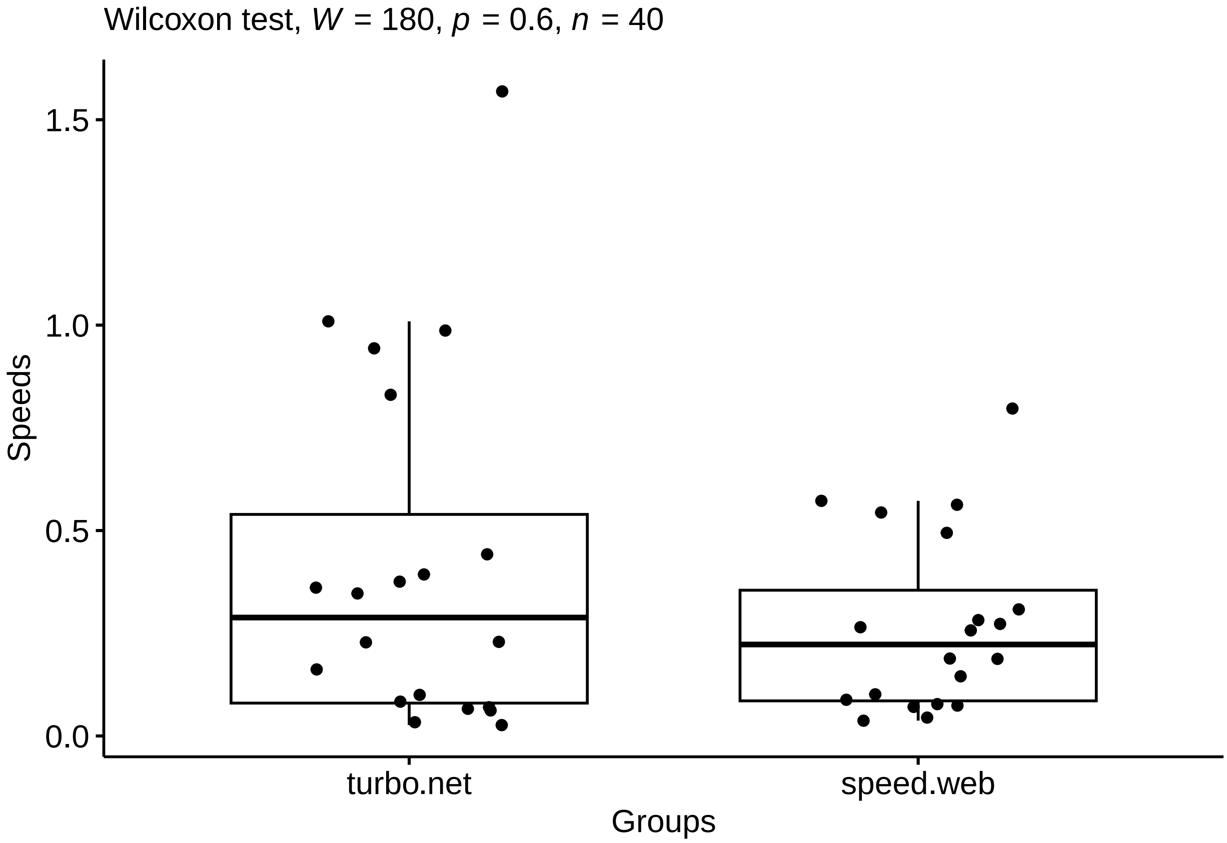
pass but we aren’t supposed to do this with our data

| group | conf_interval |
|---|---|
| speed.web | -2.0899827 |
| speed.web | -1.2400592 |
| turbo.net | -2.0333104 |
| turbo.net | -0.9000789 |

| .y. | group1 | group2 | effsize | n1 | n2 | magnitude |
|---|---|---|---|---|---|---|
| log.speed1 | speed.web | turbo.net | -0.1853344 | 20 | 20 | negligible |

### 2.23 3.c Effect Size non parametric

### 2.24 3.c Report Results Wilcoxin Test



### 2.25 3.d report results

the confidence intervals for the parametric t-test of the logged data are speed.web -2.0899827
speed.web -1.2400592
turbo.net -2.0333104
turbo.net -0.9000789

| .y. | group1 | group2 | n1 | n2 | statistic | p |
|---|---|---|---|---|---|---|
| log.speed1 | speed.web | turbo.net | 20 | 20 | 180 | 0.602 |

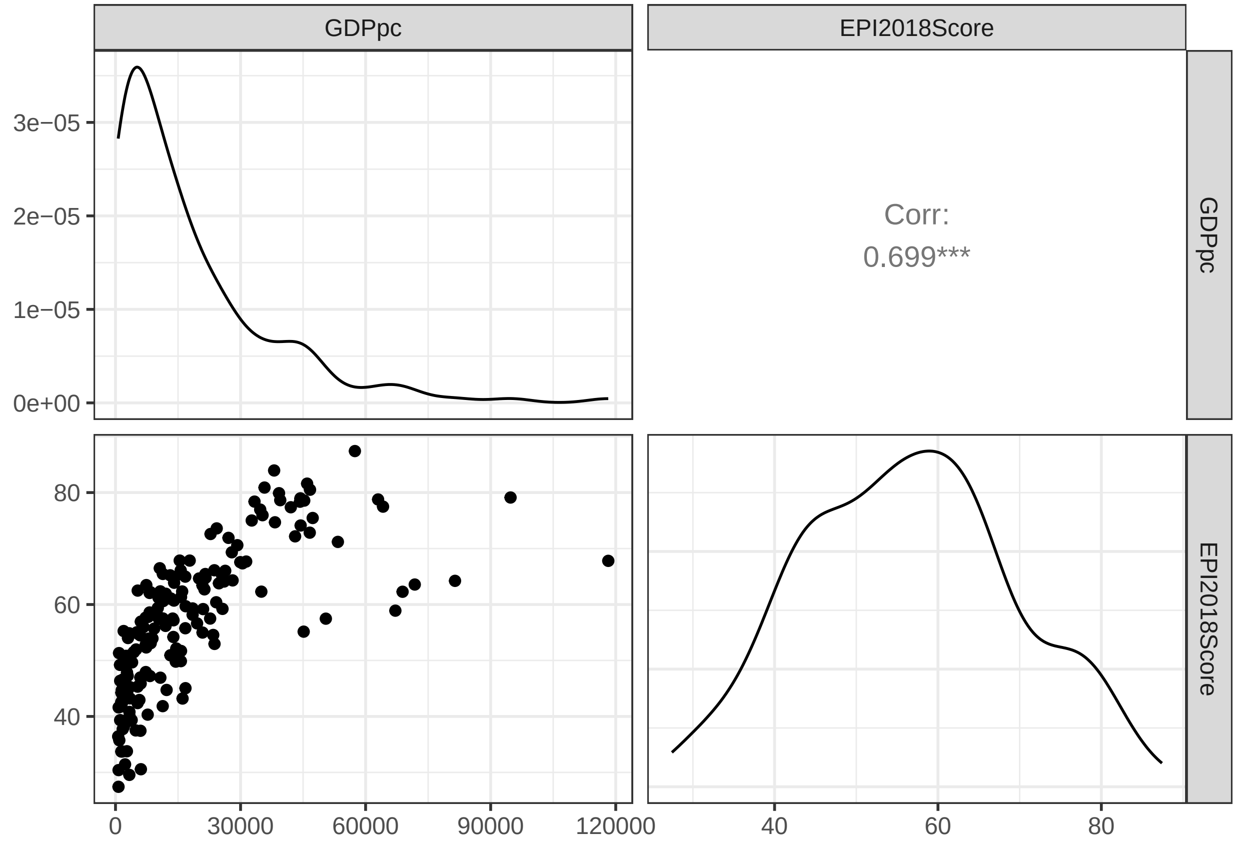
| .y. | group1 | group2 | effsize | n1 | n2 | magnitude |
|---|---|---|---:|---|---|---|
| log.speed1 | speed.web | turbo.net | 0.0855399 | 20 | 20 | small |

| x |
|---:|
| 0.6985719 |

the effect size is negligible and the p value is non significant(p=.56) which suggest that we fail to reject the null for the logged data and that the difference between the loading speed between turbo net and speed web does not change.

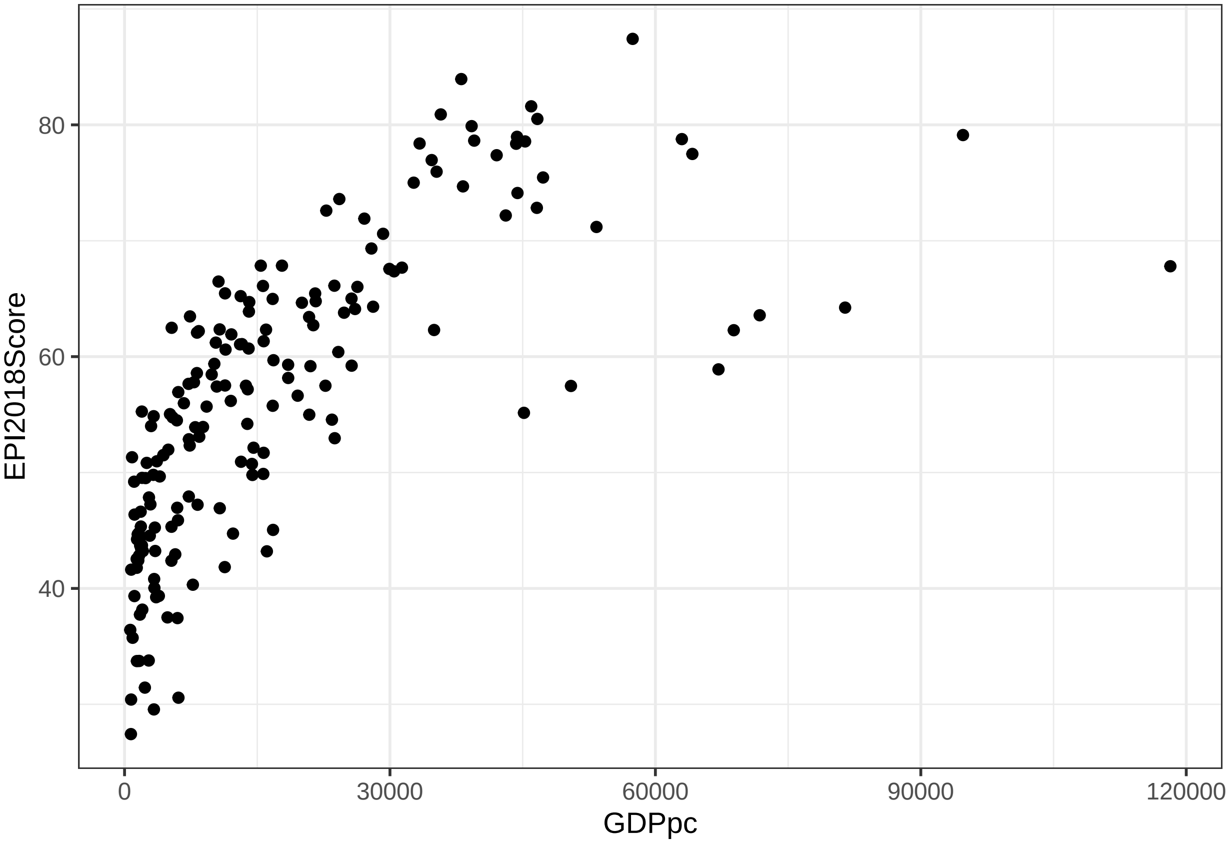
## 2.26 Problem 4

## 2.27 4.b and c GGpairs



- coorelation coeficcient .699 but data is heavily skewed and needs to be log transformed ## 4.b and c Coorelation # A tibble: 1 x 8 var1 var2 cor statistic p conf.low conf.high method
  1 GDPpc EPI2018Score 0.7 13.0 1.13e-27 0.615 0.767 Pearson

