Homework 1

Allan Kimaina

February 5, 2018

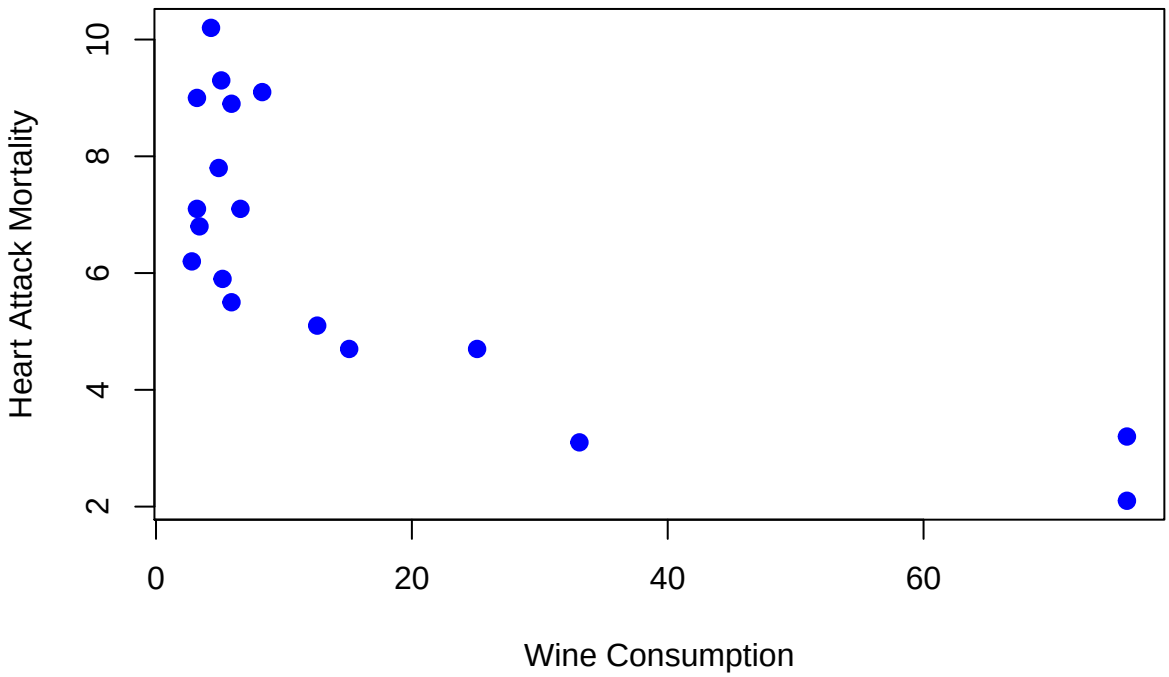
Question 1

The data in the file wine.csv (in the data sets folder on Canvas) give the average wine consumption rates (in liters per person) and number of ischemic heart attack deaths (per 1000 men aged 55 to 64 years) for 18 industrialized countries. Do these data suggest that heart disease death rates are associated with average wine consumption? If so, how can that be described?

Do any countries have substantially higher or lower death rates than others with similar wine consumption rates?

Analyze the data and write a brief report that includes a summary of findings, a graphical display and a section describing the methods used to answer the questions of interest.

The goal of the study is to determine the significance of the relationship between wine consumption and ischemic heart attack mortality

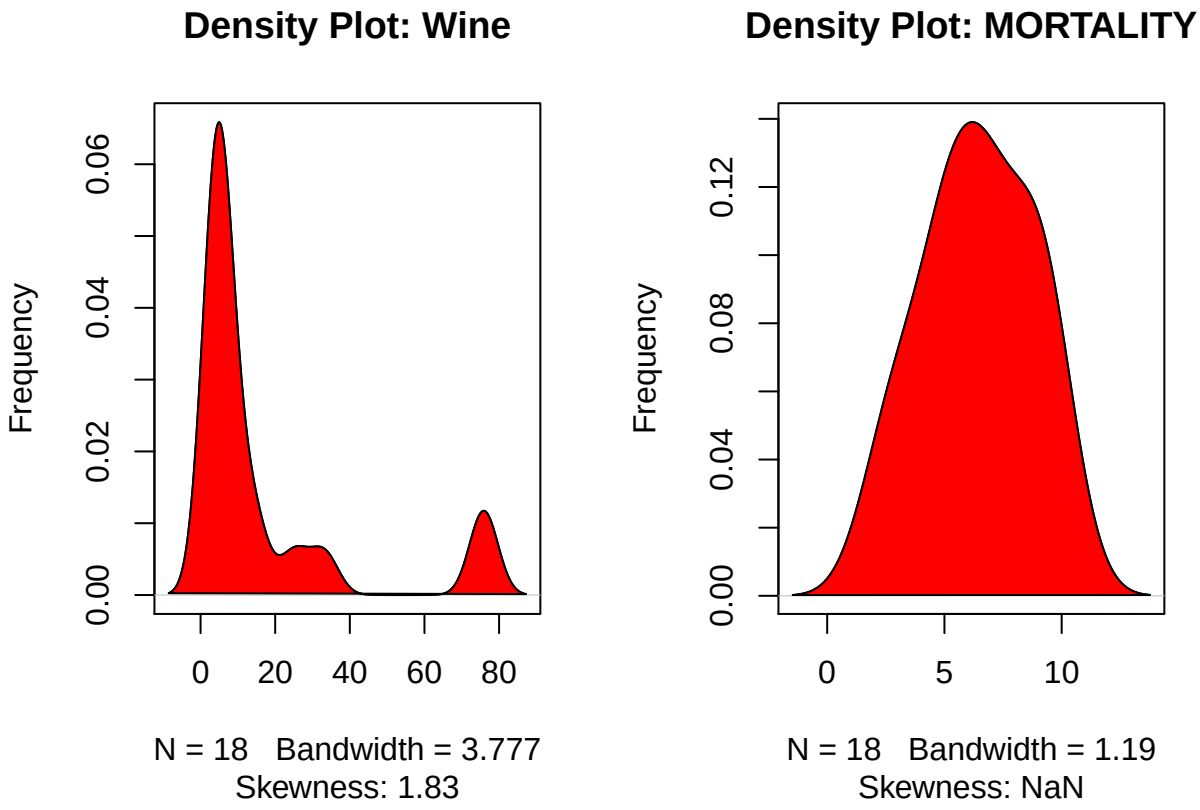


From the scatter plot, the data points suggest that there is a strong negative association between wine consumption and Ischemic heart attack deaths. Countries with low wine consumption have high heart attack mortality rates compared to countries with high wine consumption. Computing the correlation coefficient we get -0.7455682 indicating that heart attack mortality is inversely proportional to wine consumption. However,

this could be misleading since the plot indicates that there is a possibility of nonlinear relationship (possibly exponential) between the predictor variable and the response variable, i.e heart attack mortality rate increases exponentially as wine consumption per liter decreases

But first before establishing the best model that fits this association, lets try to understand the characteristics of predictor and response variables graphically.

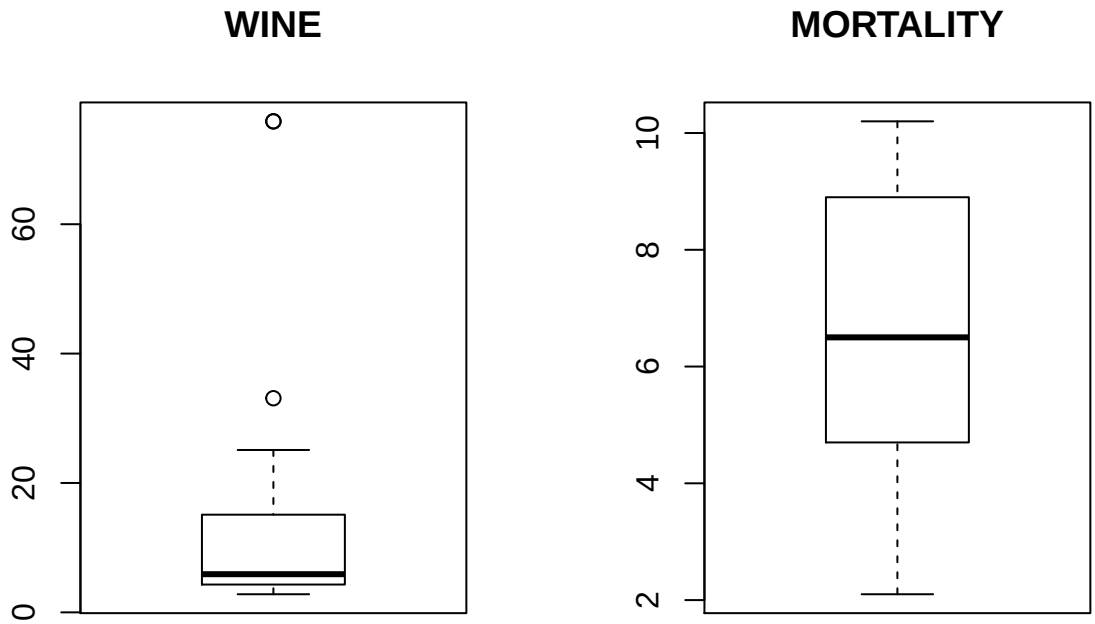
Data Examination & Exploration



Looking at the density plot, the response variable is normal without any skewness, however, the predictor variable is skewed by 1.83. This plot suggest that we could be having an outlier in the predictor variable (Wine). This can drastically distort or bias our model. Therefore, this prompts for further outlier detection and possibly outlier treatment.

Let's see if there is any data-point which lies outside the 1.5 * distance between the 25th percentile and 75th percentile values (IQR).

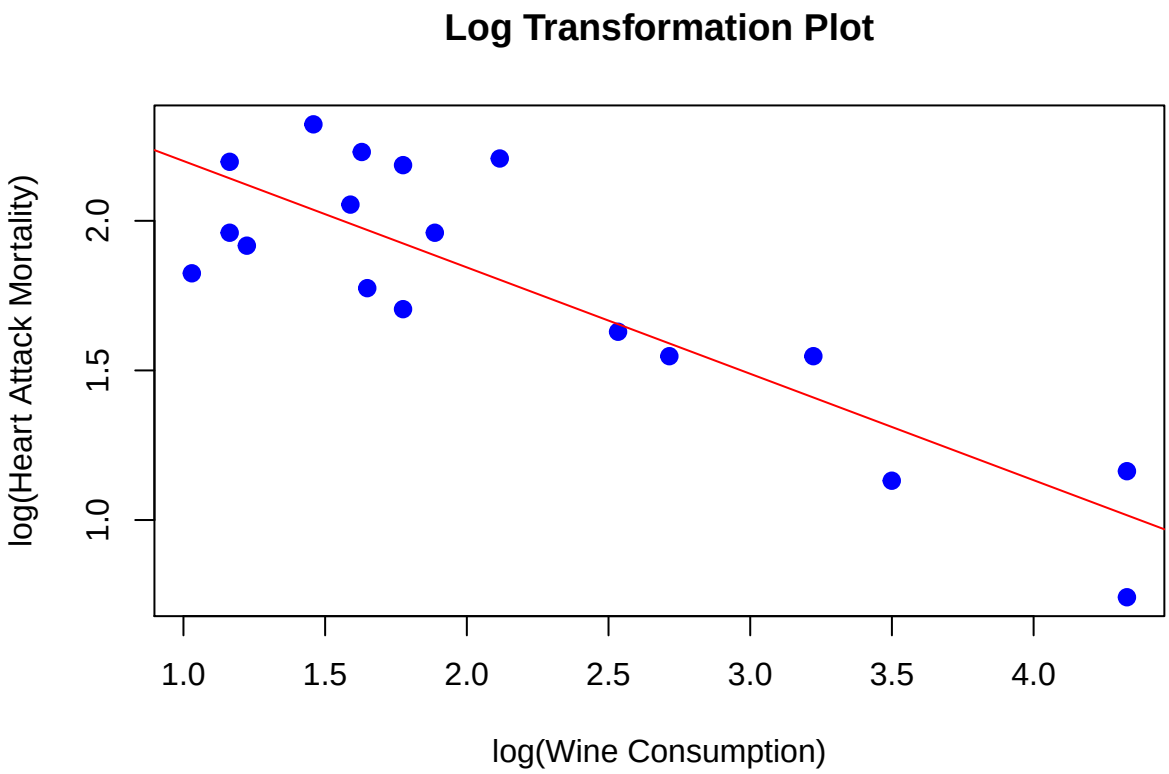
Outlier Detection



The box plot suggest that there are two outliers, France and Italy. Even though France and Italy are separated from the rest of the data-points, we do not have sufficient information to conclude that these 2 data points are outliers. For all we know there could be more data-points in the population (other countries) that would have changed this notion if they were included. In fact, an exponential relationship can be clearly visible on the original scatter plot.

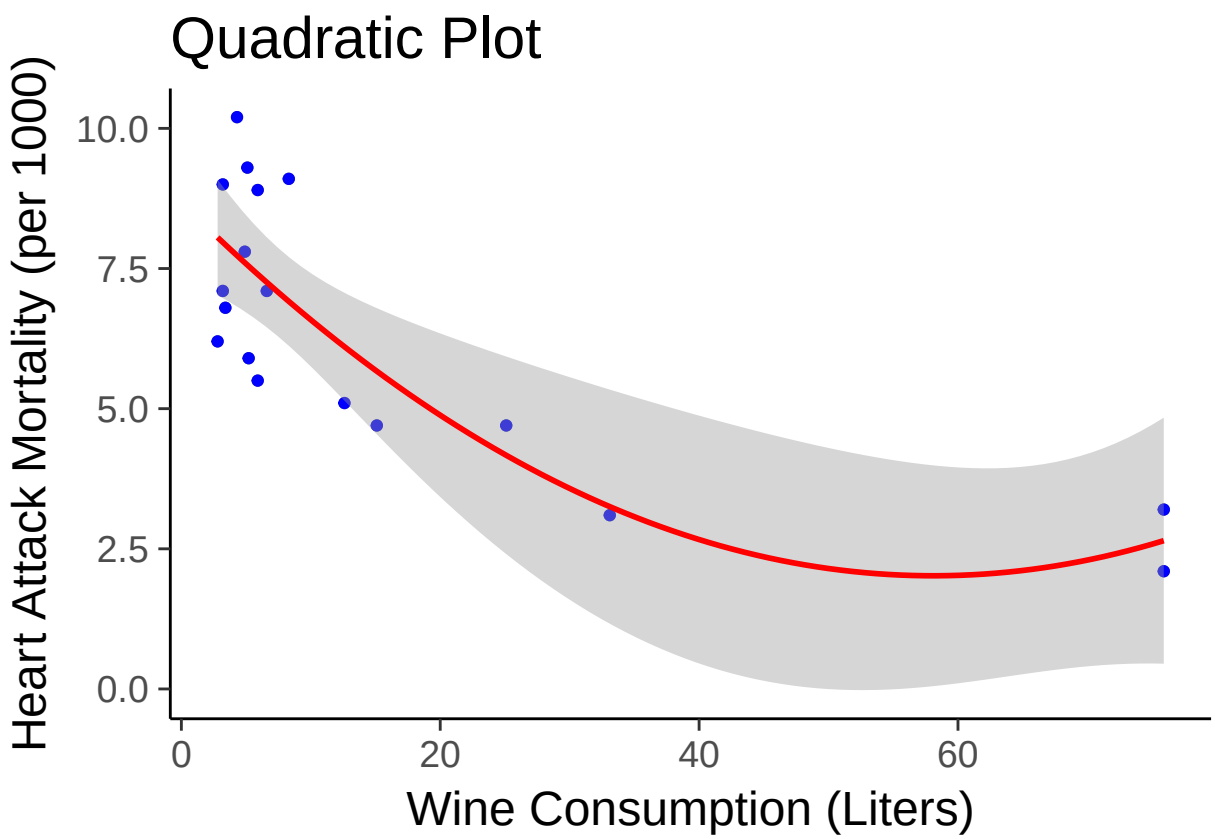
Exponential relationship

From the above scatter plot and density plot, the explanatory variable (WINE CONSUMPTION) is strongly positively skewed with many data points near zero and fewer values at the extreme right. This indicates a multiplicative effect on the response (Heart Attack Mortality). Therefore, it is possible that the log transformation would make the association between heart attack mortality and wine consumption closer to a linear relationship. Let's asses this possibility by plotting a scatter of log transformed data:



Transforming both predictor and response variable we get a perfect linear relationship implying a linear model can be fitted easily. Before fitting this model let us assess other possible models.

Another possible explanation for the curvilinear relationship could be because the relationship between predictor and response variable is modeled as 2 degree polynomial (quadratic) in the predictor.

