Multi-Linear-Regression.R

Ng’ang’a Gathii

2023-09-09

data <-read.csv("plant\_growth\_data.csv",header = TRUE)  
head(data)

## plant\_id sunlight\_exposure soil\_ph precipitation  
## 1 1 7.370958 7.360290 5.998142  
## 2 2 5.435302 7.313425 10.667554  
## 3 3 6.363128 6.699037 12.342650  
## 4 4 6.632863 7.554545 14.119078  
## 5 5 6.404268 6.799968 7.246277  
## 6 6 5.893875 7.031654 7.698289  
## average\_temperature growth\_rate  
## 1 82.10825 6.740358  
## 2 77.06615 5.989911  
## 3 79.12838 6.444611  
## 4 80.63366 6.716804  
## 5 79.06633 6.079483  
## 6 77.62374 5.895819

# Fit the multiple linear regression model  
model <- lm(growth\_rate ~ sunlight\_exposure + soil\_ph + precipitation + average\_temperature, data = data)  
  
# Print the summary of the model  
summary(model)

##   
## Call:  
## lm(formula = growth\_rate ~ sunlight\_exposure + soil\_ph + precipitation +   
## average\_temperature, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.38418 -0.12227 -0.00270 0.09157 0.68755   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.596262 1.484467 -0.402 0.68883   
## sunlight\_exposure 0.088836 0.073641 1.206 0.23068   
## soil\_ph 0.001137 0.074966 0.015 0.98793   
## precipitation 0.059012 0.010105 5.840 7.27e-08 \*\*\*  
## average\_temperature 0.072785 0.023207 3.136 0.00228 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2016 on 95 degrees of freedom  
## Multiple R-squared: 0.7417, Adjusted R-squared: 0.7308   
## F-statistic: 68.2 on 4 and 95 DF, p-value: < 2.2e-16

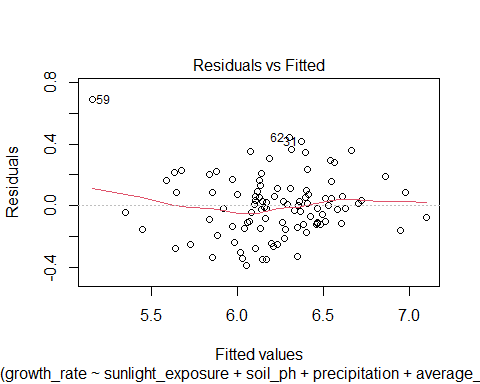
# Assess assumptions of the linear model  
# 1. Check for multicollinearity using variance inflation factor (VIF)  
library(car)  
vif(model)

## sunlight\_exposure soil\_ph precipitation   
## 14.321344 1.006971 1.028793   
## average\_temperature   
## 14.340471

# 2. Check for normality of residuals  
residuals <- residuals(model)  
  
shapiro.test(residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: residuals  
## W = 0.97834, p-value = 0.09868

# 3. Check for homoscedasticity (constant variance)  
plot(model, which = 1)

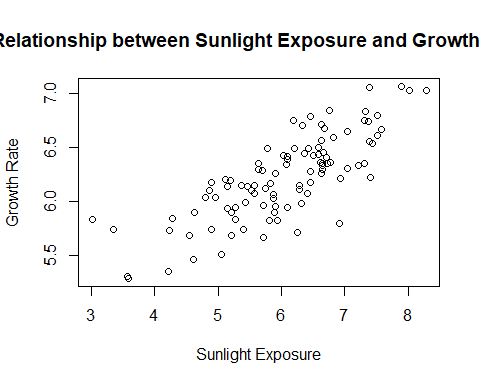


# Interpret and present the coefficients and their statistical significance  
library(knitr)  
# Create a data frame with the coefficients  
coefficients <- coef(summary(model))  
coefficients\_df <- data.frame(  
 Variable = rownames(coefficients),  
 Estimate = coefficients[, "Estimate"],  
 Std\_Error = coefficients[, "Std. Error"],  
 t\_value = coefficients[, "t value"],  
 p\_value = coefficients[, "Pr(>|t|)"]  
)  
  