coorelation could be closer ## 4.b and c Scatterplot



scatterplot shows real heteroskedasticity

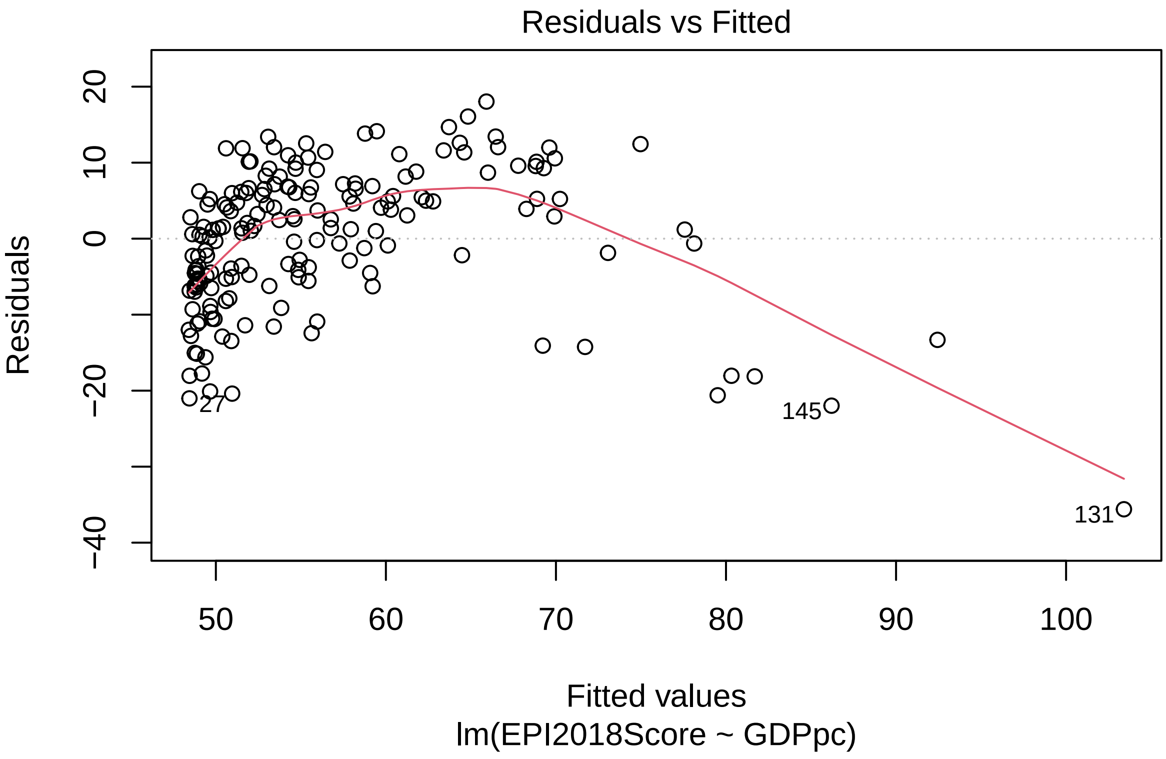
### 2.28 4.c Making Linear Model 1 [raw data with outliers]

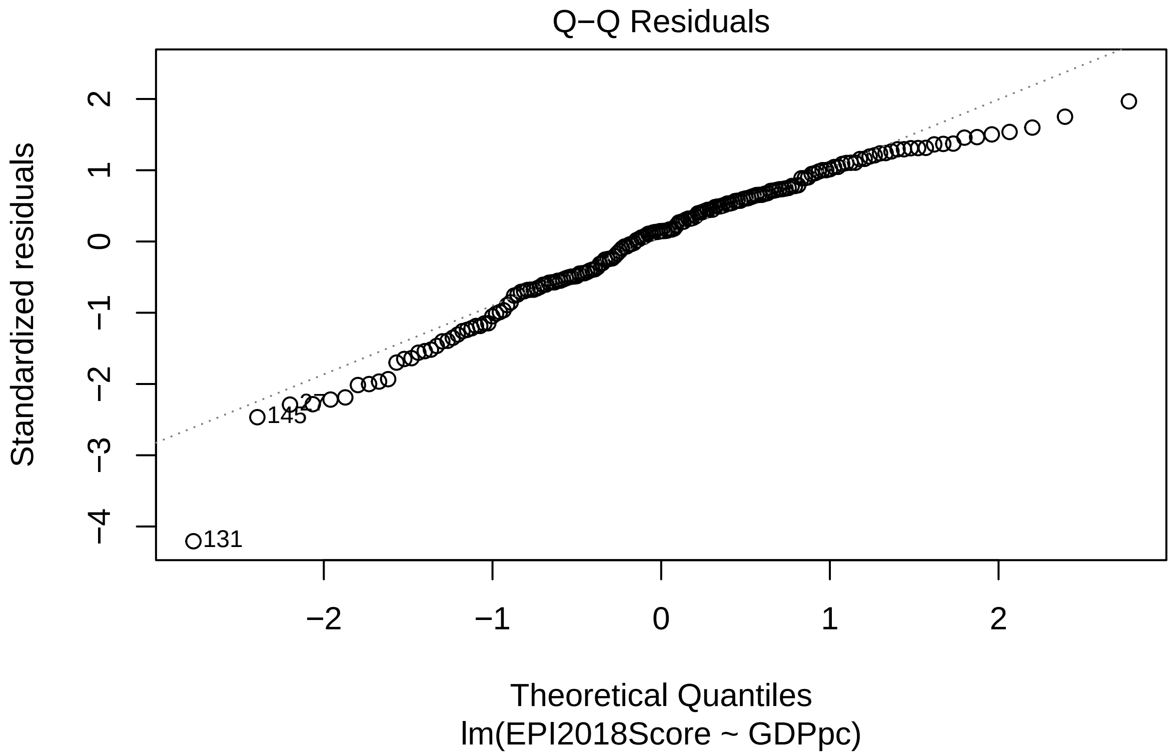
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 48.1085024 | 0.9375314 | 51.31402 | 0 |
| GDPpc | 0.0004678 | 0.0000359 | 13.02529 | 0 |

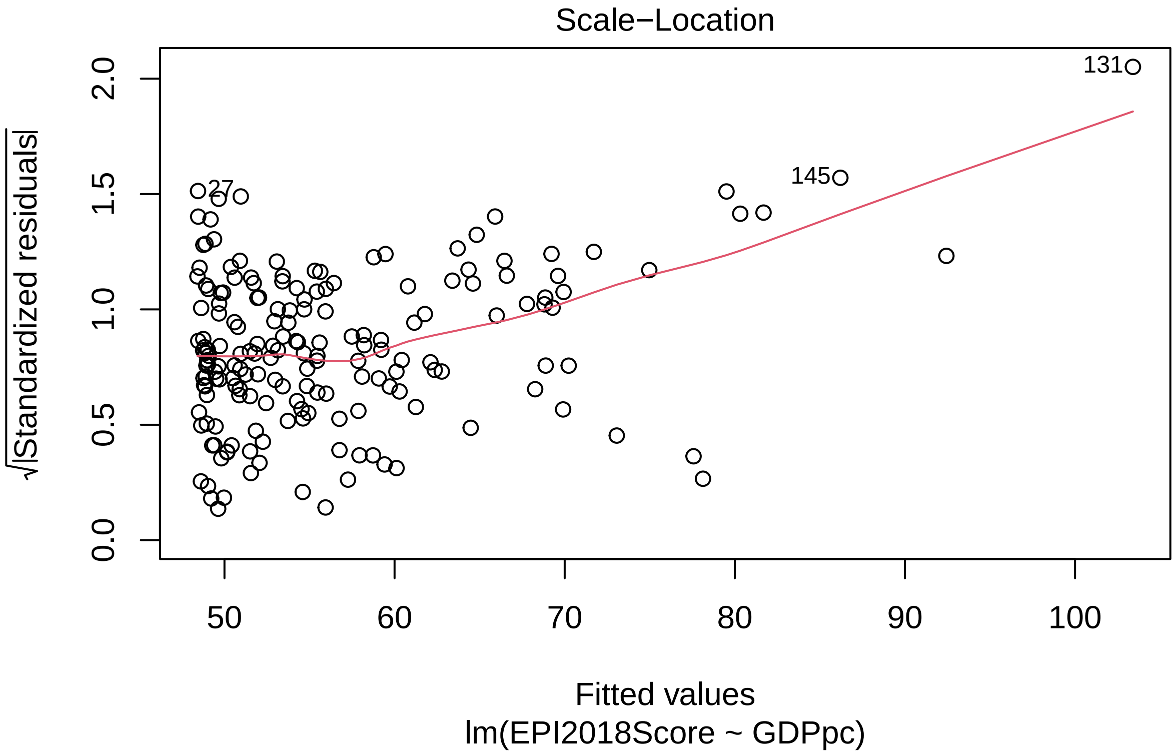
- The coefficient estimate for GDP per capita is 0.0004678, indicating that for every one-unit increase in GDP per capita, the EPI2018Score is expected to increase by approximately 0.0004678 units, holding other variables constant. Both the intercept and the coefficient for GDP per capita have extremely small p-values (close to zero), suggesting strong evidence against the null hypothesis that their coefficients are zero, indicating that both variables are statistically significant predictors of the Yale Environmental Performance Index. The t-values for both the intercept and GDP per capita are also notably high, indicating that their estimates are significantly different from zero, reinforcing their significance in predicting the EPI2018Score.