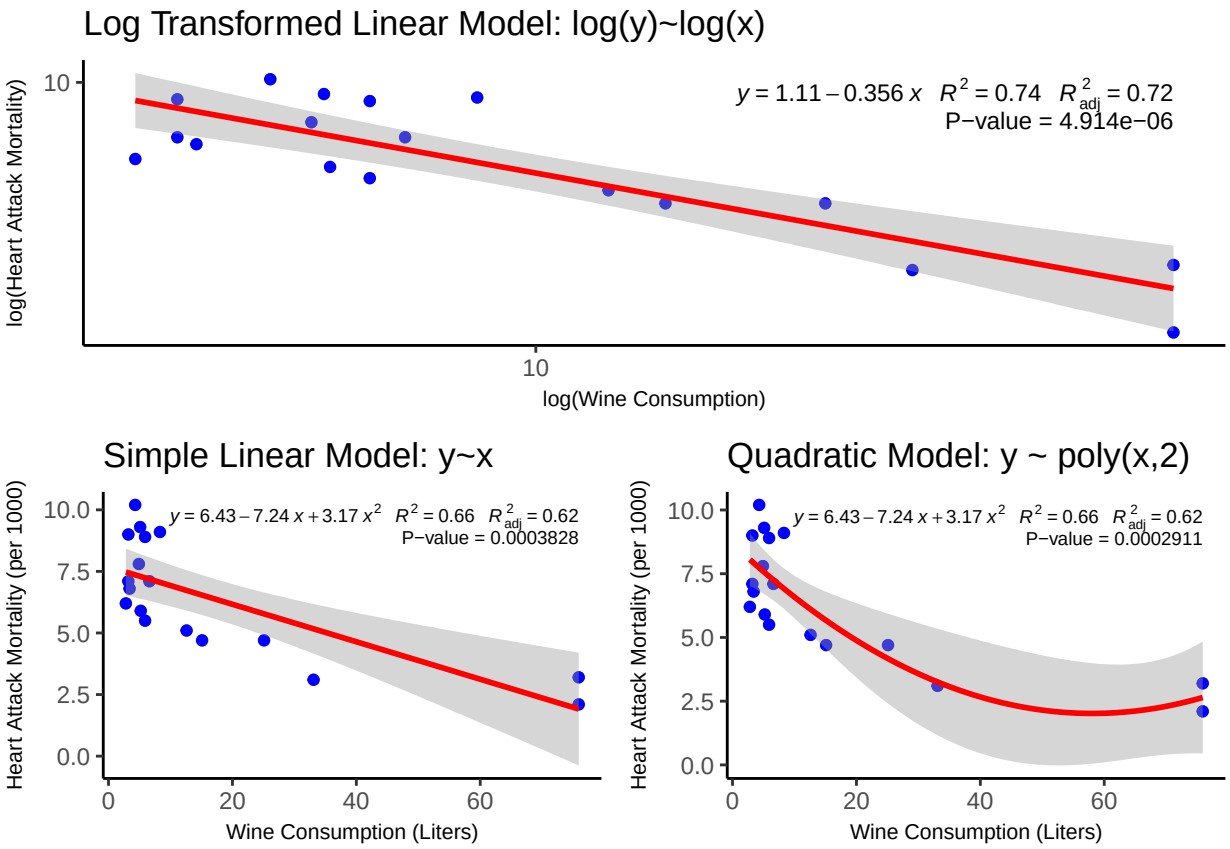


There is little reason to believe that the relationship between wine consumption and heart attack mortality is quadratic. In fact this is depicted by r-squared value explained in the next section.

Possible Models

From the previous section, we have been able to identify 3 possible models

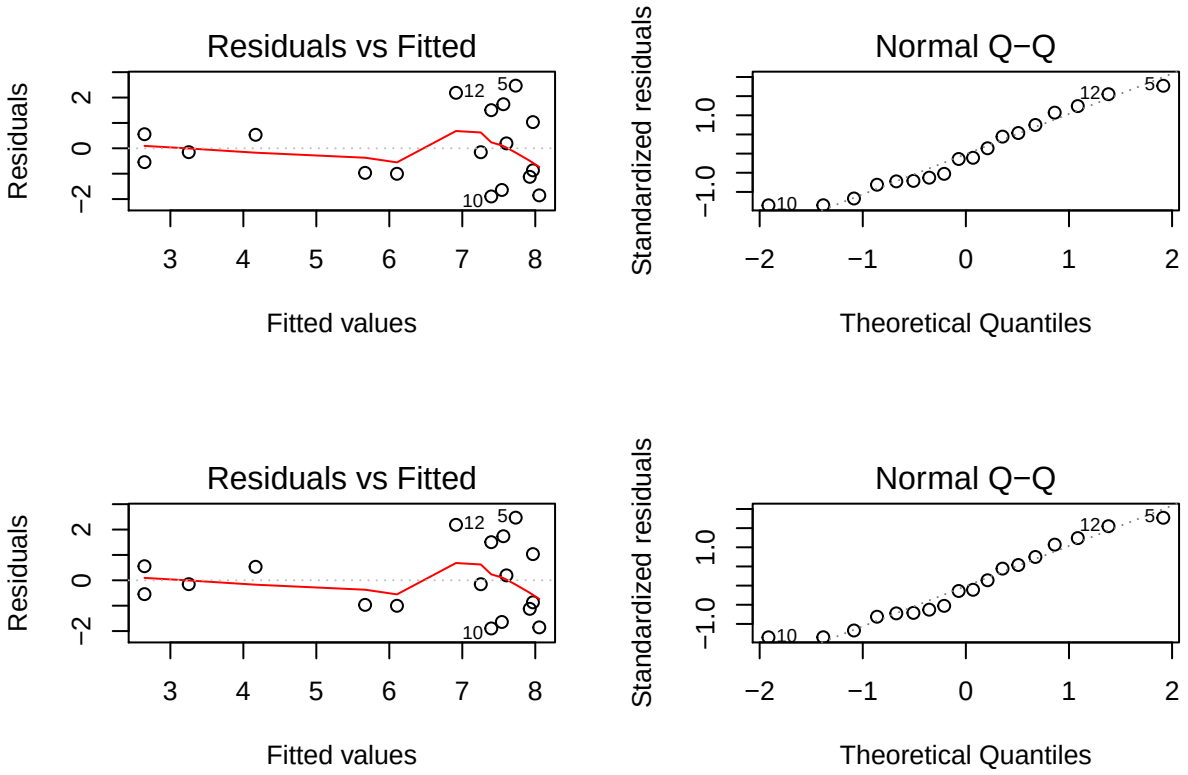
1. Simple Linear model
2. Log Transformed Linear Model (exponential)
3. Quadratic model (highly unlikely)



well, explanatory power and residual plot can help us determine which model fits well the relationship between wine consumption and heart attack

Model Selection:

Both simple Linear model (y~x) and quadratic model (y~poly(x,2)) have really low explanatory power (R-squared and adj-R squared). In simple Linear model, only 66% of the variability in heat attack mortality is explained by wine consumption while in quadratic model on 64% of the variability in the response is explained by the predictor.



Furthermore, the residual plot in both model seems to follow a consistent trend at the same time exhibiting residual heteroskedasticity. Randomness and unpredictability is a crucial components of any regression model therefore for this reason among others, we are not going to do further diagnostics on them. The rule of thumb has always been to transform any nonlinear relationship. In this case linearity and equality of standard deviations assumptions have not been met. This gives us more reason to select log transformed model as the best fit model

Log Transformed Linear Model (exponential)

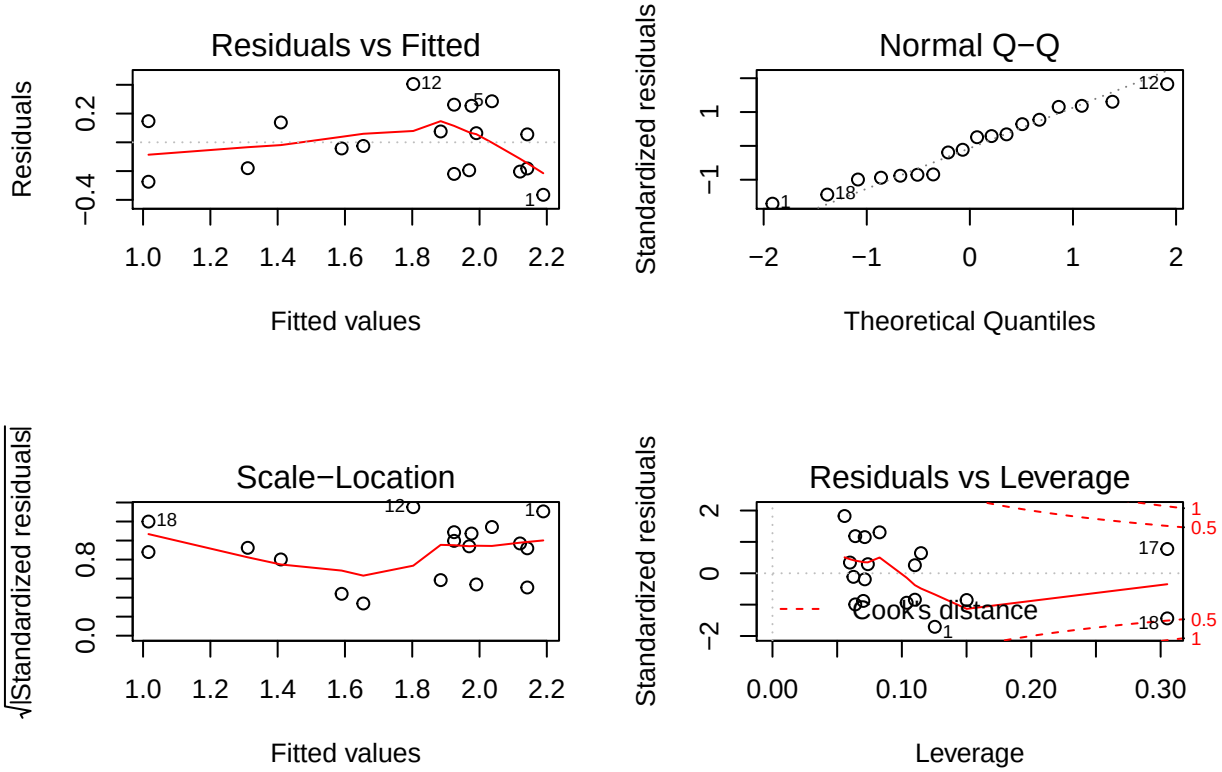
Table 1:

	Dependent variable:
	MORTALITY)
WINE)	-0.356^{***} (0.053)
Constant	2.556^{***} (0.127)
Observations	18
R^2	0.738
Adjusted R^2	0.722
Residual Std. Error	0.229 (df = 16)
F Statistic	45.170^{***} (df = 1; 16)
Note:	*p<0.1; **p<0.05; ***p<0.01

Graphically we can notice that the log transformed regression line covers most data-points as well as having a more precise and uniform CI band compared with the other two models. In fact this model p-value is very

significant (approximately 4.914e-06) and at the same time having a good explanatory power (over 73.6% of the variability in heart attack mortality is explained by wine consumption)

Model Diagnostic:



Residuals vs. fitted values

- We observe an acceptable Scatter-plot in the sense that the fitted values doesn't show any noticeable systematic pattern hence doesn't contain any "leaking" explanatory information. Unlike this model, the other linear model had noticeable pattern.
- From the plot, the residuals bounce randomly around the 0 line and roughly forms a horizontal band around the 0 line. This suggests that the variances of the error terms are approximately equal.
- The plot suggest we have noticeable outliers. This is not concerning since not all outliers are influential. The last plot (Residual vs Leverage) confirms this assumption.

Normal Q-Q Plot

- This plot shows that the residuals are normally distributed since the data points are well arranged on the straight dashed line. No noticeable deviation from the straight line can be observed.

Scale-Location Plot

- This plot confirms the assumption of equal variance (homoscedasticity). The residuals appear randomly scattered with no concerning patterns. We have a smooth horizontal red line, although it's not a perfect straight line.

Residual vs Levarage Plot

- We have no datapoints outside the Cook's distance lines (red dashed lines). This suggest we have no influential outliers on the regression (leverage points).

Summary

Do these data suggest that heart disease death rates are associated with average wine consumption? If so, how can that be described?

The data suggest that wee have a highly significant negative relationship between heart attack mortality and wine consumption. The extremely low p-value (<0.05) indicates that it is highly significant. Since the relationship depicted exponential pattern we had to do transformation so that the linearity and equality of variance assumptions can be achieved. After transformation, the data suggested that an increase in Wine consumption by 1 unit (log scale) resulted to a decrease in the number of Ischemic heart attack mortality by 0.356 (log scale).

$$log(y) = \beta_0 + \beta_1 log(x) + \epsilon$$

$$log(HeartAttackMortality) = \beta_0 + \beta_1 log(WineConsumption) + \epsilon$$

$$log(HeartAttackMortality) = 2.556 - 0.356 * log(WineConsumption) + \epsilon$$

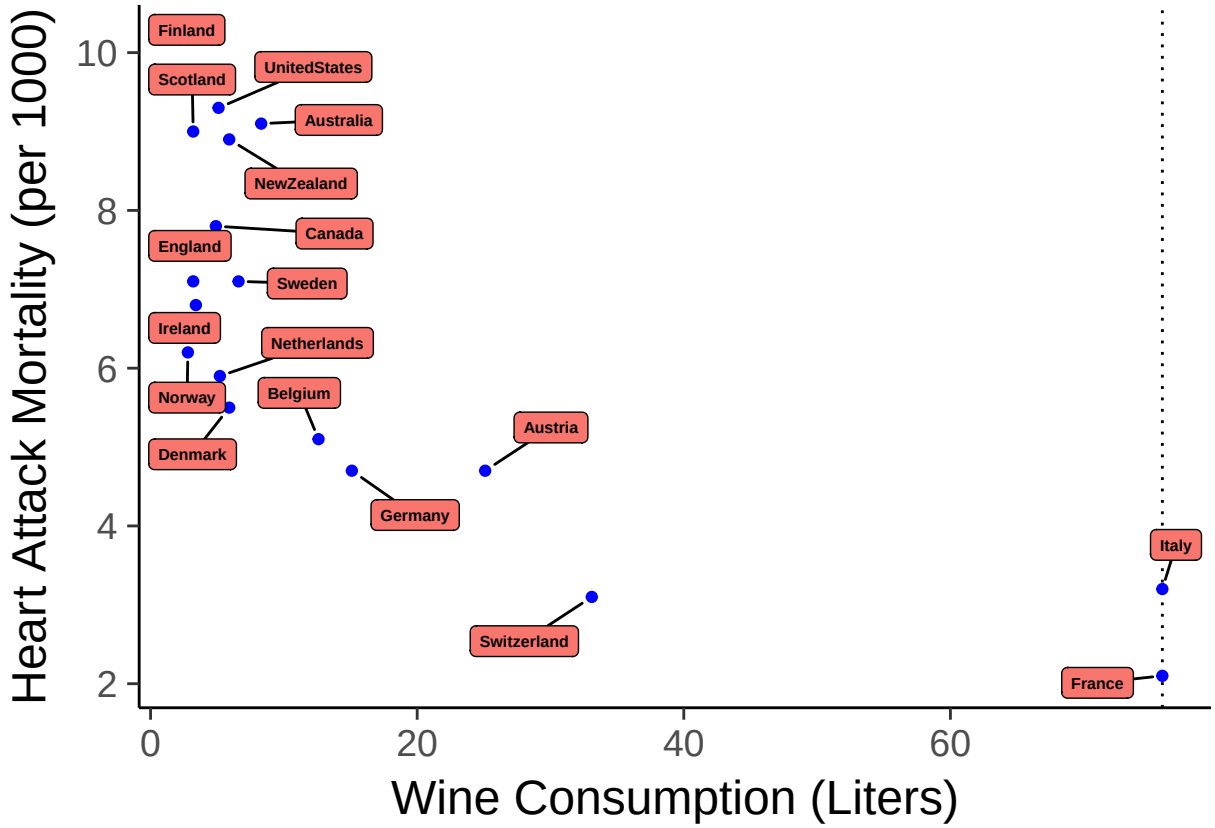
What does this tell us?

This implies that there is a multiplicative change between the response and predictor.

- **Multiplicative changes in e** - Multiplying WineConsumption by "e" will decrease HeartAttackMortality by 0.356%
- **A 1% increase in WineConsumption** - A 1% increase in Wine Consumption will decrease Heart Attack Mortality by 0.356/100 = 0.00356%
- **A 10% increase in WineConsumption** - A 10% increase in Wine Consumption will decrease Heart Attack Mortality by 0.356*log(1.10) = 0.03393042%

This tells us that an increase of wine consumption by 10% will decrease Ischemic heart attack mortality by 0.03393042%. However, it is important to note that this relationship could have been confounded by other factors not included in this dataset.

Do any countries have substantially higher or lower death rates than others with similar wine consumption rates?



Yes, both the data table and the plot above indicate that some countries have substantially higher or lower death rates than others with similar wine consumption rates.

- Australia and German have same heart mortality (4.7) yet wine consumption in Austria (25.1) is way higher than Germany (15.1)
- France and Italy have same wine consumption of (75.9) yet heart mortality in France (2.1) is lower than Itality (3.2)
- Scotland and England have same wine consumption of (3.2) yet heart mortality in Scotland (9.0) is way higher than England (7.1)
- United States (5.1) and Netherlands (5.2) have similar wine consumption yet heart mortality in United States (9.3) is way higher than Netherlands (5.3)

What does this tell us?

There are some variables that were not collected or included in the dataset that would aid us in extrapolating these differences

Question 2

Meadowfoam is a small plant that grows in Pacific Northwest and is domesticated for its seed oil. A study was set up to determine if meadowfoam can be made into a profitable crop. In a controlled growth chamber, the plant was grown at 6 different light intensities and two different timings of onset of light treatment. The outcome of interest is the number of flowers per plant which was measured by averaging numbers of flowers produced by 10 seedlings in each group. Growth was replicated at each combination of time and light intensity.

a. First put the data into a dataset with four variables: number of flowers, light intensity, timing and replicate.