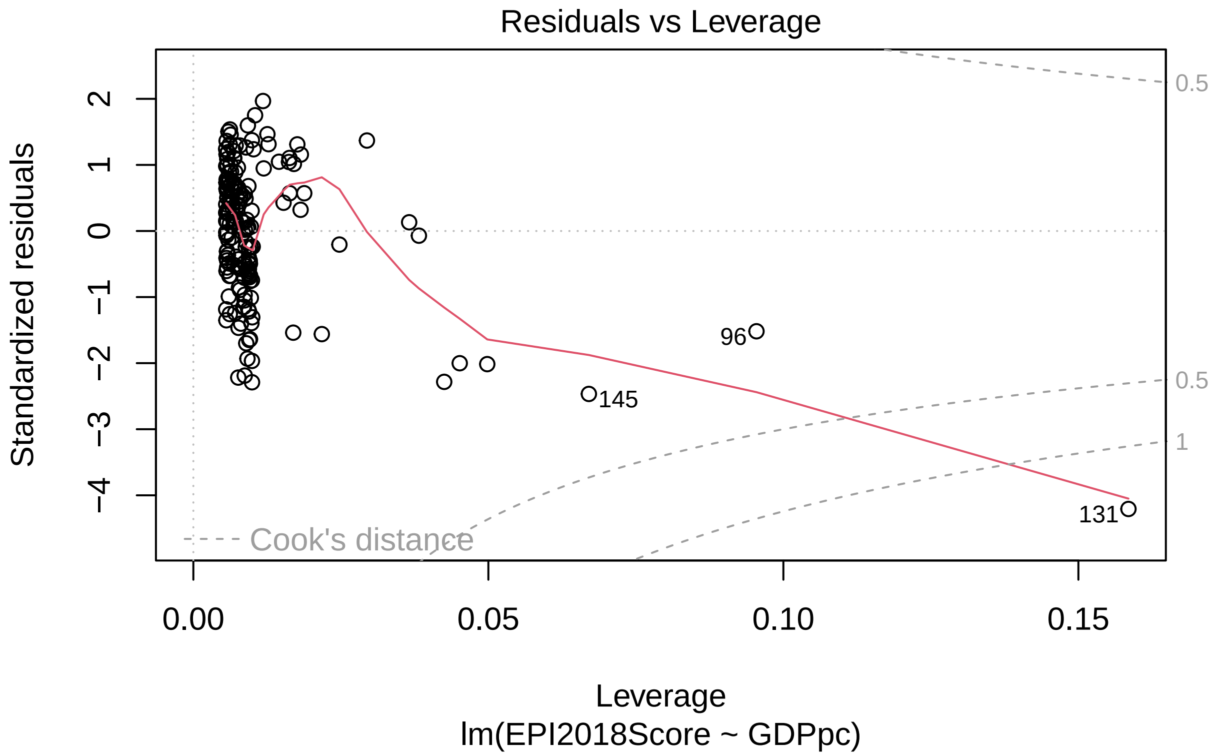
### 2.29 4.c Calculating prediction and confidence intervals

### 2.30 4.c Plot Linear Model 1 [raw data with outliers]




Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
−4
131
145
27
−2
−1
0
1
2
Theoretical Quantiles
lm(EPI2018Score ~ GDPpc)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
50
60
70
80
90
100
27
145
131
Fitted values
lm(EPI2018Score ~ GDPpc)



You can clearly see from these plots the skewness of the data, something that only a log transformation

## 2.31 4.c Non-constant Error Variance Test Linear Model 1 [raw data with outliers]

Non-constant Variance Score Test Variance formula: ~ fitted.values Chisquare = 47.84389, Df = 1, p = 4.6154e-12

```
Breusch-Pagan test
```

data: lm1 BP = 47.844, df = 1, p-value = 4.615e-12

```
studentized Breusch-Pagan test
```

data: lm1 BP = 39.782, df = 1, p-value = 2.839e-10 P values (p = 4.6154e-12, 2.61e-17 and 2.83e-10) highly stat significant indicating heteroskedascitcity the results of both tests indicate that there is significant heteroscedasticity in the regression model, suggesting that the assumption of constant variance is violated. This implies that the variability of the residuals changes across different levels of GDP per capita, which could potentially affect the reliability of the regression model's estimates and inference.

## 2.32 4.c Durbin Watson Test Linear Model 1 [raw data with outliers]

lag Autocorrelation D-W Statistic p-value 1 -0.08035092 2.150395 0.336 Alternative hypothesis: rho != 0 p value pass

## 2.33 4.c Shapiro Wilks Test Linear Model 1 [raw data with outliers]

```
Shapiro-Wilk normality test
```

data: residuals_lm1 W = 0.96707, p-value = 0.0002981 We fitted a linear model (estimated using OLS) to predict EPI2018Score with GDPpc (formula: EPI2018Score ~ GDPpc). The model explains a statistically significant and substantial proportion of variance (R2 = 0.49, F(1, 178) = 169.66, p < .001, adj. R2 = 0.49). The model's intercept, corresponding to GDPpc = 0, is at 48.11 (95% CI [46.26, 49.96], t(178) = 51.31, p < .001). Within this model:

- The effect of GDPpc is statistically significant and positive (beta = 4.68e-04, 95% CI [3.97e-04, 5.39e-04], t(178) = 13.03, p < .001; Std. beta = 0.70, 95% CI [0.59, 0.80])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

## 2.34 4.c Report Results Simple Linear Model 1 [raw data with outliers]

We fitted a linear model (estimated using OLS) to predict EPI2018Score with GDPpc (formula: EPI2018Score ~ GDPpc). The model explains a statistically significant and substantial proportion of variance (R2 = 0.49, F(1, 178) = 169.66, p < .001, adj. R2 = 0.49). The model's intercept, corresponding to GDPpc = 0, is at 48.11 (95% CI [46.26, 49.96], t(178) = 51.31, p < .001). Within this model:

- The effect of GDPpc is statistically significant and positive (beta = 4.68e-04, 95% CI [3.97e-04, 5.39e-04], t(178) = 13.03, p < .001; Std. beta = 0.70, 95% CI [0.59, 0.80])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

## 2.35 4.c logging data

## 2.36 4.d Linear Model Logged Data

Call: lm(formula = EPI2018Score ~ log_GDPpc, data = yaledat)

Residuals: Min 1Q Median 3Q Max -21.817 -5.354 1.398 4.613 16.046

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.2845 4.2689 -5.454 1.63e-07 ***log_GDPpc 8.6831 0.4612 18.828 < 2e-16*** — Signif. codes: 0 '***' **0.001** '**' *0.01* '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.455 on 178 degrees of freedom Multiple R-squared: 0.6657, Adjusted R-squared: 0.6639 F-statistic: 354.5 on 1 and 178 DF, p-value: < 2.2e-16

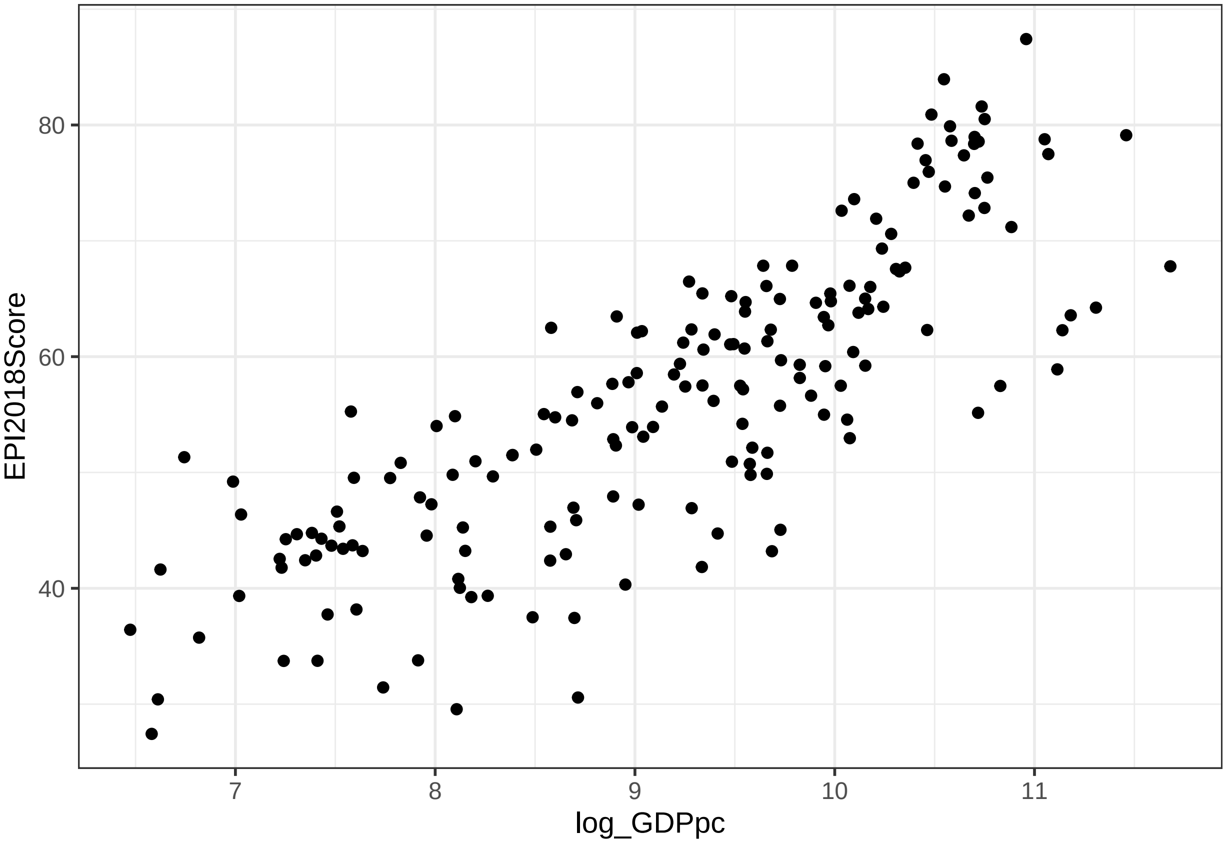
|  | x |
|---|---:|
| (Intercept) | -23.284513 |
| log_GDPpc | 8.683111 |