Table 2:

	FLOWERS	TIME	INTENS	REPLICATE
1	62.300	1	150	1
2	77.400	1	150	2
3	55.300	1	300	1
4	54.200	1	300	2
5	49.600	1	450	1
6	61.900	1	450	2
7	39.400	1	600	1
8	45.700	1	600	2
9	31.300	1	750	1
10	44.900	1	750	2
11	36.800	1	900	1
12	41.900	1	900	2
13	77.800	2	150	1
14	75.600	2	150	2
15	69.100	2	300	1
16	78	2	300	2
17	57	2	450	1
18	71.100	2	450	2
19	62.900	2	600	1
20	52.200	2	600	2
21	60.300	2	750	1
22	45.600	2	750	2
23	52.600	2	900	1
24	44.400	2	900	2

T.test Value

```
Welch Two Sample t-test
```

data: FLOWERS by REPLICATE t = -0.56369, df = 21.949, p-value = 0.5787 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -15.013752 8.597085 sample estimates: mean in group 1 mean in group 2 54.53333 57.74167

We fail to reject the null hypothesis since the p-value of .5787 indicates that the two replicate groups have no significant difference. In order to achieve parsimony, we will not include this predictor in our model.

b. Create a categorical form of the light intensity with 6 categories.

c. The research questions are: What are effects of intensity and timing? Is there an interaction between the two factors?

Table 3:

	INTENS	INTENS_CATEGORY_
1	150	Dimmest
2	300	Dimmer
3	450	Dim
4	600	Bright
5	750	Brighter
6	900	Brightest

d. First create an analysis of variance using timing and the categorical form of the light intensity variable. Determine if there is an effect of each factor.

Table 4:

	Dependent variable:
	FLOWERS
INTENS_CATEGORY_Brighter	−4.525 (4.751)
INTENS_CATEGORY_Brightest	−6.125 (4.751)
INTENS_CATEGORY_Dim	9.850* (4.751)
INTENS_CATEGORY_Dimmer	14.100*** (4.751)
INTENS_CATEGORY_Dimmest	23.225*** (4.751)
TIME_CATEGORY_LATE	−12.158*** (2.743)
Constant	56.129*** (3.629)
Observations	24
R^2	0.823
Adjusted R^2	0.761
Residual Std. Error	6.719 (df = 17)
F Statistic	13.181*** (df = 6; 17)
Note:	*p<0.1; **p<0.05; ***p<0.01

- *INTENS_CATEGORY_Brighter* - If we held other predictors constant, under a "brighter" light intensity condition results to a decrease of 4.5 flowers. However this is insignificant.
- *INTENS_CATEGORY_Brightest* - If we held other predictors constant, under the "brightest" light intensity condition results to a decrease of 6 flowers. However this is insignificant.
- *INTENS_CATEGORY_Dim* - If we held other predictors constant, under a "Dim" light intensity condition results to an increase of 9 flowers. However this is insignificant.
- *INTENS_CATEGORY_Dimmer* - If we held other predictors constant, under a "Dimmer" light intensity condition results to an increase of 14 flowers.
- *INTENS_CATEGORY_Dimmest* - If we held other predictors constant, under the "Dimmest" light intensity condition results to an increase of 23 flowers.
- *TIME_CATEGORY_LATE* - If we hold other predictors constant, "Late" onset timing of light treatment results to a decrease in 23 flowers.

Looking at the p-values, light intensity condition ("Dimmer", "Dimmest") and Timing ("Late") are statistically significant. The Dimmer it gets, results to an increase in the number of flowers. "Late" onset timing of light treatment results to a decrease in number of flowers.

Significance of REPLICATE in the model

Table 5:

	Dependent variable:
	FLOWERS
INTENS_CATEGORY_Brighter	−4.525 (4.696)
INTENS_CATEGORY_Brightest	−6.125 (4.696)
INTENS_CATEGORY_Dim	9.850* (4.696)
INTENS_CATEGORY_Dimmer	14.100*** (4.696)
INTENS_CATEGORY_Dimmest	23.225*** (4.696)
TIME_CATEGORY_LATE	−12.158*** (2.711)
REPLICATE2	3.208 (2.711)
Constant	54.525*** (3.834)
Observations	24
R^2	0.837
Adjusted R^2	0.766
Residual Std. Error	6.641 (df = 16)
F Statistic	11.764*** (df = 7; 16)
Note:	*p<0.1; **p<0.05; ***p<0.01

Adding the REPLICATE factor as one of the predictors in the model exhibited insignificance in the model with a p-value of 0.253968 which greater than our alpha level of .05. This confirmed the t-test

e. Then create an interaction between light intensity and timing by multiplying the two variables and test for the presence of an interaction.

Table 6:

	Dependent variable:
	FLOWERS
INTENS_CATEGORY_Brighter	−4.600 (7.393)
INTENS_CATEGORY_Brightest	−9.050 (7.393)
INTENS_CATEGORY_Dim	6.500 (7.393)
INTENS_CATEGORY_Dimmer	16.000* (7.393)
INTENS_CATEGORY_Dimmest	19.150** (7.393)
TIME_CATEGORY_LATE	−15.000* (7.393)
INTENS_CATEGORY_Brighter:TIME_CATEGORY_LATE	0.150 (10.456)
INTENS_CATEGORY_Brightest:TIME_CATEGORY_LATE	5.850 (10.456)
INTENS_CATEGORY_Dim:TIME_CATEGORY_LATE	6.700 (10.456)
INTENS_CATEGORY_Dimmer:TIME_CATEGORY_LATE	−3.800 (10.456)
INTENS_CATEGORY_Dimmest:TIME_CATEGORY_LATE	8.150 (10.456)
Constant	57.550*** (5.228)
Observations	24
R^2	0.849
Adjusted R^2	0.710
Residual Std. Error	7.393 (df = 12)
F Statistic	6.124*** (df = 11; 12)
Note:	*p<0.1; **p<0.05; ***p<0.01

The p-values are greater than the significant level of .05, therefore none of the interaction terms are statistically significant. This means light intensity level (as a categorical variable) and timing of onset of light treatment

do not interact. The effect of timing doesn't differs based on the level of light intensity and vice versa.

f. Now repeat the process but using light intensity as a continuous variable.

Continous Without Interaction Terms

Table 7:

	Dependent variable:
	FLOWERS
INTENS	-0.040^{***} (0.005)
TIME_CATEGORY_LATE	-12.158^{***} (2.630)
Constant	83.464^{***} (3.274)
Observations	24
R^2	0.799
Adjusted R^2	0.780
Residual Std. Error	6.441 (df = 21)
F Statistic	41.780^{***} (df = 2; 21)
Note:	*p<0.1; **p<0.05; ***p<0.01

- *INTENS* - If we held other predictors constant, an increase of light intensity by 1 unit results to a decrease in the number of flowers by 0.04. The extremely low p-value (<0.05) indicates that is highly significant.
- *TIME_CATEGORY_LATE* - If we held other predictors constant, "Late" timing for light exposure results to a decrease in number of flowers by 12. the p-value of (0.000146) indicates that this is statistically significant

This means that light intensity is inversely proportional to number of flowers. At the same time, Late timing of light exposure result to a decrease in number of flowers.

Continous with Interaction Terms

Table 8:

	Dependent variable:
	FLOWERS
INTENS	-0.040^{***} (0.007)
TIME_CATEGORY_LATE	-11.523^{*} (6.142)
INTENS:TIME_CATEGORY_LATE	-0.001 (0.011)
Constant	83.147^{***} (4.343)
Observations	24
R^2	0.799
Adjusted R^2	0.769
Residual Std. Error	6.598 (df = 20)
F Statistic	26.549^{***} (df = 3; 20)
Note:	*p<0.1; **p<0.05; ***p<0.01

Looking at the low p-value of (0.9096), we have insufficient evidence of interaction between timing of exposure and light intensity as a continuous variable with a p-value greater than .05

g. Then perform F-tests to compare the four model you have created (light as continuous and categorical with and without the interaction)

Model 1 as the reference (light as categorical variabe & without interaction)

Table 9:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	17	767.472				
2	12	655.925	5	111.547	0.408	0.834
3	21	871.236	-9	-215.311	0.438	0.889
4	20	870.660	1	0.576	0.011	0.920

- There is no significant difference between the 4 models because all the p-values are extremely high i.e >.05

Model 2 as the reference (light as categorical variabe & without interaction)

Table 10:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	12	655.925				
2	17	767.472	-5	-111.547	0.408	0.834
3	21	871.236	-4	-103.764	0.475	0.754
4	20	870.660	1	0.576	0.011	0.920

- There is no significant difference between the 4 models because all the p-values are extremely high i.e >.05

Model 3 as the reference (light as continous variabe & without interaction)

Table 11:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	21	871.236				
2	17	767.472	4	103.764	0.475	0.754
3	12	655.925	5	111.547	0.408	0.834
4	20	870.660	-8	-214.735	0.491	0.841

- There is no significant difference between the 4 models because all the p-values are extremely high i.e >.05

Model 4 as the reference (light as continous variabe & with interaction)

- There is no significant difference between the 4 models because all the p-values are extremely high i.e >.05

h. Predict the number of flowers grown at each combination of light and timing for each of the four models.

Table 12:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	20	870.660				
2	17	767.472	3	103.188	0.629	0.610
3	12	655.925	5	111.547	0.408	0.834
4	21	871.236	-9	-215.311	0.438	0.889

Table 13:

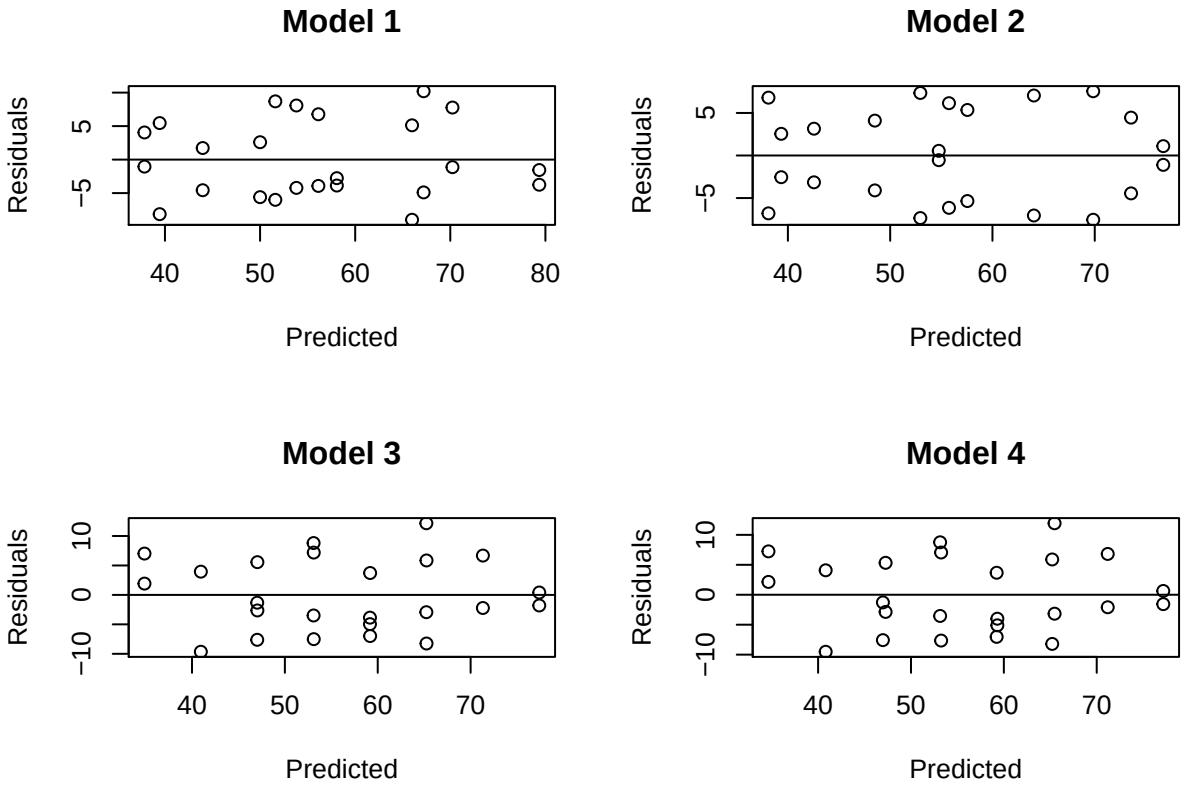
	FLOWERS	prediction1	prediction2	prediction3	prediction4
1	62.300	67.196	69.850	65.235	65.462
2	77.400	67.196	69.850	65.235	65.462
3	55.300	58.071	54.750	59.164	59.300
4	54.200	58.071	54.750	59.164	59.300
5	49.600	53.821	55.750	53.094	53.139
6	61.900	53.821	55.750	53.094	53.139
7	39.400	43.971	42.550	47.023	46.978
8	45.700	43.971	42.550	47.023	46.978
9	31.300	39.446	38.100	40.952	40.816
10	44.900	39.446	38.100	40.952	40.816
11	36.800	37.846	39.350	34.882	34.655
12	41.900	37.846	39.350	34.882	34.655
13	77.800	79.354	76.700	77.393	77.167
14	75.600	79.354	76.700	77.393	77.167
15	69.100	70.229	73.550	71.323	71.187
16	78	70.229	73.550	71.323	71.187
17	57	65.979	64.050	65.252	65.207
18	71.100	65.979	64.050	65.252	65.207
19	62.900	56.129	57.550	59.181	59.227
20	52.200	56.129	57.550	59.181	59.227
21	60.300	51.604	52.950	53.111	53.247
22	45.600	51.604	52.950	53.111	53.247
23	52.600	50.004	48.500	47.040	47.267
24	44.400	50.004	48.500	47.040	47.267

i. Compare each prediction to the observed number of flowers and calculate the difference (observed – predicted). This is the residual. Calculate the residual mean squared error for each model by adding the squared residuals together and dividing by the number of residual degrees of freedom. This should equal the mean squared error in each ANOVA table.

Table 14:

residual_mse1	residual_mse2	residual_mse3	residual_mse4
767.472	655.925	871.236	870.660

j. Now plot the residuals vs. the predicted for each model and see if there are any patterns. If you see any, what might you do to remove them?



Systematic pattern can be visible in model 2 of the residual plot. This means that predictive information is leaking into the residual due to inclusion of insignificant predictors (interraction terms). This can be removed by removing interaction terms.

k. Finally, take the model you think describes the data the best and write a short report for your grandmother who would like to grow these flowers carefully explaining to her how she should best grow them and why. Note that your grandmother is curious about how much changes in light and timing might affect her flowers and how sensitive her results will be to the settings she makes.

Report

The main goal of this study and analysis is to identify the best combination of light intensity and timing of onset of light treatment. Timing of onset of light treatment is a dummy variable which consist of (Late and Early) while light intensity is a continuous variable which consist of(150,300,450,600,750,900). We further transformed this continuous variable to form a third categorical variable consisting of ('Dimmest','Dimmer', 'Dim', 'Bright', 'Brighter', 'Brightest'). We also added another variable (REPLICATE) which uniquely identified similar combination of timing of onset of light treatment and light intensity.

Technical Analysis

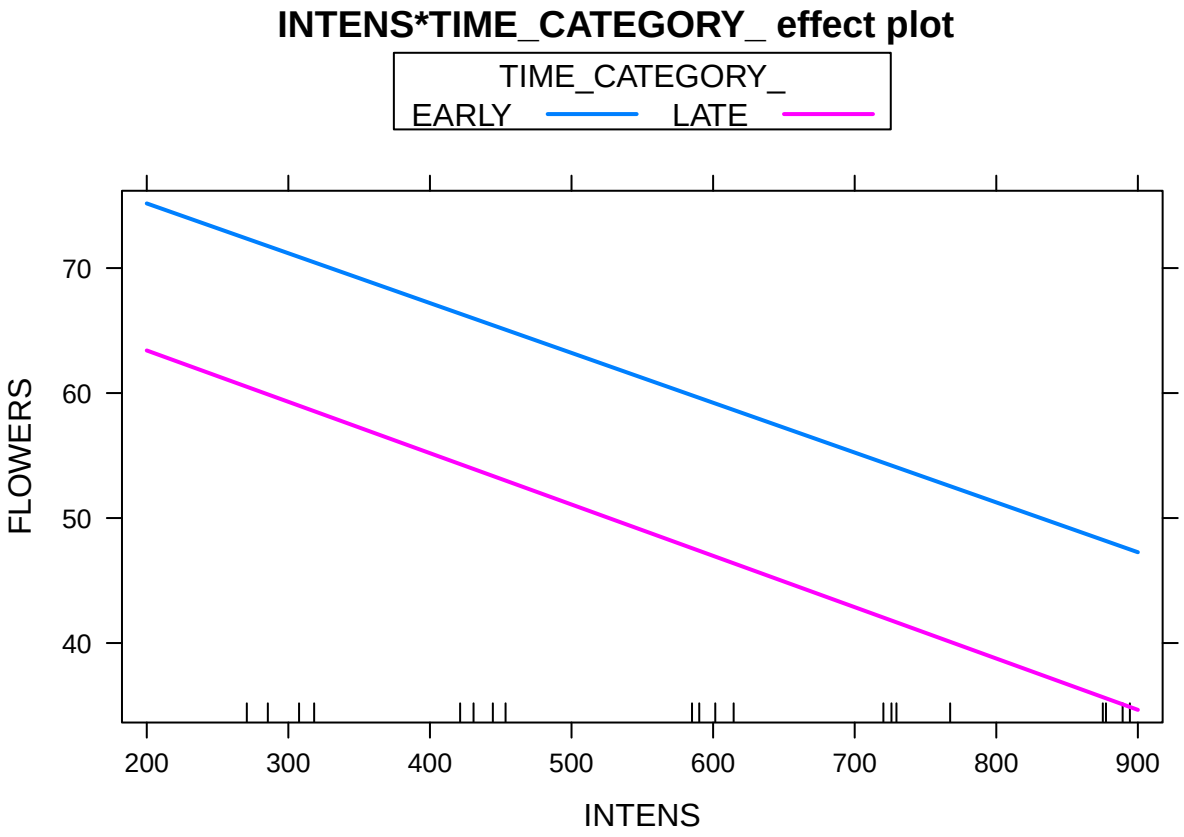
Before fitting any model we went ahead and explored our data by checking whether there was any significant difference between the replicates. T-test p-value of .5787 indicated that the two replicate groups had no significant difference.