## 2.37 4.d Equation for slope of logged data

$$\hat{y_i} = -23.28 + .086 \times GDPpc$$

## 4.e interpertation for equation above

for every one percent change in GDP per capita the EPISCORE increases by 0.086 units. ## 4.c scatterplot logged data

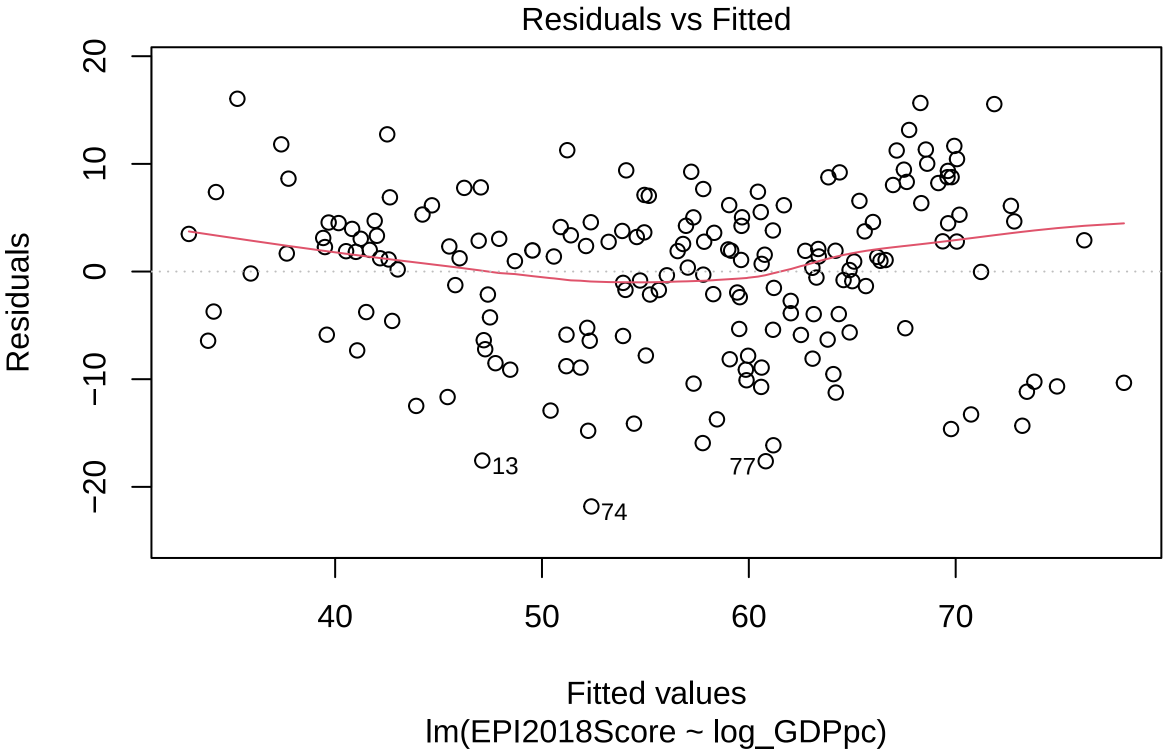


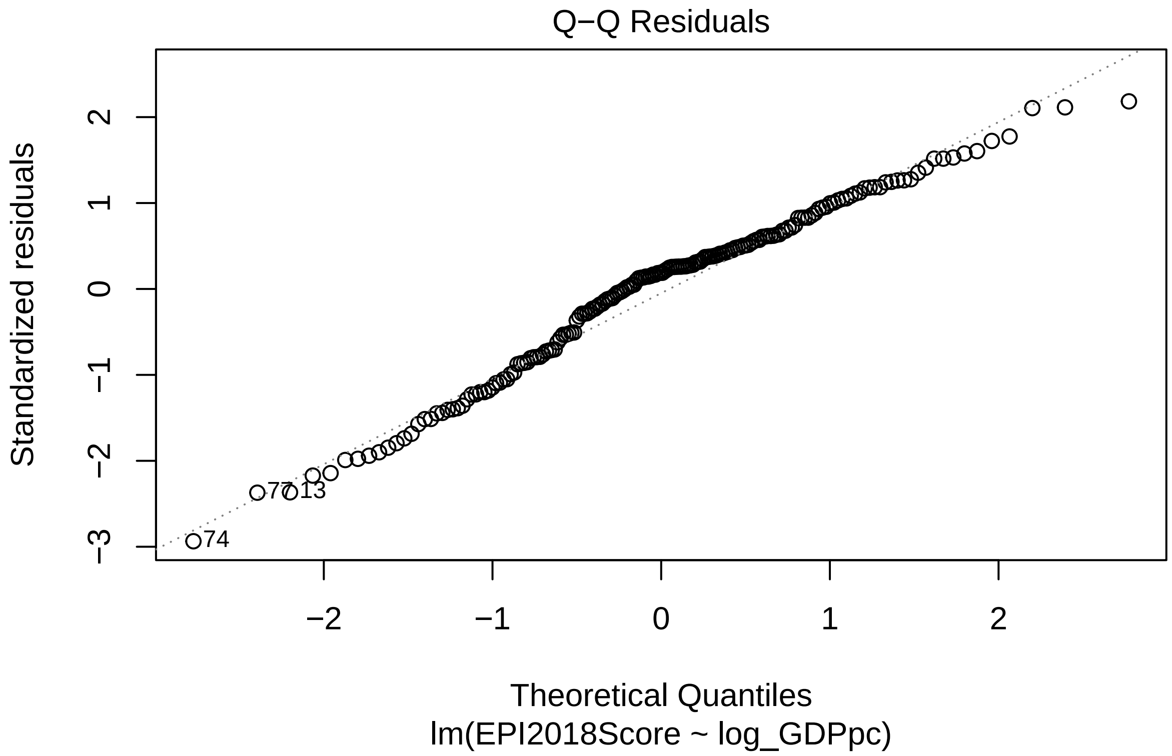
## 2.38 4.c Confidence interval Logged

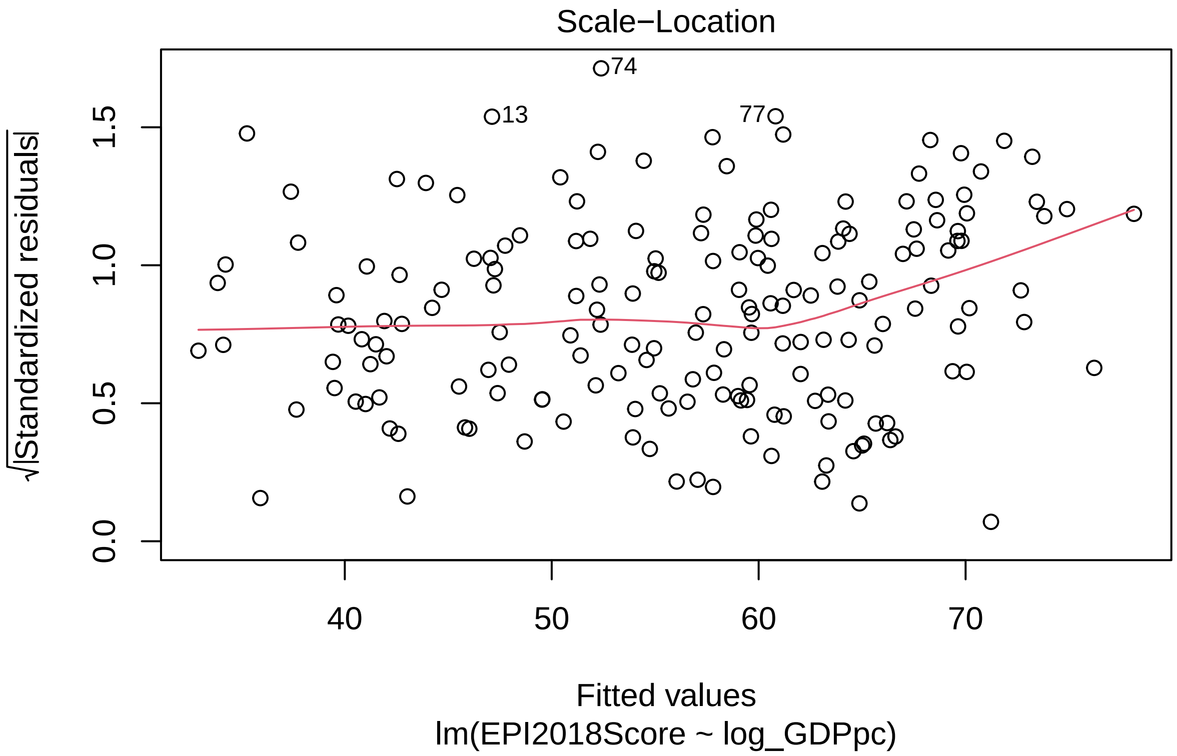
| | 2.5 % | 97.5 % |
|---|---|---|
| EPI2018Score | NA | NA |

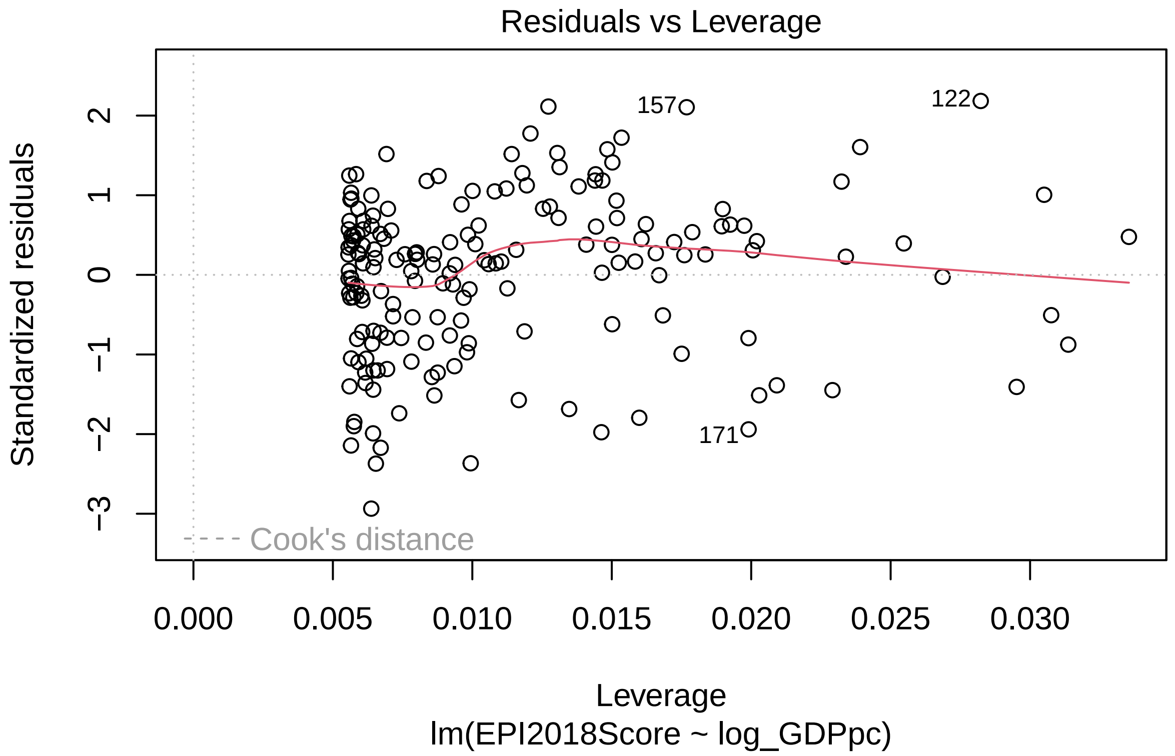
they have 97.5% and 2.5 percent confidence interval; look how pretty that looks!