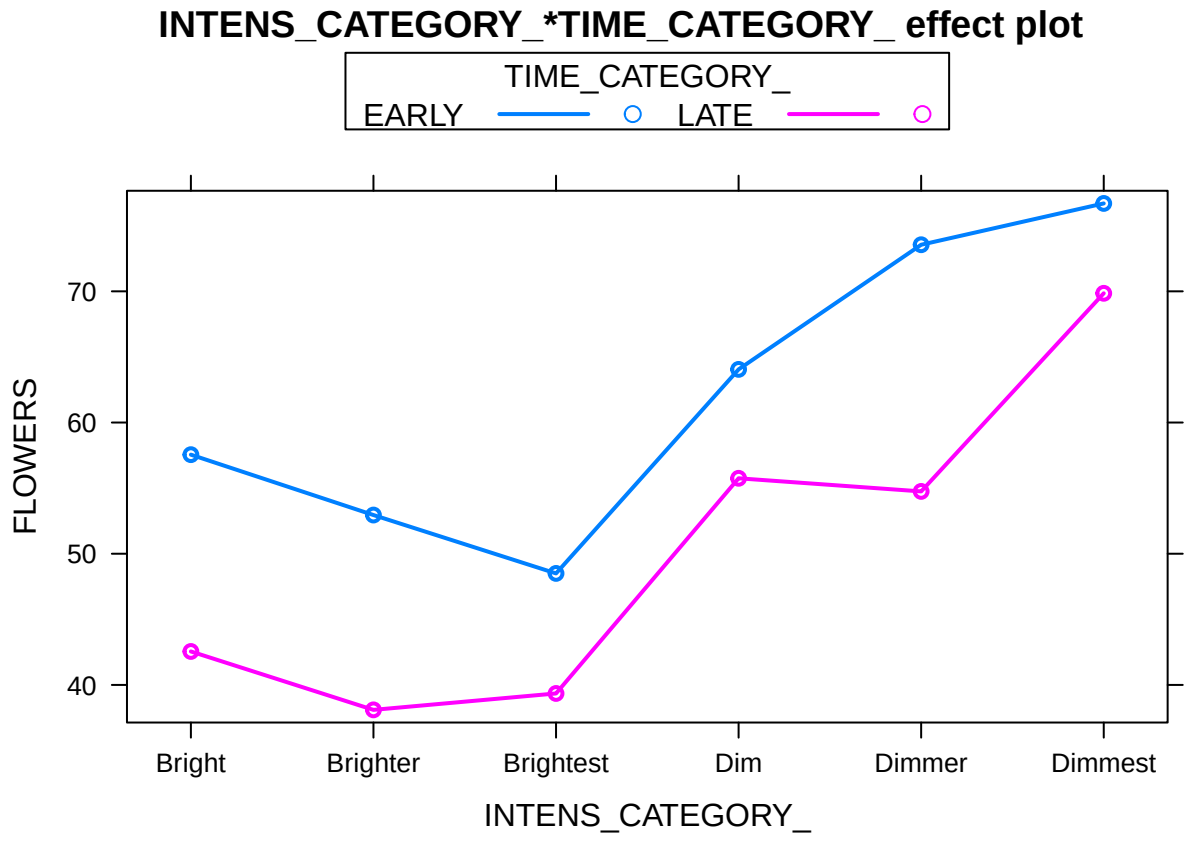
Evaluation whether results differed by replicate

Even after establishing that there was no any significant differences between the 2 groups, we went ahead and added the REPLICATE factor as one of the predictors in the model to ascertain this fact. Adding the REPLICATE factor as one of the predictors in the model exhibited insignificance in the model with a p-value of 0.253968. This confirmed the t-test hence we removed it from the model inorder to achieve a parsimonious model.

From the initial step-by-step analysis, it was well established that we have no significant evidence of interaction between timing of exposure and light intensity as a categorical variable. The p-values were greater than the significant level of .05. In fact from the interaction plot we do not see any form interactions. We do not see any of the lines crossing.

Apart from the categorical case, We also established that we did not have significant evidence of interaction between timing of exposure and light intensity as a continuous variable.


INTENS*TIME_CATEGORY_ effect plot
TIME_CATEGORY_
EARLY
LATE
FLOWERS
70
60
50
40
200
300
400
500
600
700
800
900
INTENS



In fact plotting an interaction plots, we get perfect parallel lines which do not cross (continuous model) and categorical model. These plots further confirmed that we did not have interaction.

After checking for interactions, we went ahead and assessed whether there was any significant difference among the 4 models. The extremely high p-values indicated that we did not have any significant difference among the 4 models.

Model Selection

In order to identify the best model which best describe the relationship between the dependent and independent variables, several assumptions and conditions must be taken into account.

- Linearity: The relationship between independent and the dependent is linear.
- Homoscedasticity: The variance of residual is the same for any value of independent.
- Independence: Observations are independent of each other.
- Normality: Residual must be normally distributed.

Most of the models exhibited all these properties. The next thing is to asses the best fit model with relation to the study question at hand and other factors such parsimony, R-squared, Residual Standard Error, and adj-R squared e.t.c

Since the interaction terms in the 2 models (categorical with interaction and continuous with interaction) were insignificant, we dropped them in favor of the other 2 model without interactions in order to pick the most parsimonious model.

Another compeling reason for dropping them was the fact that they had lower explanatory power (R-squared and adj-R squared) compared to the models without interaction. In the model with interaction term (categorical), only 71% of the variability in the number of flowers is explained by covariates while in the continuous case with interaction 77% of the variability in the number of flowers is explained by the covariates.

Table 15: Models With Interraction Terms!

	-	
	FLOWERS	
	Categorical Model	Continous Model
	(1)	(2)
Observations	24	24
R^2	0.849	0.799
Adjusted R^2	0.710	0.769
Residual Std. Error	7.393 (df = 12)	6.598 (df = 20)
F Statistic	6.124^{***} (df = 11; 12)	26.549^{***} (df = 3; 20)
Note:		*p<0.1; **p<0.05; ***p<0.01
		Coefficients have been removed!

Model Without interraction

Well now we are down to 2 models (Models Without interaction). For starter Both models have high explanatory power (R-squared and adj-R-squared values.

Table 16: Models Without Interraction

	Dependent variable:	
	FLOWERS	
	Categorical Model	Continous Model
	(1)	(2)
INTENS_CATEGORY_Brighter	−4.525 (−13.837, 4.787)	
INTENS_CATEGORY_Brightest	−6.125 (−15.437, 3.187)	
INTENS_CATEGORY_Dim	9.850^{*} (0.538, 19.162)	
INTENS_CATEGORY_Dimmer	14.100^{***} (4.788, 23.412)	
INTENS_CATEGORY_Dimmest	23.225^{***} (13.913, 32.537)	
INTENS		-0.040^{***} (−0.051, −0.030)
TIME_CATEGORY_LATE	-12.158^{***} (−17.535, −6.782)	-12.158^{***} (−17.312, −7.004)
Constant	56.129^{***} (49.017, 63.241)	83.464^{***} (77.048, 89.881)
Observations	24	24
R^2	0.823	0.799
Adjusted R^2	0.761	0.780
Residual Std. Error	6.719 (df = 17)	6.441 (df = 21)
F Statistic	13.181^{***} (df = 6; 17)	41.780^{***} (df = 2; 21)
Note:		*p<0.1; **p<0.05; ***p<0.01
		Notes

In the continuous model, 80% of the variability in the number of flowers is explained by covariates(Light intensity and timing of treatment) while in the categorical model 82 % of the variability in the number of flower is explained by the covariates. Adjusted r-squared in the continuous model was higher than categorical model, possibly due to the fact that we have more coefficient in the categorical model. Therefore the continuous model is more parsimonious compared to the categorical model.

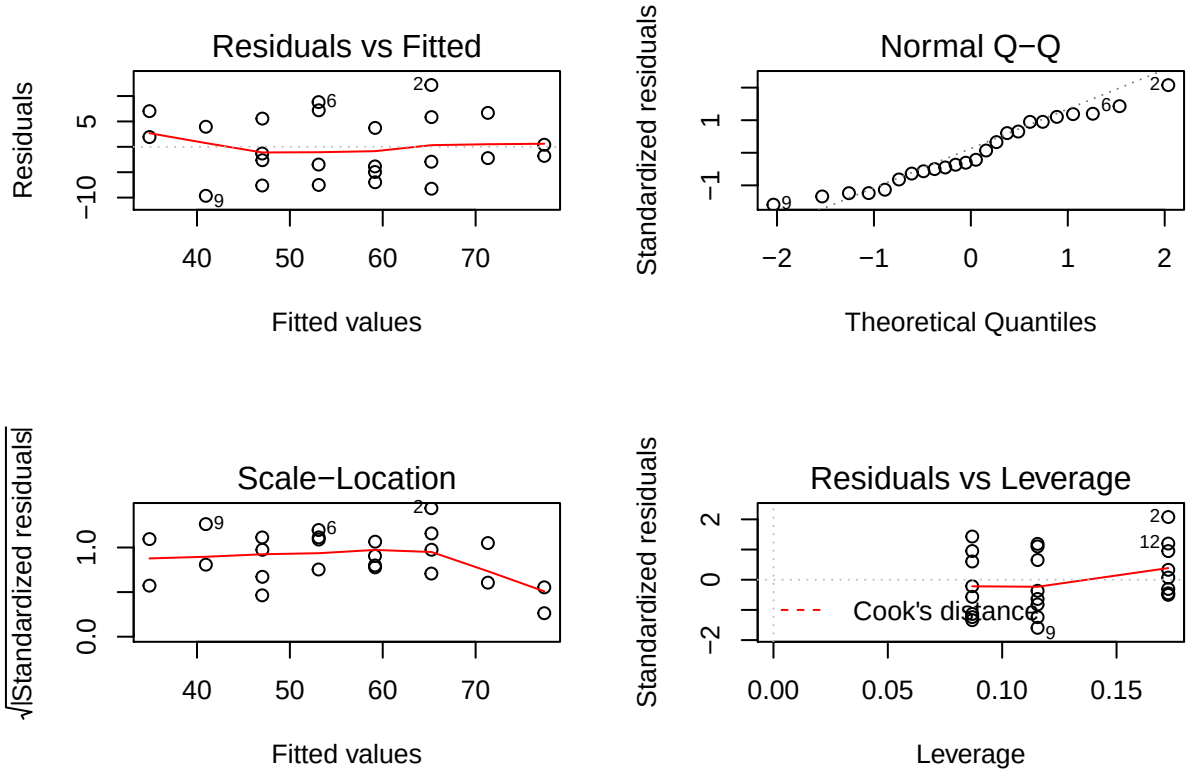
The best model that best describe the data is: **Continous model without interaction** because:

- This model accomplishes the desired level of explanatory power (80%) with as few predictor variables

as possible compared to the latter model. - parsimony

- This model has a higher adjusted R-Squared 0.78 compared to 0.76 of the latter. In the categorical model, some of the predictors were highly insignificant like (brighter and brightest). This is one of the reasons why we have a higher adjusted R-Squared in the categorical model.
- This model has lower residual standard error 6.441 compared to 6.719 in the categorical model. On average, the response variable (Flowers) will deviate from the true regression line by 6.441 compared to 6.719 in model 2
- This model has highly significant predictors. The categorical model has 2 insignificant predictors included in the model. The rule of thumb is to remove, insignificant predictors from your model.

Model diagnostics



Residuals vs. fitted values

- We observe an acceptable Scatter-plot in the sense that the fitted values doesn't show any noticeable systematic pattern hence doesn't contain any "leaking" explanatory information.
- From the plot, the residuals bounce randomly around the 0 line and roughly forms a horizontal band around the 0 line. This suggests that the variances of the error terms are approximately equal.

Normal Q-Q Plot

- This plot shows that the residuals are normally distributed since the data points are well arranged on the straight dashed line. No noticeable deviation from the straight line can be observed.

Scale-Location Plot

- This plot confirms the assumption of equal variance (homoscedasticity). The residuals appear randomly scattered with no concerning patterns. We have a smooth horizontal red line confirming this fact.

Residual vs Levarage Plot

- We do not have datapoints outside the Cook's distance lines. In fact the red dashed lines is not visible. This suggest we have no influential outliers on the regression (leverage points).

Summary

In summary the selected model tells us that:

- If we held other predictors constant, an increase of light intensity by 1 unit results to a decrease in the number of flowers by 0.04. The extremely low p-value (<0.05) indicates that is highly significant.
- If we held other predictors constant, "Late" timing for light exposure results to a decrease in number of flowers by 12. the p-value of (0.000146) indicates that this is statistically significant

This means that light intensity is inversely proportional to the number of flowers. At the same time, Late timing of onset of light treatment result to a decrease in number of flowers.

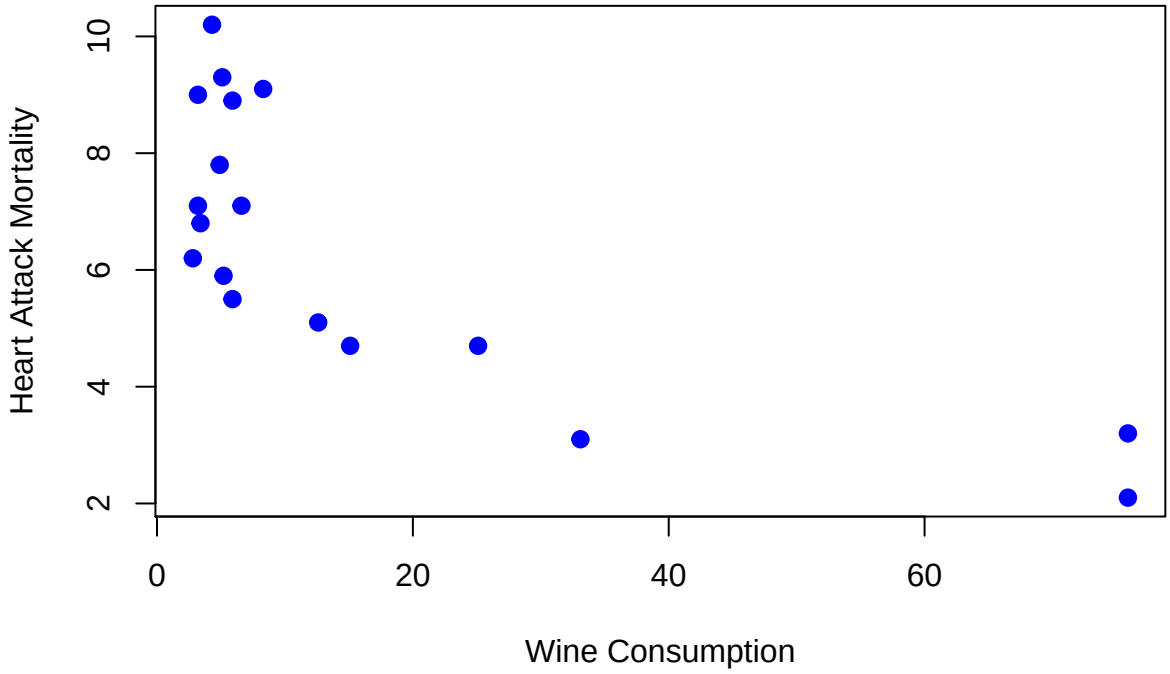
Grandmothers Report

The goal of the analysis was to identify the best combination of light intensity and timing of onset of light treatment that would yield highest number of flowers. Therefore, in order to get the highest number of flowers, grandmother needs to expose her flowers to the dimmest light intensity as early (timing) as possible.

Source Code

```
#' ---
#' title: "Homework 1"
#' author: "Allan Kimaina"
#' date: "February 5, 2018"
#' output:
#'   pdf_document: default
#'   html_document: default
#' ---
#'
#'


# load package
library(sjPlot)
library(sjmisc)
library(sjlabelled)
library(ggpubr)
library(ggpmisc)
library("gridExtra")
library(stargazer)
library(e1071)
library(jtools)
library(effects)
library(multcompView)
library(ggplot2)
library(ggrepel)
library(dplyr)
library(car)

# import data
wine = read.csv("Wine.csv")
colnames(wine)[1] <- "COUNTRY"


#'
#'
#' # Question 1
#'
#' The data in the file wine.csv (in the data sets folder on Canvas) give the average wine consumption
#' Do these data suggest that heart disease death rates are associated with average wine consumption? I
#'
#' Do any countries have substantially higher or lower death rates than others with similar wine consum
#'
#' Analyze the data and write a brief report that includes a summary of findings, a graphical display a
#'
#'
#'
#' The goal of the study is to determine the significance of the relationship between wine consumption
#'
#'
```

```r
select_similar <- wine
plot(wine$WINE, wine$MORTALITY,
      #main= "Absolute Losses vs. Relative Losses(in%)",
      xlab= "Wine Consumption",
      ylab= "Heart Attack Mortality",
      col= "blue", pch = 19, cex = 1, lty = "solid", lwd = 2)
```



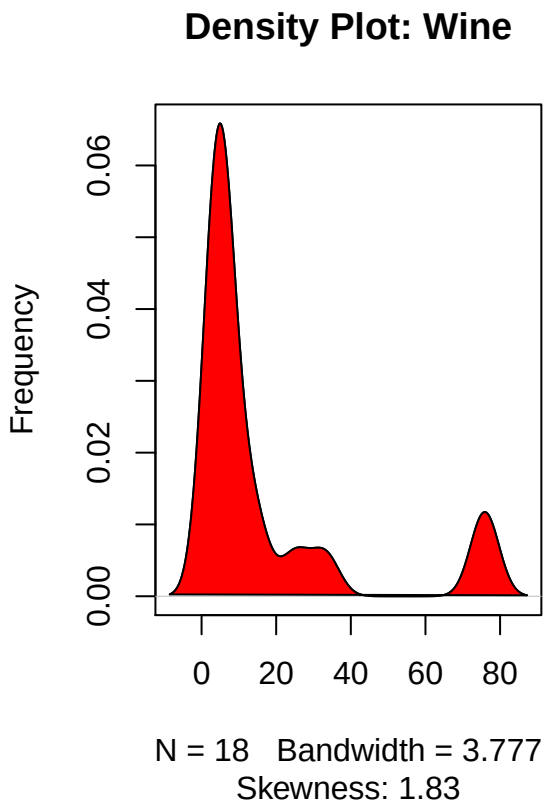

```r
#'
#'
#'
#'
#'
wine.model = lm(MORTALITY ~ WINE, data = wine)
#cor(wine$WINE, wine$MORTALITY)  # calculate correlation between WINE and MORTALITY
#'
#'
#' From the scatter plot, the data points suggest that there is a strong negative  association between
#'
#' But first before establishing the best model that fits this association, lets try to understand the
#'
#' #### Data Examination & Exploration
#'
#'

par(mfrow=c(1, 2))  # divide graph area in 2 columns
plot(density(wine$WINE), main="Density Plot: Wine", ylab="Frequency", sub=paste("Skewness:", round(e107
```

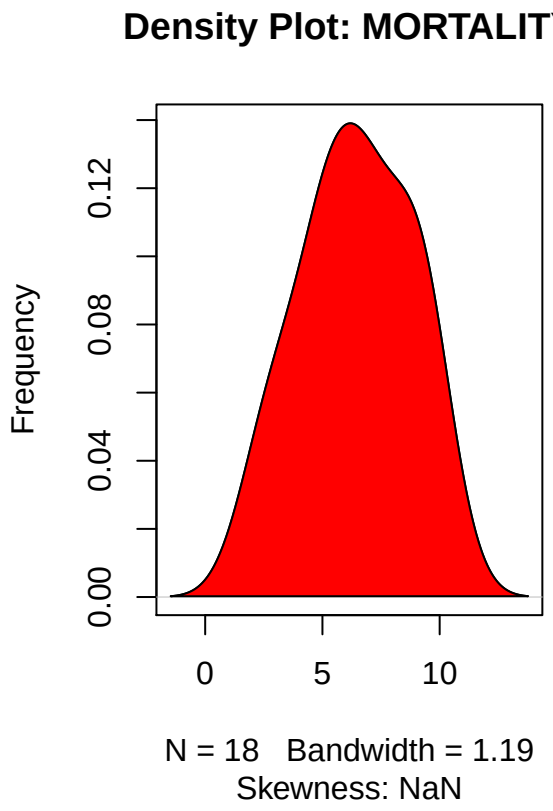
```
polygon(density(wine$WINE), col="red")
plot(density(wine$MORTALITY), main="Density Plot: MORTALITY", ylab="Frequency", sub=paste("Skewness:", r
```

```
## Warning in is.na(x): is.na() applied to non-(list or vector) of type 'NULL'
```

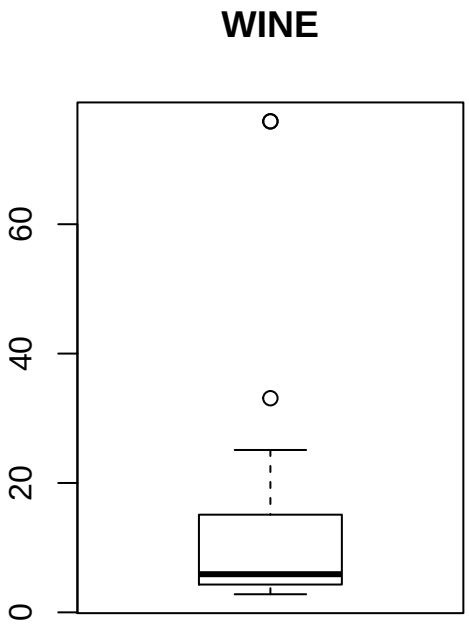
```
## Warning in mean.default(x): argument is not numeric or logical: returning
## NA
```

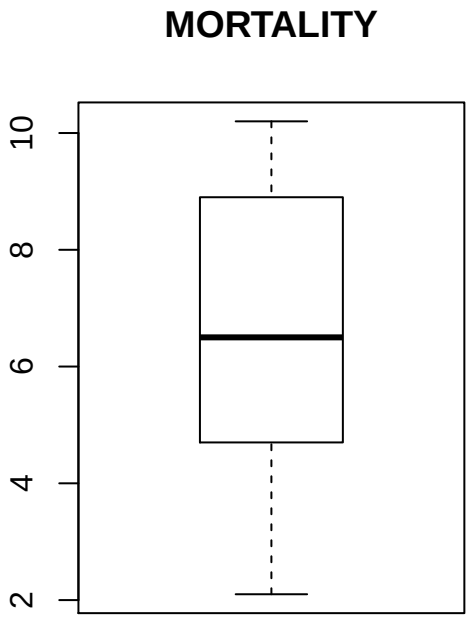
```
polygon(density(wine$MORTALITY), col="red")
```






```
#'
#'
#' Looking at the density plot, the response variable is normal without any skewness, however, the pred
#'
#' Let's see if there is any data-point which lies outside the 1.5 * distance between the 25th percenti
#'
#' ##### Outlier Detection
#'
#'
par(mfrow=c(1, 2))  # divide graph area in 2 columns
boxplot(wine$WINE, main="WINE", sub=paste("Outlier rows: ", boxplot.stats(wine$WINE)$out))  # box plot
boxplot(wine$MORTALITY, main="MORTALITY", sub=paste("Outlier rows: ", boxplot.stats(wine$MORTALITY)$out
```





Outlier rows: 35 9

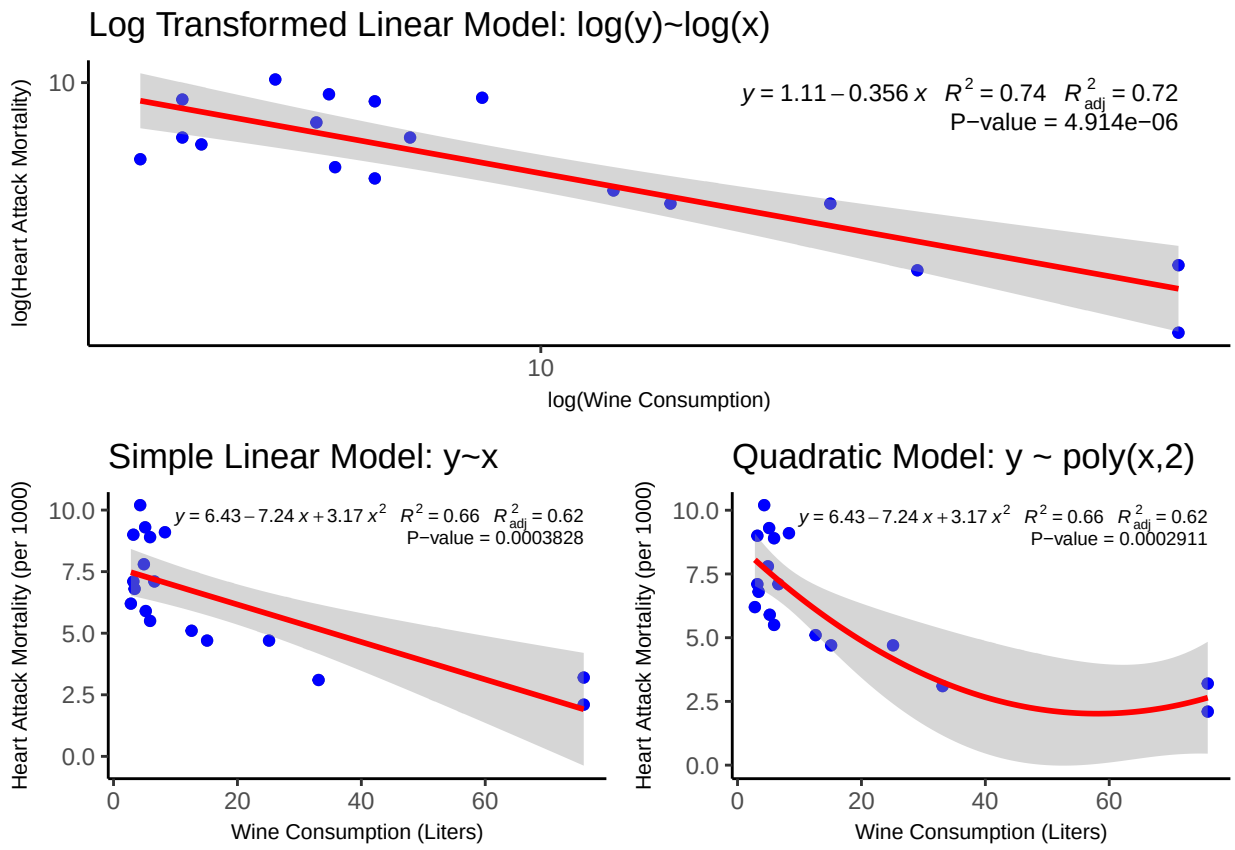
Outlier rows:

```
#'
#'
#' The box plot suggest that there are two outliers, France and Italy. Even though France and Italy are
#'
#' ##### Exponential relationship
#' From the above scatter plot and density plot, the explanatory variable (WINE CONSUMPTION) is strongl
#'
#'
wine.model.transformed <- lm( log(wine$MORTALITY)~log(wine$WINE))
plot(log(wine$WINE), log(wine$MORTALITY),
      #main= "Absolute Losses vs. Relative Losses(in%)",
      xlab= "log(Wine Consumption)",
      ylab= "log(Heart Attack Mortality)",
      main= "Log Transformation Plot",
      col= "blue", pch = 19, cex = 1, lty = "solid", lwd = 2)
abline(wine.model.transformed,  col="red")

#summary(wine.model.transformed)

#'
#'
#' Transforming both predictor and response variable we get a perfect linear relationship implying a li
#'
#' Another possible explanation for the curvilinear relationship could be because the relationship betw
#'
wine.model.poly<-lm(wine$MORTALITY~ poly(wine$WINE, 2) )
```

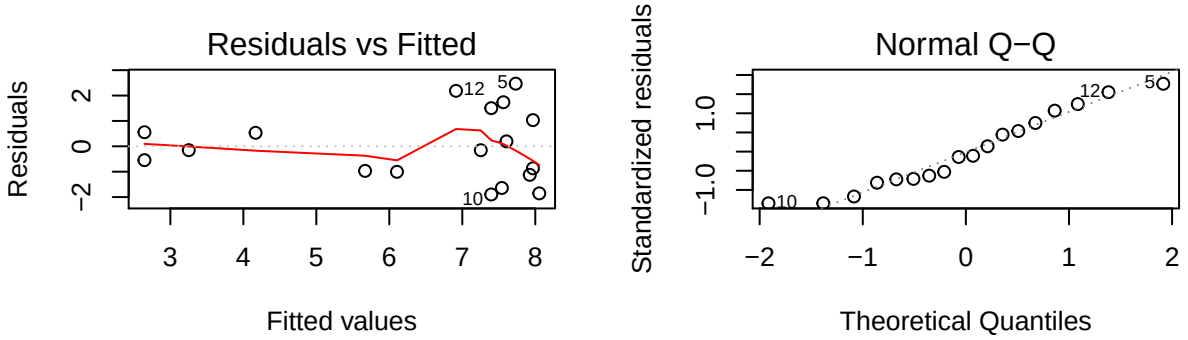
```
ggplot(wine, aes(y=MORTALITY, x=WINE)) +
  geom_point(alpha = .5) +
  geom_point(color = "blue") +
  stat_smooth(method = "lm", formula = y ~ poly(x,2), color="red")+
  labs(x='Wine Consumption (Liters)', y='Heart Attack Mortality (per 1000)')+
  labs(title = "Quadratic Plot")+
  theme_classic(base_size = 18)+ guides(fill=FALSE)


#'
#'
#' There is little reason to believe that the relationship between wine consumption and heart attack mo
#'
#' #### Possible Models
#'
#' From the previous section, we have been able to identify 3 possible models
#'
#'  1. Simple Linear model
#'  2. Log Transformed Linear Model (exponential)
#'  3. Quadratic model (highly unlikely)
#'
#'
#'
  simple_lm_plot <- ggplot(wine, aes(y=MORTALITY, x=WINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    stat_smooth(method = "lm",  color="red")+
    labs(x='Wine Consumption (Liters)', y='Heart Attack Mortality (per 1000)',
         title = "Simple Linear Model: y~x")+
    stat_poly_eq(aes(label = paste(..eq.label.., ..rr.label.., ..adj.rr.label.., sep = "~~~")),
                 label.x.npc = "right", label.y.npc = 0.87, geom = "text",
                 formula =  y ~ poly(x,2), parse = TRUE, size = 2.5)+
    stat_fit_glance(method = "lm",
                    method.args = list(formula = y ~ x),
                    geom = "text", size = 2.5,
                    label.y.npc = 0.85, label.x.npc = "right",
                    aes(label = paste("P-value = ", signif(..p.value.., digits = 4), sep = "")))+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)

    )

  log_lm_plot<-ggplot(wine, aes(y=MORTALITY, x=WINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    stat_smooth(method = "lm",  color="red")+
    labs(x='log(Wine Consumption)', y='log(Heart Attack Mortality)',
         title = "Log Transformed Linear Model: log(y)~log(x)")+
    stat_poly_eq(aes(label = paste(..eq.label.., ..rr.label.., ..adj.rr.label.., sep = "~~~")),
                 label.x.npc = "right", label.y.npc = 0.87, geom = "text",
```

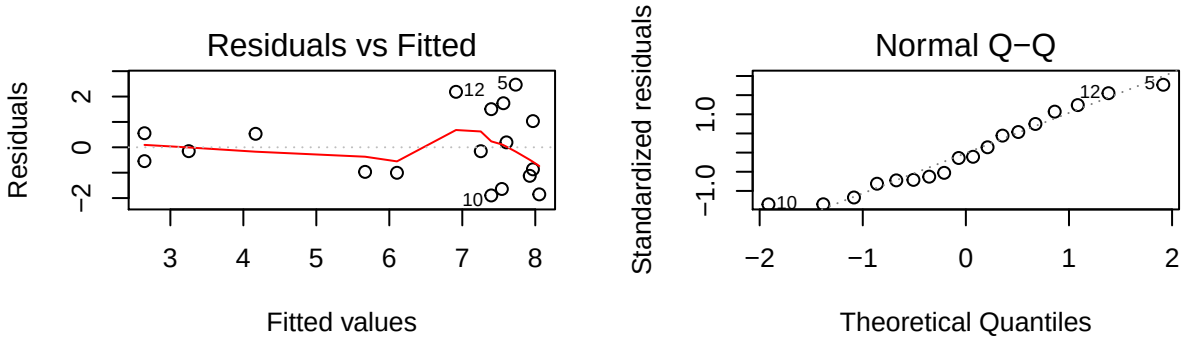
```
                formula =  y ~x, parse = TRUE, size = 3)+
    stat_fit_glance(method = "lm",
                    method.args = list(formula = y~x),
                    geom = "text", size = 3,
                    label.y.npc = 0.85, label.x.npc = "right",
                    aes(label = paste("P-value = ", signif(..p.value.., digits = 4), sep = "")))+
    scale_x_log10() +
    scale_y_log10()+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )

  quadratic_plot <- ggplot(wine, aes(y=MORTALITY, x=WINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    stat_smooth(method = "lm", formula = y ~ poly(x,2), color="red")+
    labs(x='Wine Consumption (Liters)', y='Heart Attack Mortality (per 1000)',
         title = "Quadratic Model: y ~ poly(x,2)")+
    stat_poly_eq(aes(label = paste(..eq.label.., ..rr.label.., ..adj.rr.label.., sep = "~~~")),
                 label.x.npc = "right", label.y.npc = 0.87, geom = "text",
                 formula =  y ~ poly(x,2), parse = TRUE, size = 2.5)+
    stat_fit_glance(method = "lm",
                    method.args = list(formula = y ~ poly(x,2)),
                    geom = "text", size = 2.5,
                    label.y.npc = 0.85, label.x.npc = "right",
                    aes(label = paste("P-value = ", signif(..p.value.., digits = 4), sep = "")))+

    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )



  grid.arrange(log_lm_plot, arrangeGrob(simple_lm_plot,
                            quadratic_plot, ncol = 2), # Second row with 2 plots in 2 different columns
               nrow = 2.)                          # Number of rows

#'
#'
#' well, explanatory power and residual plot can help us  determine which model fits well the relations
#'
#'
#' ### Model Selection:
#'
#' Both simple Linear model (y~x) and quadratic model (y~poly(x,2)) have really low explanatory power (
#' the response is explained by the predictor.
#'
```

```
#'
par(mfrow = c(2, 2))
```




```
plot(wine.model.poly, which=c(2,1))
plot(wine.model.poly, which=c(2,1))
```