## 2.39 4.c Plot [Logged comparison with outliers for comparisson]




Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
77
13
74
−2
−1
0
1
2
Theoretical Quantiles
lm(EPI2018Score ~ log_GDPpc)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
40
50
60
70
Fitted values
lm(EPI2018Score ~ log_GDPpc)
74
13
77



Looking much more fitted after logged transformed, significantly reduced heteroskedasticity

## 2.40 4.c [Logged comparison with outliers for comparisson]

Non-constant Variance Score Test Variance formula: ~ fitted.values Chisquare = 2.619678, Df = 1, p = 0.10555

```
Breusch-Pagan test
```

data: lm2_logged BP = 2.6197, df = 1, p-value = 0.1055

```
studentized Breusch-Pagan test
```

data: lm2_logged BP = 2.9208, df = 1, p-value = 0.08745 owever, for Model 2, there is no significant evidence of heteroscedasticity, suggesting that the assumption of constant variance holds for this model.(p = .06), (p = .1) (p= .06) all above .05! ## 4.c Cooks Distnace 1 2 3 4 5 6 7 8 9 10 11 12 13 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE 14 15 16 17 18 19 20 21 22 23 24 25 26 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE 27 28 29 30 31 32 33 34 35 36 37 38 39 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE 40 41 42 43 44 45 46 47 48 49 50 51 52 FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE 53 54 55 56 57 58 59 60 61 62 63 64 65 FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE 66 67 68 69 70 71 72 73 74 75 76 77 78 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE 79 80 81 82 83 84 85 86 87 88 89 90 91 FALSE FALSE FALSE

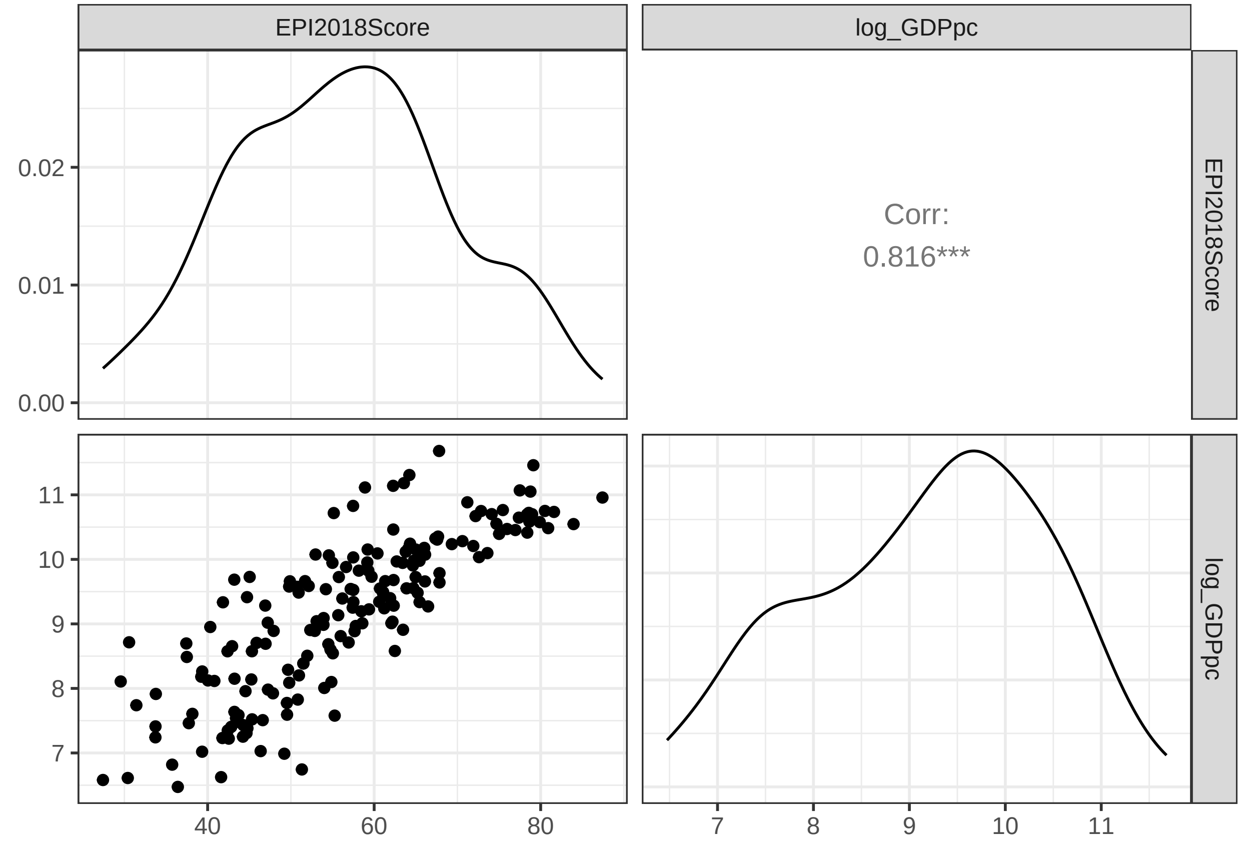
FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE 92 93 94 95 96 97 98 99 100 101 102 103 104 FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE 105 106 107 108 109 110 111 112 113 114 115 116 117 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE 118 119 120 121 122 123 124 125 126 127 128 129 130 FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE 131 132 133 134 135 136 137 138 139 140 141 142 143 TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE 144 145 146 147 148 149 150 151 152 153 154 155 156 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE 157 158 159 160 161 162 163 164 165 166 167 168 169 TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE 170 171 172 173 174 175 176 177 178 179 180 FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE 12 13 24 45 58 74 77 0.02900325 0.02811810 0.02056972 0.01872446 0.02879247 0.02766095 0.01851294 86 87 99 103 115 122 131 0.02312832 0.02370142 0.03151576 0.01926446 0.01939926 0.06925773 0.03012524 140 145 157 171 0.02617492 0.02459143 0.03988811 0.03825439

## 2.41 4 [Logged comparison with outliers for comparisson]

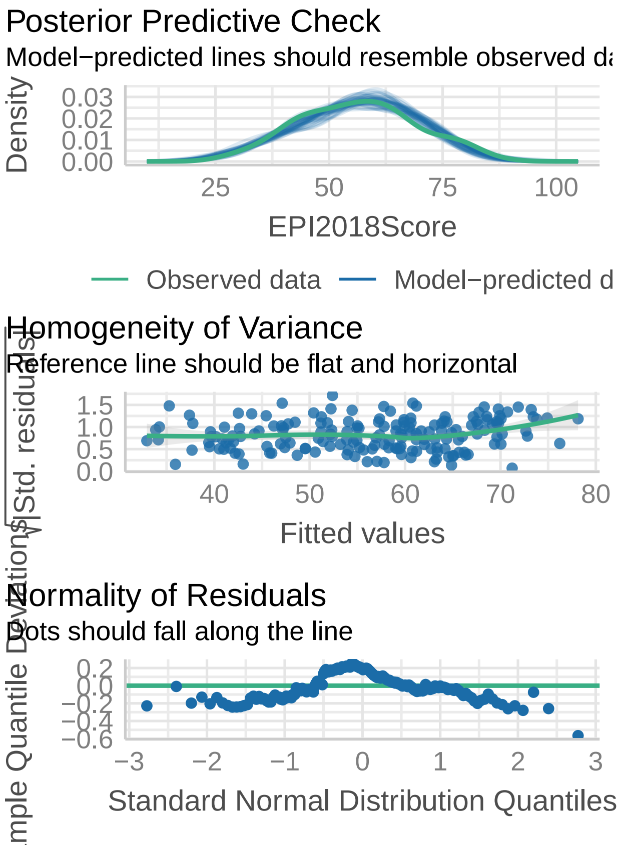
lag Autocorrelation D-W Statistic p-value 1 -0.05294 2.104294 0.488 Alternative hypothesis: rho != 0

the p-value is 0.302, which is greater than 0.05. Therefore, there is insufficient evidence to reject the null hypothesis, suggesting that there is no significant autocorrelation in the residuals at lag 1.

In this case, the p-value is very small (p = .3 ). This suggests strong evidence against the null hypothesis of normality. Typically, if the p-value is more than a chosen significance level (such as 0.05), we fail to reject the null hypothesis. So, in this case, we would fail to reject the null hypothesis and conclude that the residuals are normally distributed. pass



passed with much higher coorelation.



## 2.42 4.f R^2 interpertation:

Call: lm(formula = EPI2018Score ~ log_GDPpc, data = yaledat)

Residuals: Min 1Q Median 3Q Max -21.817 -5.354 1.398 4.613 16.046

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.2845 4.2689 -5.454 1.63e-07 ***log_GDPpc 8.6831 0.4612 18.828 < 2e-16*** — Signif. codes: 0 ‘***’ 0.001 ’**’ 0.01 ’*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.455 on 178 degrees of freedom Multiple R-squared: 0.6657, Adjusted R-squared: 0.6639 F-statistic: 354.5 on 1 and 178 DF, p-value: < 2.2e-16 R^2 value is 0.6657, which means that approximately 66.57% of the variance in the Environmental Performance Index (EPI) can be explained by the variation in the log of GDP per capita (GDPpc).

slope interpertation: for every one percent change in GDP per capita the EPISCORE increases by 0.086 units.