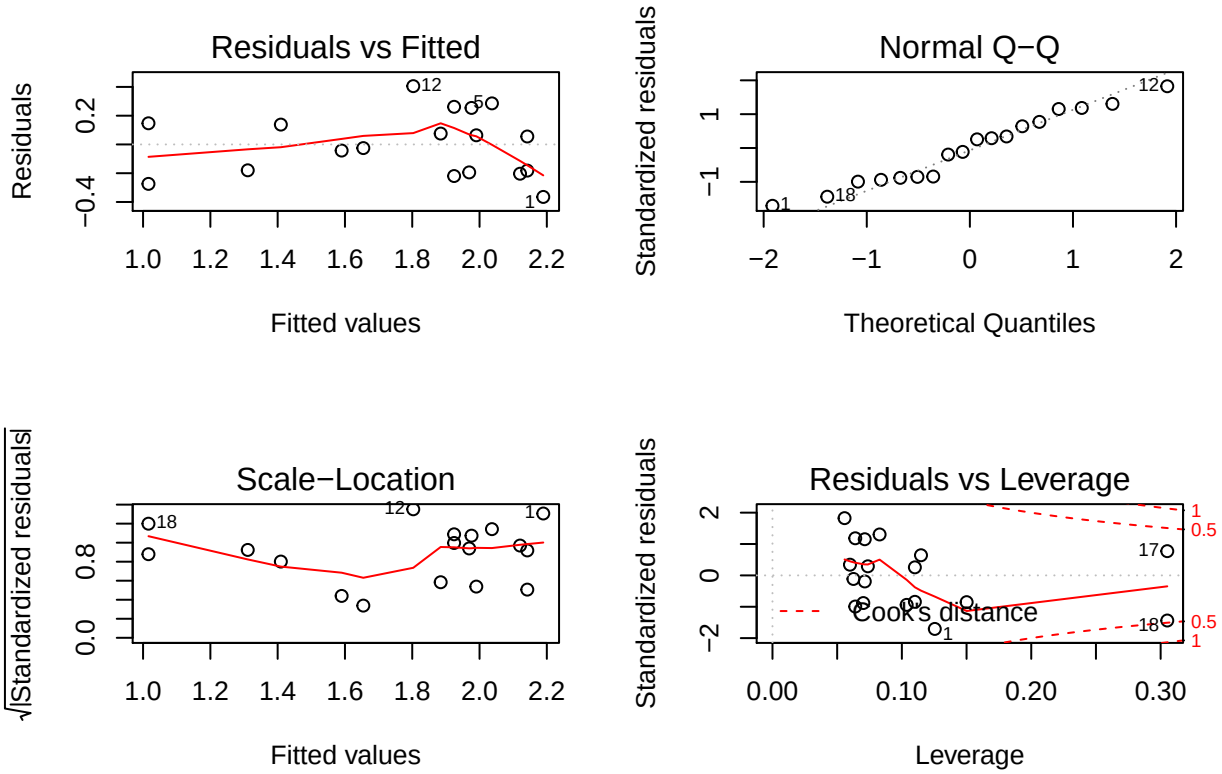



```
#'
#'
#' Furthermore, the residual plot in both model seems to follow a consistent trend at the same time exh
#'
#' ####  Log Transformed Linear Model (exponential)
#'
#'
#'

stargazer(wine.model.transformed,
          header=F,
          type = "latex",
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & MORTALITY) \\
## \hline \\[-1.8ex]
```

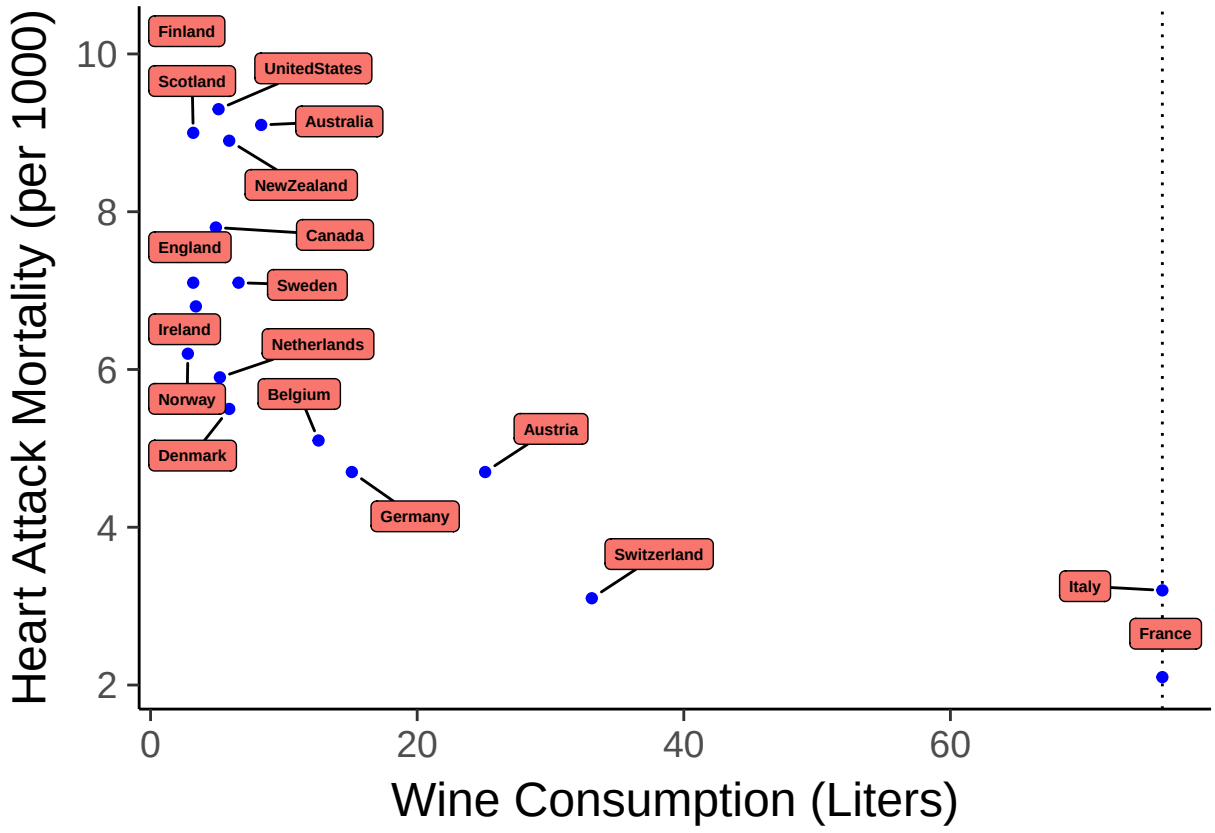
```
##  WINE) & $-$0.356$^{***}$ (0.053) \\
##   Constant & 2.556$^{***}$ (0.127) \\
##  \hline \\[-1.8ex]
## Observations & 18 \\
## R$^{2}$ & 0.738 \\
## Adjusted R$^{2}$ & 0.722 \\
## Residual Std. Error & 0.229 (df = 16) \\
## F Statistic & 45.170$^{***}$ (df = 1; 16) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
#'
#'
#'
#' Graphically we can notice that the log transformed regression line covers most data-points as well a
#'
#'
#'
#' ### Model Diagnostic:
#'
#'
par(mfrow = c(2, 2))
plot(wine.model.transformed)
```



```
#'
#'
#'
#' **Residuals vs. fitted values**
#'
#' * We observe an acceptable Scatter-plot in the sense that the fitted values doesn't show any noticea
#' * From the plot, the residuals bounce randomly around the 0 line and roughly forms a horizontal band
#' * The plot suggest we have noticeable outliers. This is not concerning since not all outliers are in
#'
#' **Normal Q-Q Plot**
#'
#' * This plot shows that the residuals are normally distributed since the data points are  well arrang
#'
#' **Scale-Location Plot**
#'
#' * This plot confirms the assumption of equal variance (homoscedasticity). The residuals appear rando
#'
#' **Residual vs Levarage Plot**
#'
#' *  We have no datapoints outside the Cook's distance lines (red dashed lines). This suggest we have
#'
#' ### Summary
#'
#'
#' **Do these data suggest that heart disease death rates are associated with average wine consumption?
#'
#' The data suggest that wee have a highly significant negative relationship between  heart attack mort
#'
#' \begin{center}
#'
#'  \(log(y) =  \beta_{0} + \beta_{1} log(x) + \epsilon\)
#'
#'  \(log(HeartAttackMortality) =  \beta_{0} + \beta_{1} log(WineConsumption) + \epsilon\)
#'
#'   \(log(HeartAttackMortality) =  2.556 - 0.356*log(WineConsumption) + \epsilon\)
#'
#' \end{center}
#'
#' **What does this tell us?**
#'
#' This implies that there is a multiplicative change between the response and predictor.
#'
#' *  **Multiplicative changes in e** - Multiplying WineConsumption by "e" will decrease HeartAttackMor
#' *  **A 1% increase in WineConsumption** - A 1% increase in Wine Consumption will decrease Heart Atta
#' *  **A 10% increase in WineConsumption** - A 10% increase in Wine Consumption will decrease Heart At
#'
#'
#' This tells us that an increase of wine consumption by 10% will decrease Ischemic heart attack mortal
#'
#'
#'
#'
#' **Do any countries have substantially higher or lower death rates than others with similar wine cons
```

```
#'
#'
#'

ggplot(wine, aes(WINE, MORTALITY, label = COUNTRY)) +
   geom_vline(xintercept =  c(75.90), linetype = 3) +
  geom_point(color = "blue") +
  geom_label_repel(
     aes(
    fill = "black"),
  fontface = 'bold', color = 'black',
  size = 2,
  box.padding = unit(0.25, "lines"),
  point.padding = unit(0.5, "lines")

  ) +
  labs(x='Wine Consumption (Liters)', y='Heart Attack Mortality (per 1000)')+
  theme_classic(base_size = 18)+ guides(fill=FALSE)
```




```
#'
#'
#' Yes, both the data table and the plot above indicate that some countries have substantially higher o
#'
#' * Australia and German have same heart mortality (4.7) yet wine consumption in Austria (25.1) is way
#' * France and Italy have same wine consumption of (75.9) yet heart mortality in France (2.1) is lower
#' * Scotland and England have same wine consumption of (3.2) yet heart mortality in Scotland (9.0) is
```

```r
#' * United States (5.1) and Netherlands (5.2) have similar wine consumption  yet heart mortality in Un
#'
#' **What does this tell us?**
#'
#' There are some variables that were not collected or included in the dataset that would aid us in ext
#'
#'
#' \onecolumn
#'
#' # Question 2
#'
#' Meadowfoam is a small plant that grows in Pacific Northwest and is domesticated for its seed oil. A
#'
#' a. First put the data into a dataset with four variables: number of flowers, light intensity, timing
#'
flowers = read.csv("Flowers.csv")
#replicate <- flowers%>%mutate(replicate = match(INTENS, unique(INTENS)))

flowers$REPLICATE=as.factor(rep(c(1,2),12))

t.test.value <- t.test(data =flowers, FLOWERS~REPLICATE)
# Since the p-value
# print
stargazer(flowers,
          header=F,
          type = "latex",
          no.space = T,
          summary = F,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} ccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & FLOWERS & TIME & INTENS & REPLICATE \\
## \hline \\[-1.8ex]
## 1 & $62.300$ & $1$ & $150$ & 1 \\
## 2 & $77.400$ & $1$ & $150$ & 2 \\
## 3 & $55.300$ & $1$ & $300$ & 1 \\
## 4 & $54.200$ & $1$ & $300$ & 2 \\
## 5 & $49.600$ & $1$ & $450$ & 1 \\
## 6 & $61.900$ & $1$ & $450$ & 2 \\
## 7 & $39.400$ & $1$ & $600$ & 1 \\
## 8 & $45.700$ & $1$ & $600$ & 2 \\
## 9 & $31.300$ & $1$ & $750$ & 1 \\
## 10 & $44.900$ & $1$ & $750$ & 2 \\
## 11 & $36.800$ & $1$ & $900$ & 1 \\
## 12 & $41.900$ & $1$ & $900$ & 2 \\
## 13 & $77.800$ & $2$ & $150$ & 1 \\
## 14 & $75.600$ & $2$ & $150$ & 2 \\
```

```
## 15 & $69.100$ & $2$ & $300$ & 1 \\
## 16 & $78$ & $2$ & $300$ & 2 \\
## 17 & $57$ & $2$ & $450$ & 1 \\
## 18 & $71.100$ & $2$ & $450$ & 2 \\
## 19 & $62.900$ & $2$ & $600$ & 1 \\
## 20 & $52.200$ & $2$ & $600$ & 2 \\
## 21 & $60.300$ & $2$ & $750$ & 1 \\
## 22 & $45.600$ & $2$ & $750$ & 2 \\
## 23 & $52.600$ & $2$ & $900$ & 1 \\
## 24 & $44.400$ & $2$ & $900$ & 2 \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
#'
#'
#' ##### *T.test Value*
#'

t.test.value <- t.test(data =flowers, FLOWERS~REPLICATE)

t.test.value
```

```
##
##  Welch Two Sample t-test
##
## data:  FLOWERS by REPLICATE
## t = -0.56369, df = 21.949, p-value = 0.5787
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.013752   8.597085
## sample estimates:
## mean in group 1 mean in group 2
##        54.53333        57.74167
```

```
#'
#'
#' We fail to reject the null hypothesis since the p-value of .5787 indicates that the two replicate gr
#'
#'
#'
#' b. Create a categorical form of the light intensity with 6 categories.
#'
# cat form
INTENS_CATEGORY <- data.frame(
   INTENS=c(150,300,450,600,750,900),
  INTENS_CATEGORY_ =c('Dimmest','Dimmer', 'Dim', 'Bright', 'Brighter', 'Brightest')
)


# light intensity
 flowers.df<- flowers%>%inner_join(INTENS_CATEGORY, by='INTENS') %>% # add categories
  mutate(TIME_CATEGORY_=as.factor(ifelse(TIME==2,  "EARLY","LATE"))) # add

 # print out the new categories
```

```
stargazer(INTENS_CATEGORY,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} ccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & INTENS & INTENS\_CATEGORY\_ \\
## \hline \\[-1.8ex]
## 1 & $150$ & Dimmest \\
## 2 & $300$ & Dimmer \\
## 3 & $450$ & Dim \\
## 4 & $600$ & Bright \\
## 5 & $750$ & Brighter \\
## 6 & $900$ & Brightest \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
#'
#'
#'
#' c. The research questions are: What are effects of intensity and timing? Is there an interaction bet
#'
#' d. First create an analysis of variance using timing and the categorical form of the light intensity
#'


flowers.model.cat <- lm(data=flowers.df, FLOWERS~INTENS_CATEGORY_+TIME_CATEGORY_)

stargazer(flowers.model.cat,
          header=F,
          type = "latex",
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & FLOWERS \\
```

```
## \hline \\[-1.8ex]
##  INTENS\_CATEGORY\_Brighter & $-$4.525 (4.751) \\
##   INTENS\_CATEGORY\_Brightest & $-$6.125 (4.751) \\
##   INTENS\_CATEGORY\_Dim & 9.850$^{*}$ (4.751) \\
##   INTENS\_CATEGORY\_Dimmer & 14.100$^{***}$ (4.751) \\
##   INTENS\_CATEGORY\_Dimmest & 23.225$^{***}$ (4.751) \\
##   TIME\_CATEGORY\_LATE & $-$12.158$^{***}$ (2.743) \\
##   Constant & 56.129$^{***}$ (3.629) \\
##  \hline \\[-1.8ex]
## Observations & 24 \\
## R$^{2}$ & 0.823 \\
## Adjusted R$^{2}$ & 0.761 \\
## Residual Std. Error & 6.719 (df = 17) \\
## F Statistic & 13.181$^{***}$ (df = 6; 17) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
  #summary(flowers.model.cat)

# from the p-value it seems that very bright and slightly bright do not have any significant effects on

# talk about timing?



#'
#'
#' * *INTENS_CATEGORY_Brighter* - If we held  other predictors constant, under a "brighter" light inten
#' * *INTENS_CATEGORY_Brightest* - If we held  other predictors constant, under the "brightest" light i
#' * *INTENS_CATEGORY_Dim* - If we held  other predictors constant, under a "Dim" light intensity condi
#' * *INTENS_CATEGORY_Dimmer* - If we held  other predictors constant, under a "Dimmer" light intensity
#' * *INTENS_CATEGORY_Dimmest* - If we held  other predictors constant, under the "Dimmest" light inten
#' * *TIME_CATEGORY_LATE* - If we hold  other predictors constant,  "Late" onset timing of light treatm
#'
#' Looking at the p-values, light intensity condition ("Dimmer", "Dimmest" ) and Timing ("Late") are st
#'
#' ##### *Significance of REPLICATE in the model*
#'

# What is the effect of the treatment on the value ?
flowers.model.with.replicate <- lm(data=flowers.df, FLOWERS~INTENS_CATEGORY_+TIME_CATEGORY_+REPLICATE)

stargazer(flowers.model.with.replicate,
          header=F,
          type = "latex",
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
```

```
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & FLOWERS \\
## \hline \\[-1.8ex]
##  INTENS\_CATEGORY\_Brighter & $-$4.525 (4.696) \\
##   INTENS\_CATEGORY\_Brightest & $-$6.125 (4.696) \\
##   INTENS\_CATEGORY\_Dim & 9.850$^{*}$ (4.696) \\
##   INTENS\_CATEGORY\_Dimmer & 14.100$^{***}$ (4.696) \\
##   INTENS\_CATEGORY\_Dimmest & 23.225$^{***}$ (4.696) \\
##   TIME\_CATEGORY\_LATE & $-$12.158$^{***}$ (2.711) \\
##   REPLICATE2 & 3.208 (2.711) \\
##   Constant & 54.525$^{***}$ (3.834) \\
##  \hline \\[-1.8ex]
## Observations & 24 \\
## R$^{2}$ & 0.837 \\
## Adjusted R$^{2}$ & 0.766 \\
## Residual Std. Error & 6.641 (df = 16) \\
## F Statistic & 11.764$^{***}$ (df = 7; 16) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
#summary(flowers.model.with.replicate)

#'
#'
#' Adding the REPLICATE factor as one of the predictors in the model exhibited insignificance in the mo
#'
#'
#' e. Then create an interaction between light intensity and timing by multiplying the two variables an
#'
flowers.model.cat.int <- lm(data=flowers.df, FLOWERS~INTENS_CATEGORY_*TIME_CATEGORY_)
stargazer(flowers.model.cat.int,
          header=F,
          type = "latex",
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & FLOWERS \\
```

```
## \hline \\[-1.8ex]
##  INTENS\_CATEGORY\_Brighter & $-$4.600 (7.393) \\
##   INTENS\_CATEGORY\_Brightest & $-$9.050 (7.393) \\
##   INTENS\_CATEGORY\_Dim & 6.500 (7.393) \\
##   INTENS\_CATEGORY\_Dimmer & 16.000$^{*}$ (7.393) \\
##   INTENS\_CATEGORY\_Dimmest & 19.150$^{**}$ (7.393) \\
##   TIME\_CATEGORY\_LATE & $-$15.000$^{*}$ (7.393) \\
##   INTENS\_CATEGORY\_Brighter:TIME\_CATEGORY\_LATE & 0.150 (10.456) \\
##   INTENS\_CATEGORY\_Brightest:TIME\_CATEGORY\_LATE & 5.850 (10.456) \\
##   INTENS\_CATEGORY\_Dim:TIME\_CATEGORY\_LATE & 6.700 (10.456) \\
##   INTENS\_CATEGORY\_Dimmer:TIME\_CATEGORY\_LATE & $-$3.800 (10.456) \\
##   INTENS\_CATEGORY\_Dimmest:TIME\_CATEGORY\_LATE & 8.150 (10.456) \\
##   Constant & 57.550$^{***}$ (5.228) \\
##  \hline \\[-1.8ex]
## Observations & 24 \\
## R$^{2}$ & 0.849 \\
## Adjusted R$^{2}$ & 0.710 \\
## Residual Std. Error & 7.393 (df = 12) \\
## F Statistic & 6.124$^{***}$ (df = 11; 12) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
#summary(flowers.model.cat.int)

# while for the model without interaction R-squared:  0.823,    Adjusted R-squared:  0.761
# for the model with interactions  we have an R2 of  .849  and adj. R2  of .710
# No negligable differnec between the 2 r squared hence No strong evidence of interaction between light
#'
#'
#' The p-values are greater than the significant level of .05, therefore none of the interaction terms
#'
#' f. Now repeat the process but using light intensity as a continuous variable.
#'
#' ##### *Continous Without Interaction Terms*
#'
#'
flowers.model.cont <- lm(data=flowers.df, FLOWERS~INTENS+TIME_CATEGORY_)
#summary(flowers.model.cont)
#sjt.lm(flowers.model.cont,  show.fstat = TRUE)
stargazer(flowers.model.cont,
          header=F,
          type = "latex",
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
```

```
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & FLOWERS \\
## \hline \\[-1.8ex]
##  INTENS & $-$0.040$^{***}$ (0.005) \\
##   TIME\_CATEGORY\_LATE & $-$12.158$^{***}$ (2.630) \\
##   Constant & 83.464$^{***}$ (3.274) \\
##  \hline \\[-1.8ex]
## Observations & 24 \\
## R$^{2}$ & 0.799 \\
## Adjusted R$^{2}$ & 0.780 \\
## Residual Std. Error & 6.441 (df = 21) \\
## F Statistic & 41.780$^{***}$ (df = 2; 21) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
# R2 / adj. R2      .799 / .780


#'
#'
#' * *INTENS* - If we held  other predictors constant, an increase of light intensity by 1 unit  result
#' * *TIME_CATEGORY_LATE* - If we held  other predictors constant, "Late" timing for light exposure  re
#'
#' This means that light intensity is inversely proportional to number of flowers. At the same time, La
#'
#' ##### *Continous with Interaction Terms*
#'
#'

flowers.model.cont.int <- lm(data=flowers.df, FLOWERS~INTENS*TIME_CATEGORY_)
#summary(flowers.model.cont.int)
#sjt.lm(flowers.model.cont.int,  show.fstat = TRUE)
stargazer(flowers.model.cont.int,
          header=F,
          type = "latex",
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & FLOWERS \\
## \hline \\[-1.8ex]
```

```
##  INTENS & $-$0.040$^{***}$ (0.007) \\
##   TIME\_CATEGORY\_LATE & $-$11.523$^{*}$ (6.142) \\
##   INTENS:TIME\_CATEGORY\_LATE & $-$0.001 (0.011) \\
##   Constant & 83.147$^{***}$ (4.343) \\
##  \hline \\[-1.8ex]
## Observations & 24 \\
## R$^{2}$ & 0.799 \\
## Adjusted R$^{2}$ & 0.769 \\
## Residual Std. Error & 6.598 (df = 20) \\
## F Statistic & 26.549$^{***}$ (df = 3; 20) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
#R2 / adj. R2       .799 / .769

#'
#'
#' Looking at the low p-value of (0.9096), we have insufficient evidence of  interaction between timing
#'
#' g. Then perform F-tests to compare the four model you have created (light as continuous and categori
#'
#' ##### *Model 1 as the reference (light as categorical variabe & without interaction )*
#'
#'
flowers.anova.1 <- anova(flowers.model.cat, flowers.model.cat.int, flowers.model.cont, flowers.model.co
#summary(flowers.anova.1)
stargazer(flowers.anova.1,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} ccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & Res.Df & RSS & Df & Sum of Sq & F & Pr(\textgreater F) \\
## \hline \\[-1.8ex]
## 1 & $17$ & $767.472$ & $$ & $$ & $$ & $$ \\
## 2 & $12$ & $655.925$ & $5$ & $111.547$ & $0.408$ & $0.834$ \\
## 3 & $21$ & $871.236$ & $$-$9$ & $$-$215.311$ & $0.438$ & $0.889$ \\
## 4 & $20$ & $870.660$ & $1$ & $0.576$ & $0.011$ & $0.920$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
# there is no significant difference between the 4 models beccause all the p-values are extremly high
#'
#'
#' * There is no significant difference between the 4 models because all the p-values are extremely hig
#'
#' ##### *Model 2 as the reference (light as categorical variabe & without interaction )*
#'
#'
flowers.anova.2 <- anova( flowers.model.cat.int, flowers.model.cat, flowers.model.cont, flowers.model.c
#summary(flowers.anova.2)
stargazer(flowers.anova.2,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} ccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & Res.Df & RSS & Df & Sum of Sq & F & Pr(\textgreater F) \\
## \hline \\[-1.8ex]
## 1 & $12$ & $655.925$ & $$ & $$ & $$ & $$ \\
## 2 & $17$ & $767.472$ & $$-$5$ & $$-$111.547$ & $0.408$ & $0.834$ \\
## 3 & $21$ & $871.236$ & $$-$4$ & $$-$103.764$ & $0.475$ & $0.754$ \\
## 4 & $20$ & $870.660$ & $1$ & $0.576$ & $0.011$ & $0.920$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
# there is no significant difference between the 4 models beccause all the p-values are extremly high
#'
#'
#' * There is no significant difference between the 4 models because all the p-values are extremely hig
#'
#' ##### *Model 3 as the reference (light as continous variabe & without interaction )*
#'
#'
flowers.anova.3 <- anova( flowers.model.cont, flowers.model.cat, flowers.model.cat.int, flowers.model.c
#summary(flowers.anova.3)
stargazer(flowers.anova.3,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
```

```
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} ccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & Res.Df & RSS & Df & Sum of Sq & F & Pr(\textgreater F) \\
## \hline \\[-1.8ex]
## 1 & $21$ & $871.236$ & $$ & $$ & $$ & $$ \\
## 2 & $17$ & $767.472$ & $4$ & $103.764$ & $0.475$ & $0.754$ \\
## 3 & $12$ & $655.925$ & $5$ & $111.547$ & $0.408$ & $0.834$ \\
## 4 & $20$ & $870.660$ & $$-$8$ & $$-$214.735$ & $0.491$ & $0.841$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
# there is no significant difference between the 4 models beccause all the p-values are extremly high
#'
#'
#' * There is no significant difference between the 4 models because all the p-values are extremely hig
#'
#' ##### *Model 4 as the reference (light as continous variabe & with interaction )*
#'
#'
flowers.anova.4 <- anova(flowers.model.cont.int, flowers.model.cat, flowers.model.cat.int, flowers.model
#summary(flowers.anova.4)
stargazer(flowers.anova.4,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} ccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & Res.Df & RSS & Df & Sum of Sq & F & Pr(\textgreater F) \\
## \hline \\[-1.8ex]
## 1 & $20$ & $870.660$ & $$ & $$ & $$ & $$ \\
## 2 & $17$ & $767.472$ & $3$ & $103.188$ & $0.629$ & $0.610$ \\
## 3 & $12$ & $655.925$ & $5$ & $111.547$ & $0.408$ & $0.834$ \\
## 4 & $21$ & $871.236$ & $$-$9$ & $$-$215.311$ & $0.438$ & $0.889$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
# there is no significant difference between the 4 models beccause all the p-values are extremly high
#'
#'
#' * There is no significant difference between the 4 models because all the p-values are extremely hig
#'
```

```
#' h. Predict the number of flowers grown at each combination of light and timing for each of the four
#'
flowers.df$prediction1 = predict(flowers.model.cat, flowers.df, type = "response")
flowers.df$prediction2 = predict(flowers.model.cat.int, flowers.df, type = "response")
flowers.df$prediction3 = predict(flowers.model.cont, flowers.df, type = "response")
flowers.df$prediction4 = predict(flowers.model.cont.int,flowers.df, type = "response")
stargazer(flowers.df%>%select(FLOWERS,prediction1,prediction2,prediction3, prediction4),
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} cccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & FLOWERS & prediction1 & prediction2 & prediction3 & prediction4 \\
## \hline \\[-1.8ex]
## 1 & $62.300$ & $67.196$ & $69.850$ & $65.235$ & $65.462$ \\
## 2 & $77.400$ & $67.196$ & $69.850$ & $65.235$ & $65.462$ \\
## 3 & $55.300$ & $58.071$ & $54.750$ & $59.164$ & $59.300$ \\
## 4 & $54.200$ & $58.071$ & $54.750$ & $59.164$ & $59.300$ \\
## 5 & $49.600$ & $53.821$ & $55.750$ & $53.094$ & $53.139$ \\
## 6 & $61.900$ & $53.821$ & $55.750$ & $53.094$ & $53.139$ \\
## 7 & $39.400$ & $43.971$ & $42.550$ & $47.023$ & $46.978$ \\
## 8 & $45.700$ & $43.971$ & $42.550$ & $47.023$ & $46.978$ \\
## 9 & $31.300$ & $39.446$ & $38.100$ & $40.952$ & $40.816$ \\
## 10 & $44.900$ & $39.446$ & $38.100$ & $40.952$ & $40.816$ \\
## 11 & $36.800$ & $37.846$ & $39.350$ & $34.882$ & $34.655$ \\
## 12 & $41.900$ & $37.846$ & $39.350$ & $34.882$ & $34.655$ \\
## 13 & $77.800$ & $79.354$ & $76.700$ & $77.393$ & $77.167$ \\
## 14 & $75.600$ & $79.354$ & $76.700$ & $77.393$ & $77.167$ \\
## 15 & $69.100$ & $70.229$ & $73.550$ & $71.323$ & $71.187$ \\
## 16 & $78$ & $70.229$ & $73.550$ & $71.323$ & $71.187$ \\
## 17 & $57$ & $65.979$ & $64.050$ & $65.252$ & $65.207$ \\
## 18 & $71.100$ & $65.979$ & $64.050$ & $65.252$ & $65.207$ \\
## 19 & $62.900$ & $56.129$ & $57.550$ & $59.181$ & $59.227$ \\
## 20 & $52.200$ & $56.129$ & $57.550$ & $59.181$ & $59.227$ \\
## 21 & $60.300$ & $51.604$ & $52.950$ & $53.111$ & $53.247$ \\
## 22 & $45.600$ & $51.604$ & $52.950$ & $53.111$ & $53.247$ \\
## 23 & $52.600$ & $50.004$ & $48.500$ & $47.040$ & $47.267$ \\
## 24 & $44.400$ & $50.004$ & $48.500$ & $47.040$ & $47.267$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

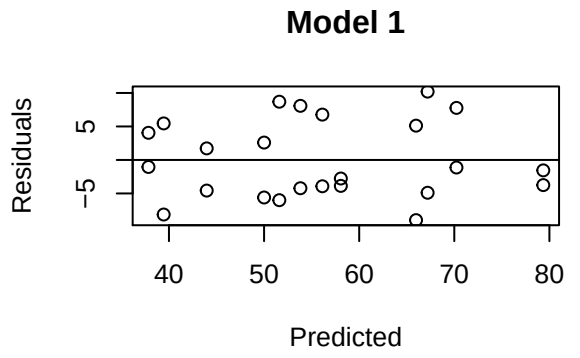
```
#'
#'
#'
```

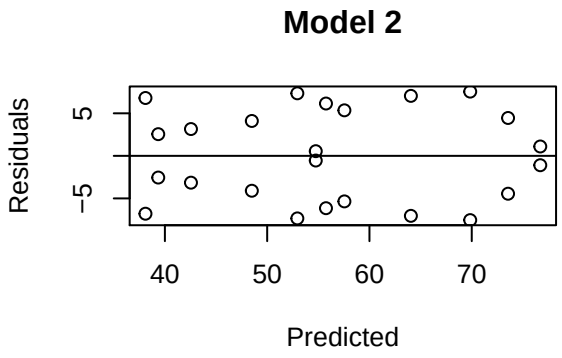
```r
#' i. Compare each prediction to the observed number of flowers and calculate the difference (observed
#'
#RESIDUE
flowers.df$residual1 = flowers.df$FLOWERS-flowers.df$prediction1
flowers.df$residual2 = flowers.df$FLOWERS-flowers.df$prediction2
flowers.df$residual3 = flowers.df$FLOWERS-flowers.df$prediction3
flowers.df$residual4 = flowers.df$FLOWERS-flowers.df$prediction4

# SQUAred residual
flowers.df$residual_mse1 = flowers.df$residual1^2
flowers.df$residual_mse2 = flowers.df$residual2^2
flowers.df$residual_mse3 = flowers.df$residual3^2
flowers.df$residual_mse4 = flowers.df$residual4^2

#MSE
flowers.mse =colSums(flowers.df[,c("residual_mse1","residual_mse2","residual_mse3","residual_mse4")])

stargazer(flowers.mse,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          single.row = T
          )
```

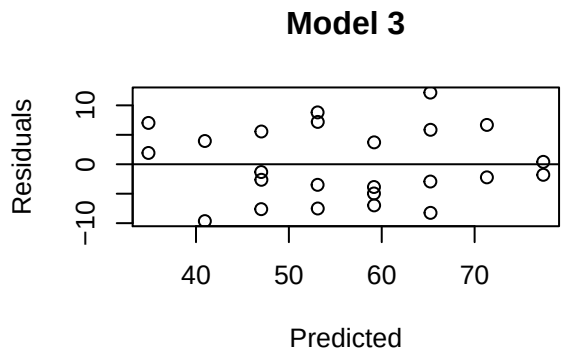
```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} cccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## residual\_mse1 & residual\_mse2 & residual\_mse3 & residual\_mse4 \\
## \hline \\[-1.8ex]
## $767.472$ & $655.925$ & $871.236$ & $870.660$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