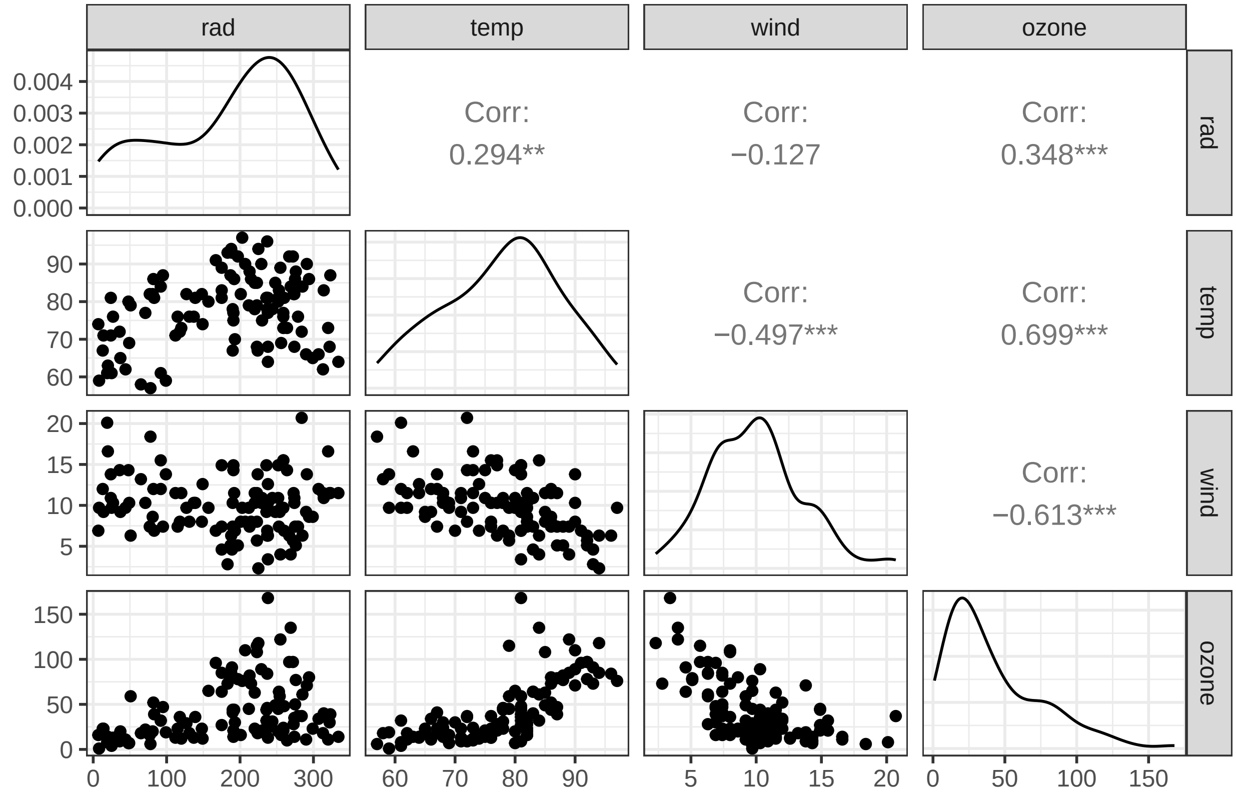
We fitted a linear model (estimated using OLS) to predict log_EPI2018Score with log_GDPpc (formula: log_EPI2018Score ~ log_GDPpc). The model explains a statistically significant and substantial proportion of variance (R2 = 0.66, F(1, 178) = 340.35, p < .001, adj. R2 = 0.65). The model's intercept, corresponding to log_GDPpc = 0, is at 2.54 (95% CI [2.38, 2.70], t(178) = 31.66, p < .001). Within this model:

- The effect of log GDPpc is statistically significant and positive (beta = 0.16, 95% CI [0.14, 0.18], t(178) = 18.45, p < .001; Std. beta = 0.81, 95% CI [0.72, 0.90])

So it is clear that the linear correlation coefficient is statisticly significant showing that we can assume that the linear coorelation value of .8 is significant, assuming that GDP has significant influence on the change in EPIscore in the cooresponding data frame. We will reject the null hypothesis and accept the alternative hypothesis that GDP has an effect on EPISCore and changes EPIscore

### 2.43 problem 5

### 2.44 5.b GGpairs



### 2.45 5.b first linear model

Call: lm(formula = ozone ~ rad + temp + wind, data = ozonedat)

Residuals: Min 1Q Median 3Q Max -40.485 -14.210 -3.556 10.124 95.600

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.23208 23.04204 -2.788 0.00628 ** rad 0.05980 0.02318 2.580 0.01124 *
temp 1.65121 0.25341 6.516 2.43e-09 ***wind -3.33760 0.65384 -5.105 1.45e-06*** — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 21.17 on 107 degrees of freedom Multiple R-squared: 0.6062, Adjusted R-squared: 0.5952 F-statistic: 54.91 on 3 and 107 DF, p-value: < 2.2e-16

We fitted a linear model (estimated using OLS) to predict ozone with rad, temp and wind (formula: ozone ~ rad + temp + wind). The model explains a statistically significant and substantial proportion of variance (R2 = 0.61, F(3, 107) = 54.91, p < .001, adj. R2 = 0.60). The model's intercept, corresponding to rad =

0, temp = 0 and wind = 0, is at -64.23 (95% CI [-109.91, -18.55], t(107) = -2.79, p = 0.006). Within this model:

- The effect of rad is statistically significant and positive (beta = 0.06, 95% CI [0.01, 0.11], t(107) = 2.58, p = 0.011; Std. beta = 0.16, 95% CI [0.04, 0.29])
- The effect of temp is statistically significant and positive (beta = 1.65, 95% CI [1.15, 2.15], t(107) = 6.52, p < .001; Std. beta = 0.47, 95% CI [0.33, 0.62])
- The effect of wind is statistically significant and negative (beta = -3.34, 95% CI [-4.63, -2.04], t(107) = -5.10, p < .001; Std. beta = -0.36, 95% CI [-0.50, -0.22])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

## 2.46 5.b vif first linear model

```
 rad     temp     wind
```

1.095241 1.431201 1.328979

Pass multicollinearity test so continue to cross validation

## 2.47 5.b GVLMA first linear model

Call: lm(formula = ozone ~ rad + temp + wind, data = ozonedat)

Coefficients: (Intercept) rad temp wind
-64.2321 0.0598 1.6512 -3.3376

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM: Level of Significance = 0.05

Call: gvlma(x = lm3)

```
                    Value   p-value                   Decision
```

Global Stat 111.30249 0.000e+00 Assumptions NOT satisfied! Skewness 34.23929 4.873e-09 Assumptions NOT satisfied! Kurtosis 50.28630 1.329e-12 Assumptions NOT satisfied! Link Function 26.72320 2.348e-07 Assumptions NOT satisfied! Heteroscedasticity 0.05369 8.168e-01 Assumptions acceptable. Initial GVLMA failed continue to cross validation

Table 4: Assumption Test Results

| Global.Stat | Skewness | Kurtosis | Link.Function | Heteroscedasticity | p.value | Decision |
|---|---|---|---|---|---|---|
| 111.3025 | 34.23929 | 50.2863 | 26.7232 | 0.053689 | 0.0000000 | Assumptions NOT satisfied! |
| 111.3025 | 34.23929 | 50.2863 | 26.7232 | 0.053689 | 0.0000000 | Assumptions NOT satisfied! |
| 111.3025 | 34.23929 | 50.2863 | 26.7232 | 0.053689 | 0.0000000 | Assumptions NOT satisfied! |
| 111.3025 | 34.23929 | 50.2863 | 26.7232 | 0.053689 | 0.0000002 | Assumptions NOT satisfied! |

| Global.Stat | Skewness | Kurtosis | Link.Function | Heteroscedasticity | p.value | Decision |
|---|---|---|---|---|---|---|
| 111.3025 | 34.23929 | 50.2863 | 26.7232 | 0.053689 | 0.8167641 | Assumptions acceptable. |

## 2.48 5.c Cross Validation and K fold

## 2.49 5.c Cross Validation Results

| Model.Formula | RMSE | R2 | MAE |
|---|---|---|---|
| log(ozone)~ 1+temp+wind+temp:wind+temp:rad | 0.5127035 | 0.6648894 | 0.4017731 |
| log(ozone)~ 1+temp+temp:wind+temp:rad+rad:wind | 0.5184218 | 0.6577820 | 0.4018344 |
| log(ozone)~ 1+temp+temp:wind+temp:rad | 0.5194998 | 0.6586985 | 0.4015029 |
| log(ozone)~ 1+rad+temp+temp:wind | 0.5197990 | 0.6489711 | 0.4013519 |
| log(ozone)~ 1+rad+temp+wind+temp:wind | 0.5203636 | 0.6585872 | 0.4081567 |
| log(ozone)~ 1+temp+temp:rad+rad:wind | 0.5209050 | 0.6649092 | 0.4006607 |
| log(ozone)~ 1+temp+wind+temp:rad+rad:wind | 0.5212435 | 0.6626349 | 0.4041197 |
| log(ozone)~ 1+temp+wind+temp:rad | 0.5253292 | 0.6639494 | 0.4045672 |
| log(ozone)~ 1+rad+temp+rad:wind | 0.5259307 | 0.6602642 | 0.4068017 |
| log(ozone)~ 1+rad+temp+wind+temp:wind+rad:wind | 0.5260944 | 0.6514639 | 0.4127110 |

## 2.50 New Linear Model from Cross Validated Results [ with outliers]

Test

Global.Stat

p.value

Decision

Global Stat

27.919015

1.295329e-05

Assumptions NOT satisfied!

Skewness

3.464400

6.270323e-02

Assumptions acceptable.

Kurtosis

14.078607

1.753266e-04

Assumptions NOT satisfied!

Link Function

1.690542

1.935296e-01

Assumptions acceptable.

Heteroscedasticity

8.685466

3.207576e-03

Assumptions NOT satisfied!