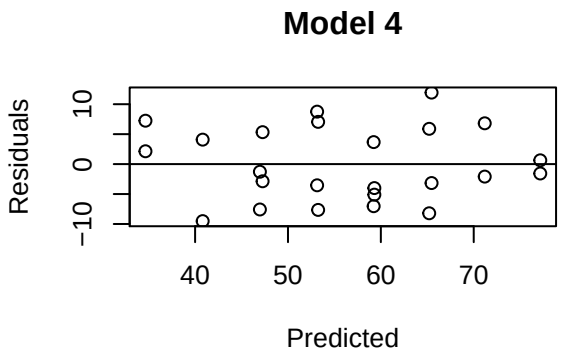
```r
#'
#'
#' j. Now plot the residuals vs. the predicted for each model and see if there are any patterns. If you
#'
#'
par(mfrow = c(2, 2))  # Split the plotting panel into a 2 x 2 grid
# model 1
plot(flowers.df$prediction1,flowers.df$residual1,
     ylab = "Residuals",
       xlab = "Predicted",
       main = "Model 1"
     )
abline(0,0)

# model 2
```

```r
plot(flowers.df$prediction2,flowers.df$residual2,
     ylab = "Residuals",
       xlab = "Predicted",
       main = "Model 2"
     )
abline(0,0)

# model 3
plot(flowers.df$prediction3,flowers.df$residual3,
     ylab = "Residuals",
       xlab = "Predicted",
       main = "Model 3"
     )
abline(0,0)

# model 4
plot(flowers.df$prediction4,flowers.df$residual4,
     ylab = "Residuals",
       xlab = "Predicted",
       main = "Model 4"
     )
abline(0,0)
```



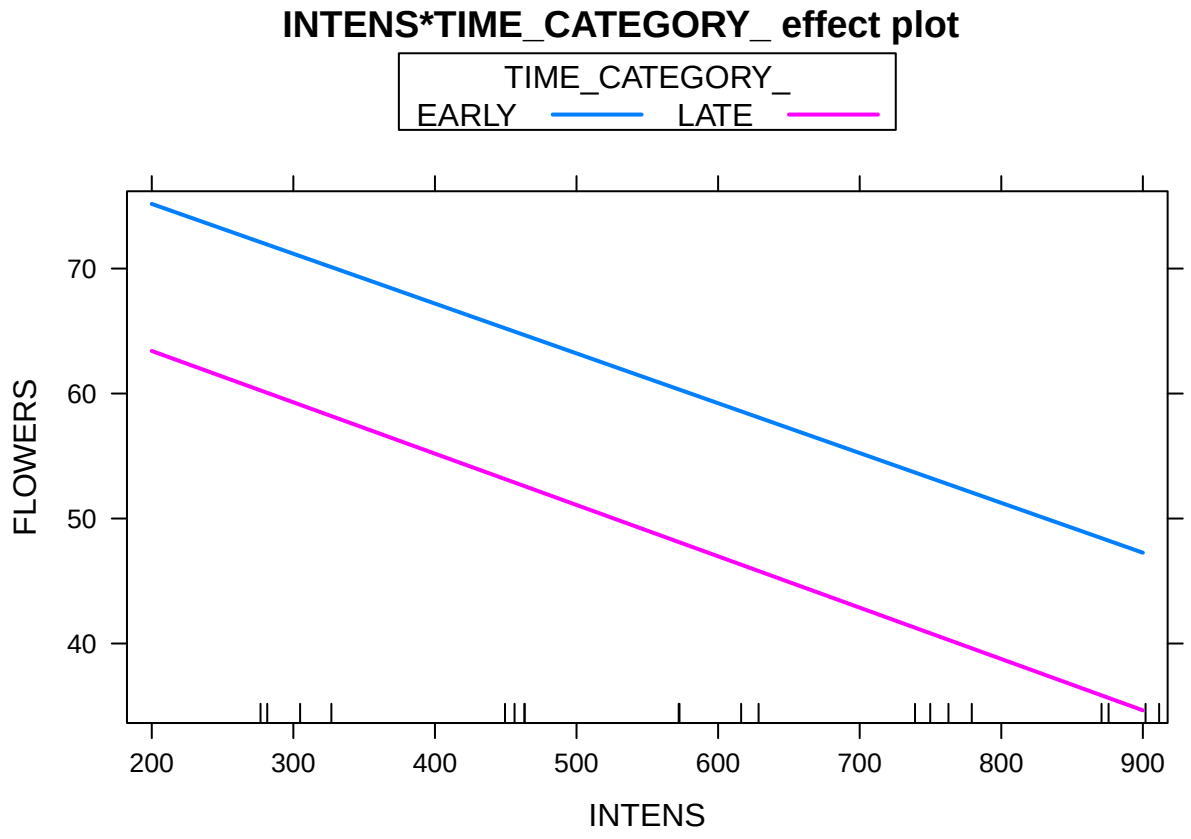




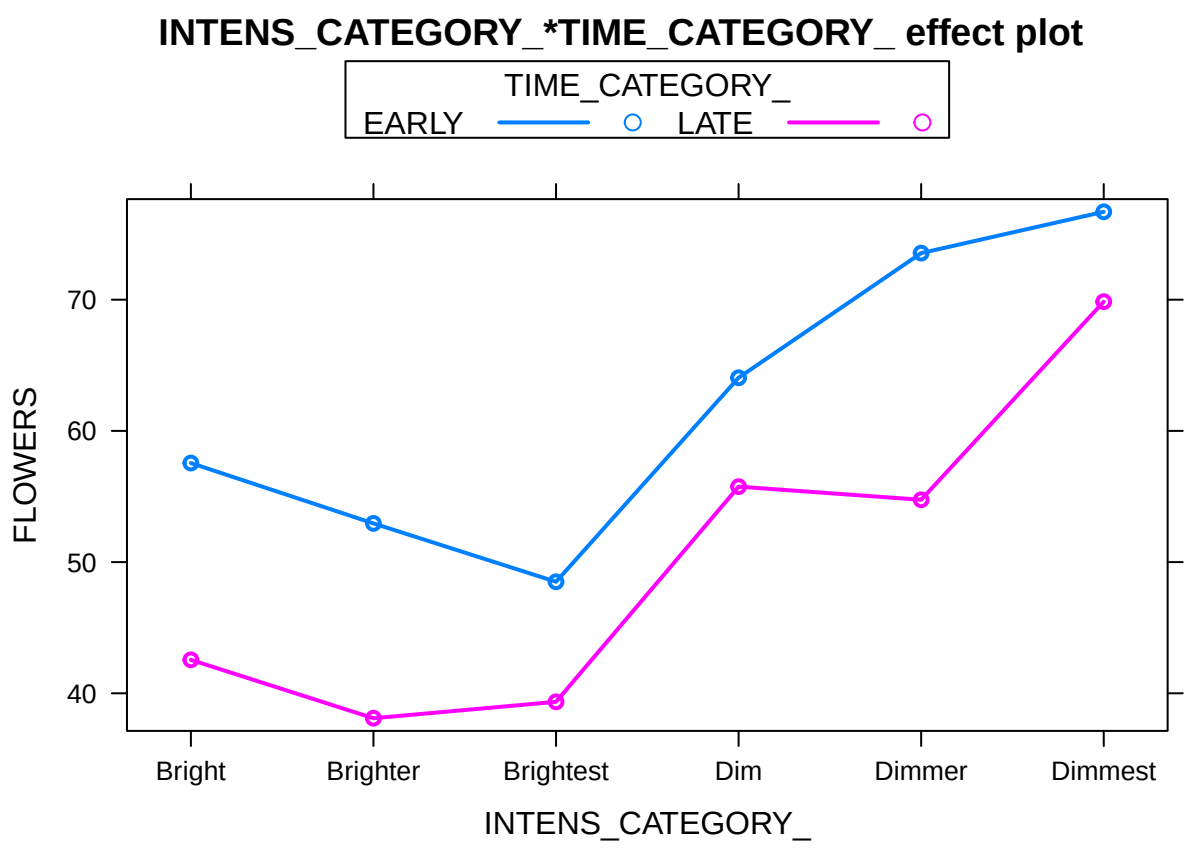



```r
#'
#'
#' Systematic pattern can be visible in model 2 of the residual plot. This means that predictive inform
```

```
#'
#' k. Finally, take the model you think describes the data the best and write a short report for your g
#'
#'
#'
#' ##Report
#'
#' The main goal of this study and analysis is to identify the best combination of light intensity and
#'
#'
#' ### Technical Analysis
#' Before fitting any model we went ahead and explored our data by checking whether there was any signi
#'
#' #### Evaluation whether results differed by replicate
#' Even after establishing that there was no  any significant differences between the 2 groups, we went
#'
#' From the initial step-by-step  analysis, it was well established that we have no significant evidenc
#'
#'
#' Apart from the categorical case, We also established that we did not have significant evidence of  i
#'
#'
#'
par(mfrow = c(2, 1))
plot(effect(term="INTENS:TIME_CATEGORY_",mod=flowers.model.cont.int),multiline=TRUE)
```



```r
plot(effect(term="INTENS_CATEGORY_:TIME_CATEGORY_",mod=flowers.model.cat.int,default.levels=20),multilin
```




```r
#'
#'
#'
#' In fact plotting an interaction plots,  we get perfect parallel lines which do not cross (continuous
#'
#' After checking for interactions, we went ahead and assessed whether there was any significant differ
#'
#' #### Model Selection
#' In order to identify the best model which best describe the relationship between the dependent and i
#'
#' * Linearity: The relationship between independent and the dependent is linear.
#' * Homoscedasticity: The variance of residual is the same for any value of independent.
#' * Independence: Observations are independent of each other.
#' * Normality: Residual must be normally distributed.
#'
#' Most of the models exhibited all these properties. The next thing is to asses the best fit model wit
#'
#' Since the interaction terms in the 2 models (categorical with interaction and continuous with intera
#'
#'
stargazer(flowers.model.cat.int,flowers.model.cont.int,
          header=F,
          type = "latex",
          summary = F,
```

```
          no.space = T,
          ci = TRUE,
         keep = c("\\bprecip\\b"),
          title = "Models With Interraction Terms!",
          column.labels = c( "Categorical Model", "Continous Model"),
          notes = "Coefficients have been removed!",
            dep.var.caption  = "-" ,  # Bold
          single.row = T

          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{Models With Interraction Terms!}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lcc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{2}{c}{-} \\
## \cline{2-3}
## \\[-1.8ex] & \multicolumn{2}{c}{FLOWERS} \\
##  & Categorical Model & Continous Model \\
## \\[-1.8ex] & (1) & (2)\\
## \hline \\[-1.8ex]
## \hline \\[-1.8ex]
## Observations & 24 & 24 \\
## R$^{2}$ & 0.849 & 0.799 \\
## Adjusted R$^{2}$ & 0.710 & 0.769 \\
## Residual Std. Error & 7.393 (df = 12) & 6.598 (df = 20) \\
## F Statistic & 6.124$^{***}$ (df = 11; 12) & 26.549$^{***}$ (df = 3; 20) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{2}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
##  & \multicolumn{2}{r}{Coefficients have been removed!} \\
## \end{tabular}
## \end{table}
```

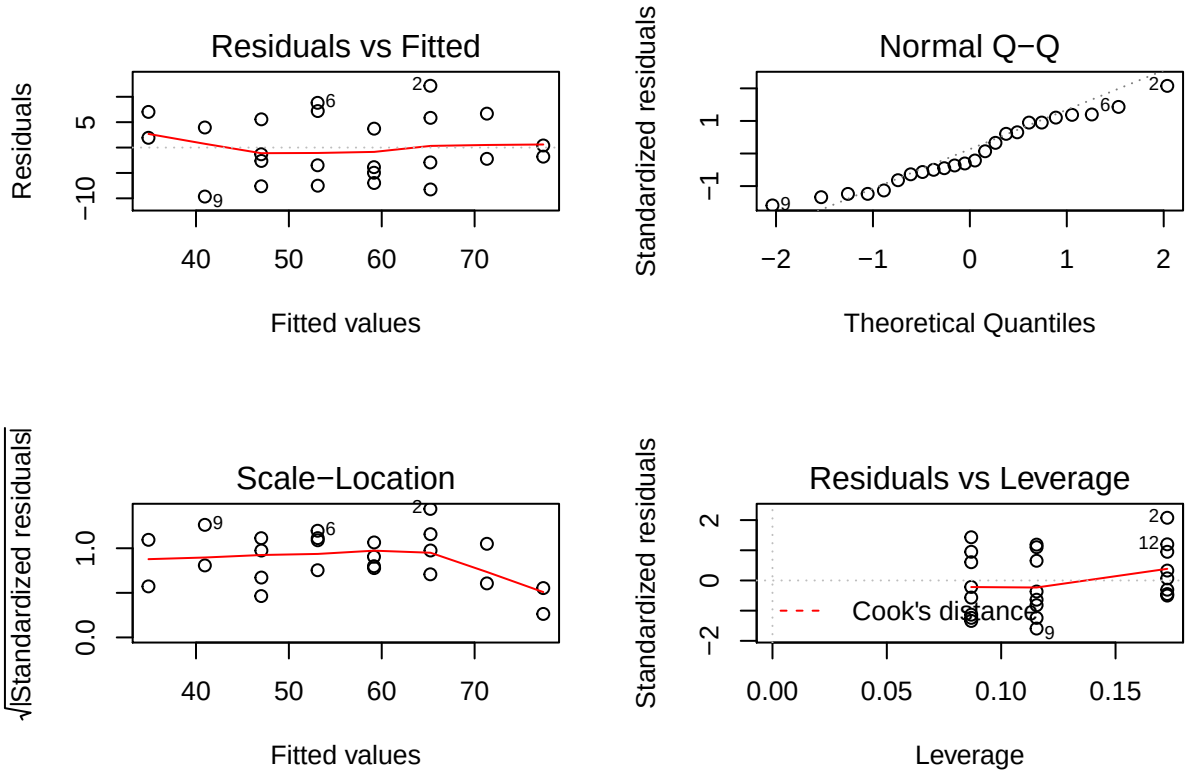
```
#'
#'
#' Another reason for dropping them was the fact that they had  lower explanatory power (R-squared and
#'
#' #### Model Without interraction
#'
#' Well now we are down to 2 models (Models Without interaction). For starter Both models have  high ex
#'
#'
stargazer(flowers.model.cat,flowers.model.cont,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          ci = TRUE,
          title = "Models Without Interraction",
          column.labels = c( "Categorical Model", "Continous Model"),
```

```
          notes = "Notes",
          single.row = T

          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{Models Without Interraction}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lcc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{2}{c}{\textit{Dependent variable:}} \\
## \cline{2-3}
## \\[-1.8ex] & \multicolumn{2}{c}{FLOWERS} \\
##  & Categorical Model & Continous Model \\
## \\[-1.8ex] & (1) & (2)\\
## \hline \\[-1.8ex]
##  INTENS\_CATEGORY\_Brighter & $-$4.525 ($-$13.837, 4.787) &  \\
##   INTENS\_CATEGORY\_Brightest & $-$6.125 ($-$15.437, 3.187) &  \\
##   INTENS\_CATEGORY\_Dim & 9.850$^{*}$ (0.538, 19.162) &  \\
##   INTENS\_CATEGORY\_Dimmer & 14.100$^{***}$ (4.788, 23.412) &  \\
##   INTENS\_CATEGORY\_Dimmest & 23.225$^{***}$ (13.913, 32.537) &  \\
##   INTENS &  & $-$0.040$^{***}$ ($-$0.051, $-$0.030) \\
##   TIME\_CATEGORY\_LATE & $-$12.158$^{***}$ ($-$17.535, $-$6.782) & $-$12.158$^{***}$ ($-$17.312, $-$7
##   Constant & 56.129$^{***}$ (49.017, 63.241) & 83.464$^{***}$ (77.048, 89.881) \\
##  \hline \\[-1.8ex]
## Observations & 24 & 24 \\
## R$^{2}$ & 0.823 & 0.799 \\
## Adjusted R$^{2}$ & 0.761 & 0.780 \\
## Residual Std. Error & 6.719 (df = 17) & 6.441 (df = 21) \\
## F Statistic & 13.181$^{***}$ (df = 6; 17) & 41.780$^{***}$ (df = 2; 21) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{2}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
##  & \multicolumn{2}{r}{Notes} \\
## \end{tabular}
## \end{table}
```

```
#'
#'
#' In the continuous model, 80% of the variability in the number of flowers is explained by covariates(
#'
#'
#' The best model that best describe the data is:  **Continous model without interaction** because:
#'
#' * This model accomplishes the desired level of explanatory power (80%) with as few predictor variabl
#' * This model has a higher adjusted R-Squared 0.78 compared to 0.76 of the latter. In the categorical
#' * This model has lower residual standard error 6.441 compared to 6.719 in the categorical model. On
#' * This model has  highly significant predictors. The categorical model has 2 insignificant predictor
#'
#'
#' ####Model diagnostics
#'
```

```
#'
par(mfrow = c(2, 2))
plot(flowers.model.cont)
```




```
#'
#'
#'
#'
#' **Residuals vs. fitted values**
#'
#' * We observe an acceptable Scatter-plot in the sense that the fitted values doesn't show any noticea
#' * From the plot, the residuals bounce randomly around the 0 line and roughly forms a horizontal band
#'
#' **Normal Q-Q Plot**
#'
#' * This plot shows that the residuals are normally distributed since the data points are  well arrang
#'
#' **Scale-Location Plot**
#'
#' * This plot confirms the assumption of equal variance (homoscedasticity). The residuals appear rando
#'
#' **Residual vs Levarage Plot**
#'
#' *  We do not have datapoints outside the Cook's distance lines. In fact the red dashed lines is not
#'
#' ####Summary
```

```
#' In summary the selected model tells us that:
#'
#' * If we held  other predictors constant, an increase of light intensity by 1 unit  results to a decr
#' * If we held  other predictors constant, "Late" timing for light exposure  results to a decrease in
#'
#' This means that light intensity is inversely proportional to the number of flowers. At the same time
#'
#' ### Grandmothers Report
#'
#' The goal of the analysis was to identify the best combination of light intensity and timing of onset
#'
#'
#'
#'
#'
#' \onecolumn
#'
#' # Source Ccde
#'
#'

#'
#'
```