## 2.51 5.Cooks Distance on Cross Validated Linear Model

Lag

Autocorrelation

11

0.0362175

17

0.2846866

18

0.0353861

19

0.0362745

20

0.0489980

30

0.0835909

45

0.0448058

77

0.0407174

85

0.0345147

## 2.52 Removing Cook's Distance points from cross validated data

| rad | temp | wind | ozone |
|---|---|---|---|
| 190 | 67 | 7.4 | 41 |
| 118 | 72 | 8.0 | 36 |
| 149 | 74 | 12.6 | 12 |
| 313 | 62 | 11.5 | 18 |
| 299 | 65 | 8.6 | 23 |
| 99 | 59 | 13.8 | 19 |

## 2.53 New Linear Model of Cross Validated with NO Outliers

Call: lm(formula = log(ozone) ~ 1 + temp + wind + temp:wind + temp:rad, data = ozonedat_withoutoutliers)

Coefficients: (Intercept) temp wind temp:wind temp:rad
-9.362e-01 5.741e-02 5.489e-02 -1.502e-03 3.217e-05

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM: Level of Significance = 0.05

Call: gvlma(x = lm11)

```
                    Value p-value                   Decision
```

Global Stat 4.91460 0.2962 Assumptions acceptable. Skewness 0.03755 0.8463 Assumptions acceptable. Kurtosis 0.80786 0.3688 Assumptions acceptable. Link Function 0.96319 0.3264 Assumptions acceptable. Heteroscedasticity 3.10600 0.0780 Assumptions acceptable.

## 2.54 5.c and d GVLMA final Linear Model

Test

Value

p_value

Decision

Global Stat

4.9146022

0.2961726

Assumptions acceptable.

Skewness

0.0375543

0.8463408

Assumptions acceptable.

Kurtosis

0.8078563

0.3687548

Assumptions acceptable.

Link Function

0.9631888

0.3263848

Assumptions acceptable.

Heteroscedasticity

3.1060028

0.0780042

Assumptions acceptable.

|  | x |
|---|---:|
| (Intercept) | -60.7884955 |
| temp | 5.9089948 |
| wind | 5.6424402 |
| temp:wind | -0.1500873 |
| temp:rad | 0.0032171 |

## 2.55 5.c and d Summary and Values for semi-final

Call: lm(formula = log(ozone) ~ 1 + temp + wind + temp:wind + temp:rad, data = ozonedat_withoutoutliers)

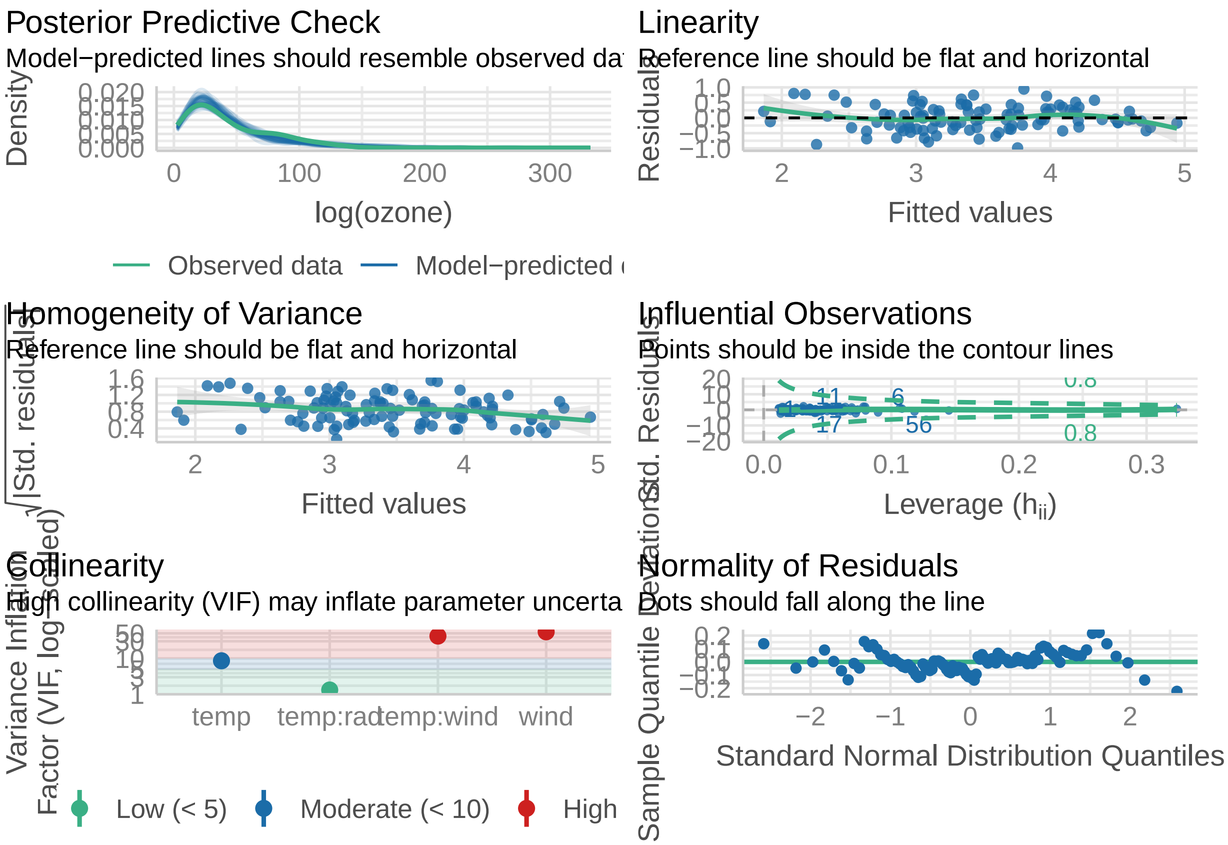
Residuals: Min 1Q Median 3Q Max -0.98332 -0.31350 -0.04297 0.27936 0.94212

Coefficients: Estimate Std. Error t value Pr(>|t|)<br>
(Intercept) -9.362e-01 1.050e+00 -0.891 0.375<br>
temp 5.741e-02 1.278e-02 4.494 1.92e-05 ***wind 5.489e-02 9.243e-02 0.594 0.554***<br>
***temp:wind -1.502e-03 1.159e-03 -1.295 0.198***<br>
***temp:rad 3.217e-05 6.589e-06 4.882 4.08e-06*** — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

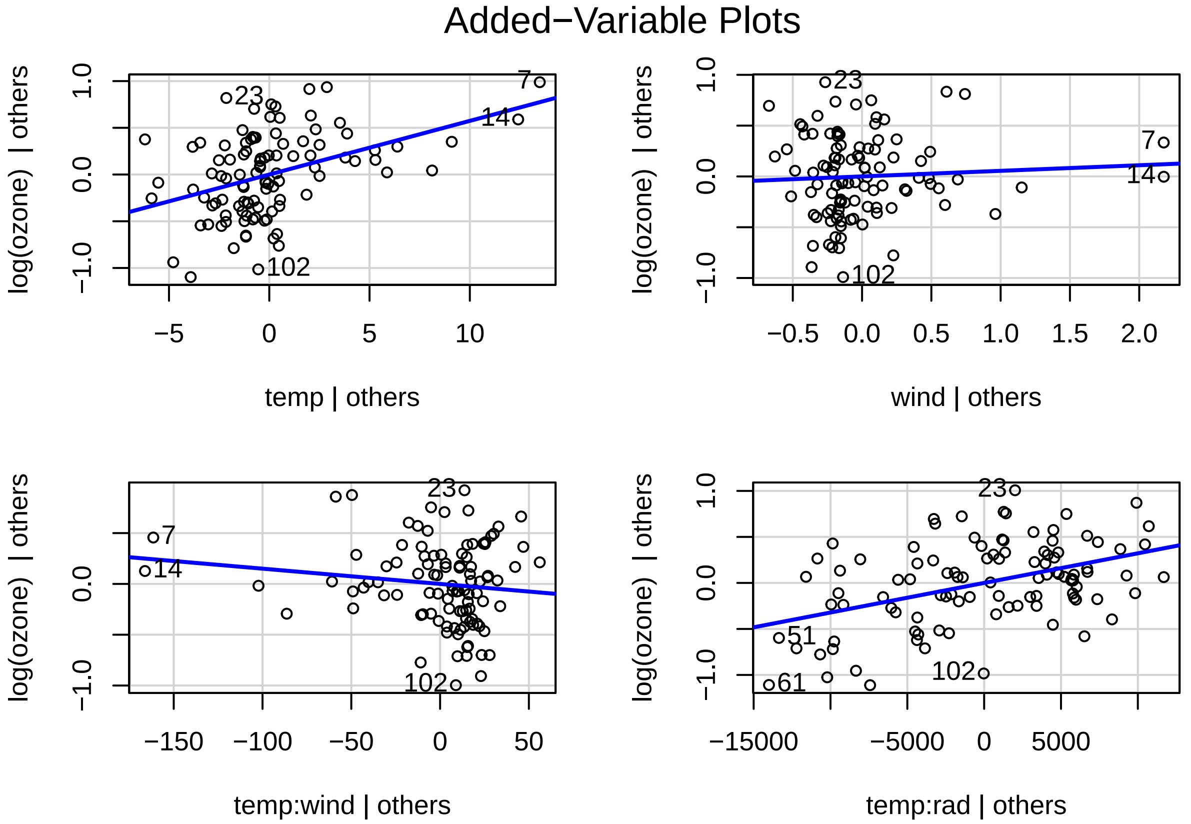
Residual standard error: 0.4146 on 98 degrees of freedom Multiple R-squared: 0.7345, Adjusted R-squared: 0.7237 F-statistic: 67.79 on 4 and 98 DF, p-value: < 2.2e-16 ## 5.c exponentiate variables of summary

Every time temperature increases by one unit ozone increases by 5.9%. Everytime wind increases by one unit ozone increases 5.64 %. Every additional rise in unit temp decreases the positive/neg effect of wind on the dependent variable of ozone by -.15%. every additional in temp decreases the positive/neg effect of solar radiation on dependent variable ofozone decreases by -0.00321%.

### 2.56 Question 5.e



### 2.57 5.e simple main effects plots

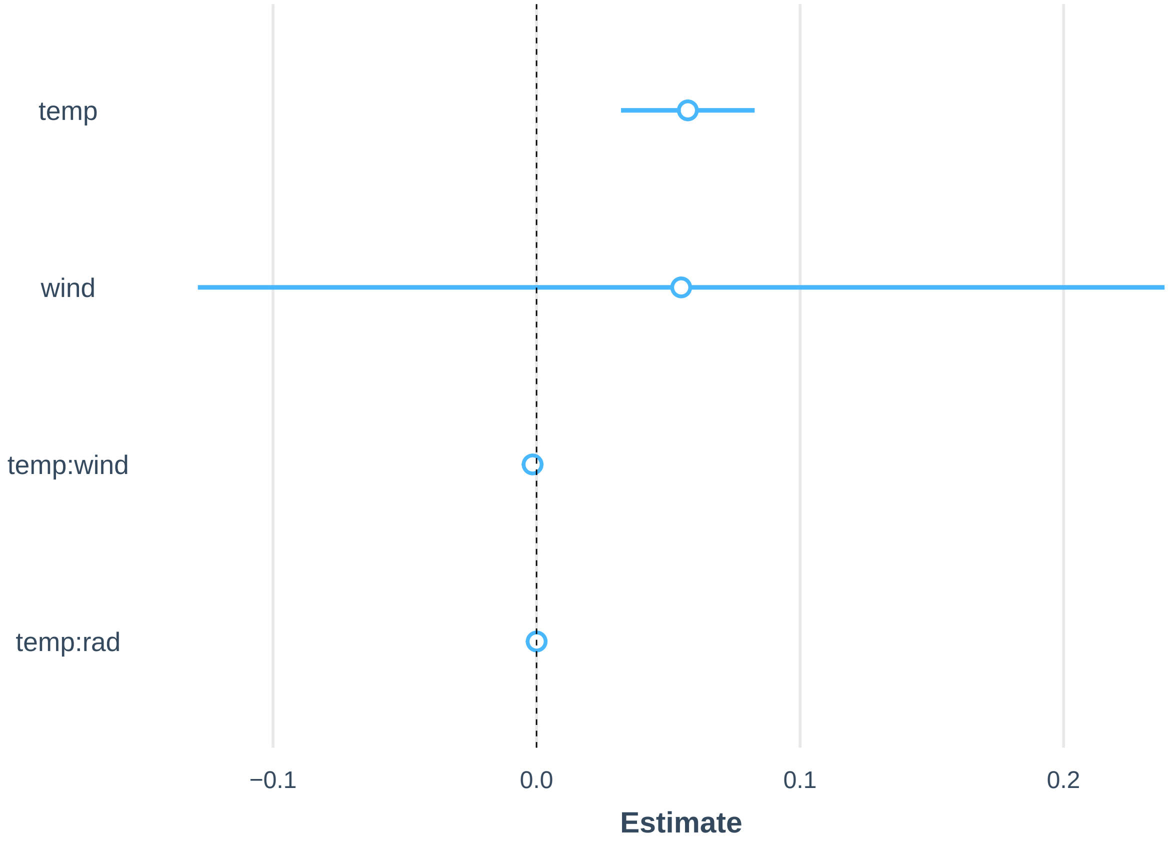


Every time temputure increases by one unit ozone increases by 5.9%, evertime wind increases by one unit ozone decreases by -.15 %,every additional in temp decreases the positive/neg effect of wind on dependent variable of ozone decreases by -.15. every additional in temp decreases the positive/neg effect of solar radiation on dependent variable of ozone decreases by -0.00321%.

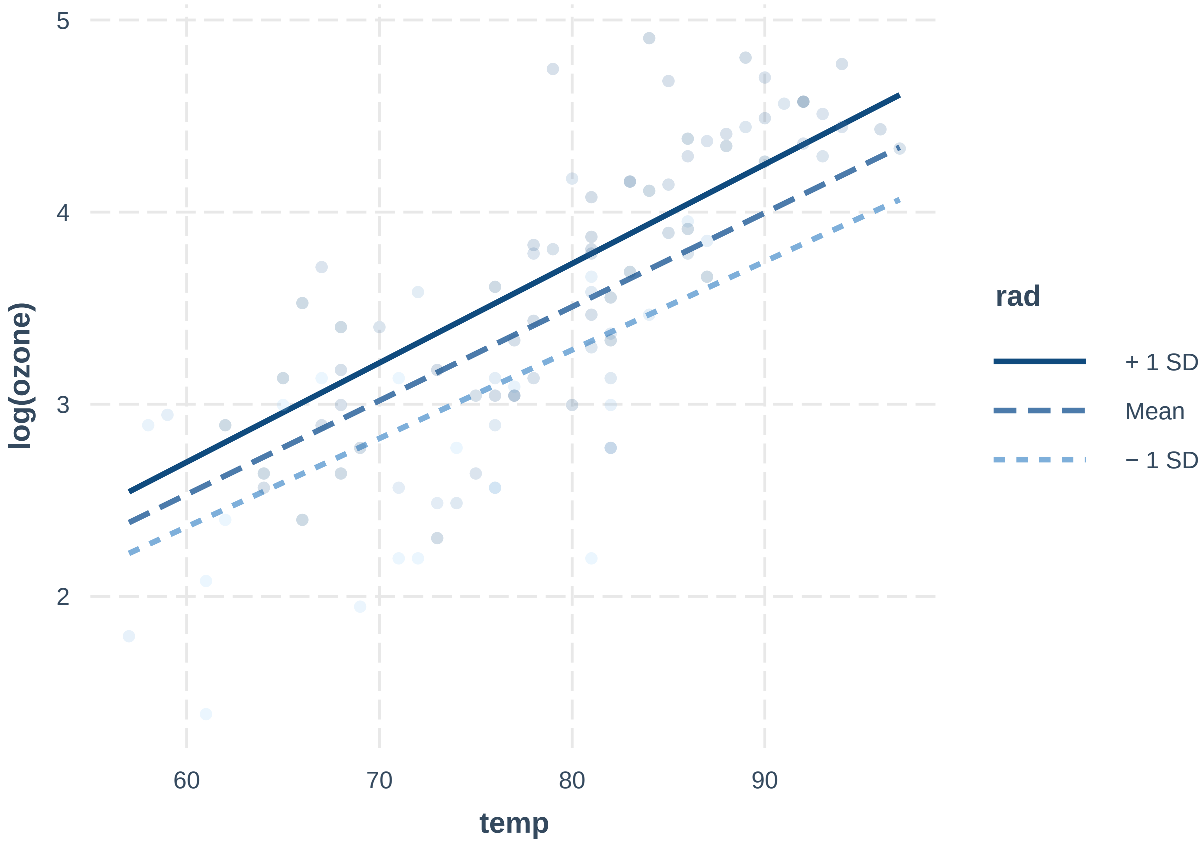
What I did From the cross validation and subsequent data wrangling we have determined several things: - Running the cross validation with the interactive effects on the logged ozone we where able to obtain more accurate results from the CV, with that and after removing outliers from that data frame, all assumptions passed in the GVLMA.

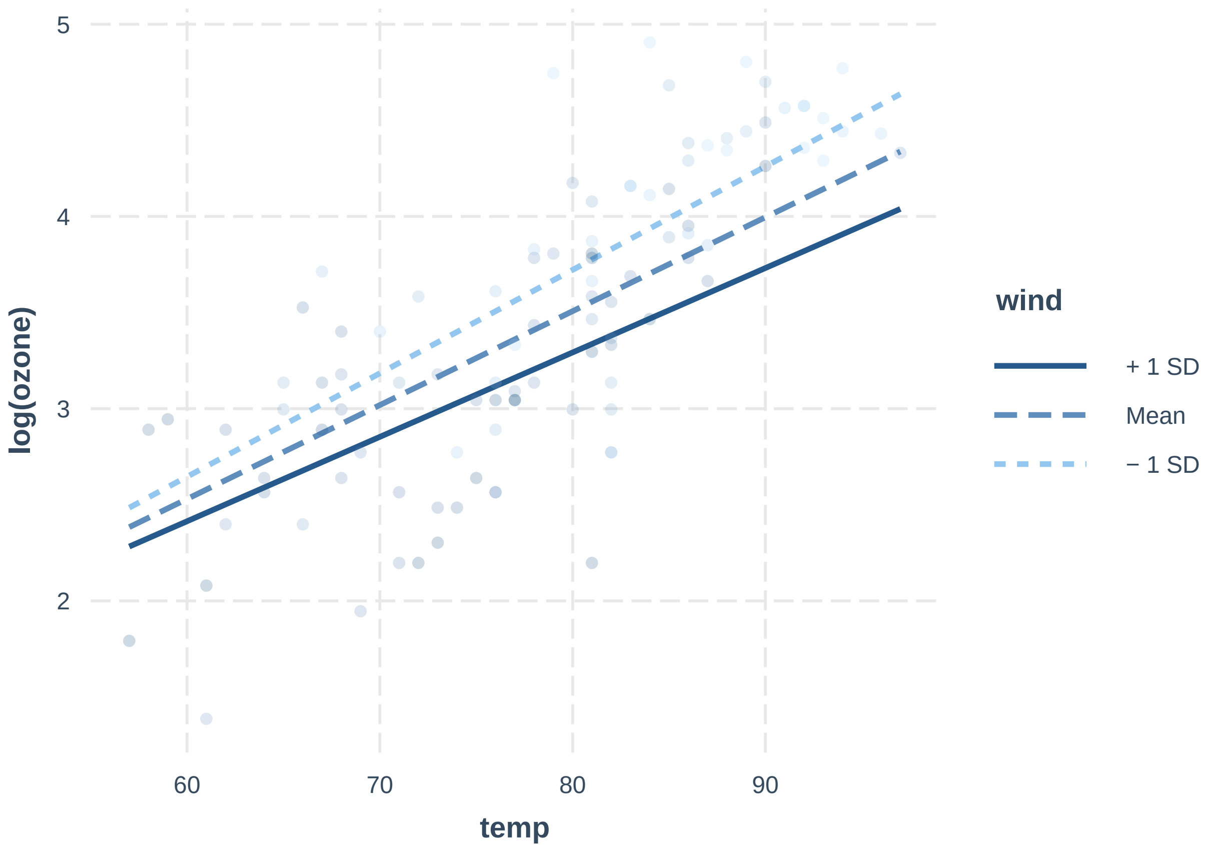
From these results we determine that the most accurate predictor variables are temp and wind with the most prominenet interaction effects being temp:wind and temp:rad on the response variable of ozone.

### 2.58 5.e plot coefecients simple main effects and interaction effect plots



### 2.59 5.e Interaction plots for ineraction effects




5
4
3
2
log(ozone)
60
70
80
90
temp
wind
+ 1 SD
Mean
− 1 SD