# Print the coefficients table  
kable(coefficients\_df, caption = "Coefficients and their Statistical Significance")

Coefficients and their Statistical Significance

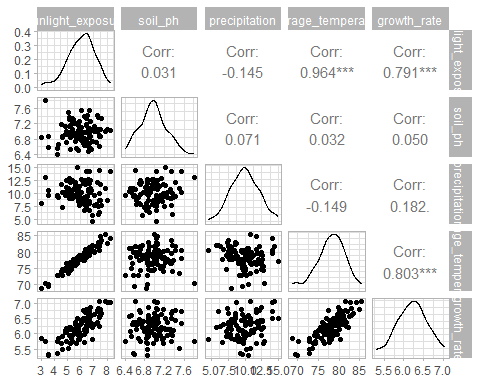
|  | Variable | Estimate | Std\_Error | t\_value | p\_value |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | (Intercept) | -0.5962621 | 1.4844669 | -0.4016675 | 0.6888309 |
| sunlight\_exposure | sunlight\_exposure | 0.0888359 | 0.0736411 | 1.2063354 | 0.2306838 |
| soil\_ph | soil\_ph | 0.0011368 | 0.0749659 | 0.0151636 | 0.9879335 |
| precipitation | precipitation | 0.0590118 | 0.0101050 | 5.8398685 | 0.0000001 |
| average\_temperature | average\_temperature | 0.0727850 | 0.0232071 | 3.1363232 | 0.0022776 |

# Visualize the relationships between variables using graphs or charts  
# Scatter plot of growth rate vs. sunlight exposure  
plot(data$sunlight\_exposure, data$growth\_rate, xlab = "Sunlight Exposure", ylab = "Growth Rate", main = "Relationship between Sunlight Exposure and Growth Rate")

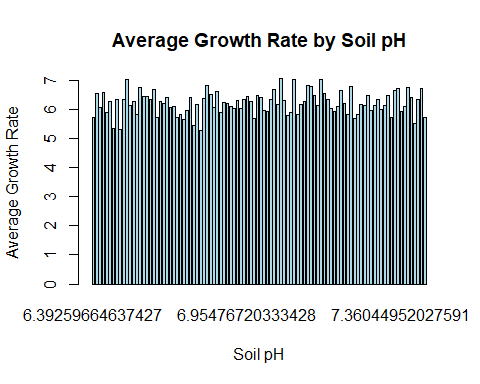


# Scatter plot matrix for all variables  
library(GGally)  
ggpairs(data, columns = 2:6)

## plot: [1,1] [=>--------------------------------------] 4% est: 0s  
## plot: [1,2] [==>-------------------------------------] 8% est: 2s  
## plot: [1,3] [====>-----------------------------------] 12% est: 2s  
## plot: [1,4] [=====>----------------------------------] 16% est: 2s  
## plot: [1,5] [=======>--------------------------------] 20% est: 2s  
## plot: [2,1] [=========>------------------------------] 24% est: 2s  
## plot: [2,2] [==========>-----------------------------] 28% est: 2s  
## plot: [2,3] [============>---------------------------] 32% est: 2s  
## plot: [2,4] [=============>--------------------------] 36% est: 2s  
## plot: [2,5] [===============>------------------------] 40% est: 2s  
## plot: [3,1] [=================>----------------------] 44% est: 2s  
## plot: [3,2] [==================>---------------------] 48% est: 2s  
## plot: [3,3] [====================>-------------------] 52% est: 2s  
## plot: [3,4] [=====================>------------------] 56% est: 2s  
## plot: [3,5] [=======================>----------------] 60% est: 2s  
## plot: [4,1] [=========================>--------------] 64% est: 1s  
## plot: [4,2] [==========================>-------------] 68% est: 1s  
## plot: [4,3] [============================>-----------] 72% est: 1s  
## plot: [4,4] [=============================>----------] 76% est: 1s  
## plot: [4,5] [===============================>--------] 80% est: 1s  
## plot: [5,1] [=================================>------] 84% est: 1s  
## plot: [5,2] [==================================>-----] 88% est: 0s  
## plot: [5,3] [====================================>---] 92% est: 0s  
## plot: [5,4] [=====================================>--] 96% est: 0s  
## plot: [5,5] [========================================]100% est: 0s



# Bar plot: Average Growth Rate by Soil pH  
barplot(tapply(data$growth\_rate, data$soil\_ph, mean),  
 xlab = "Soil pH", ylab = "Average Growth Rate",  
 main = "Average Growth Rate by Soil pH", col = "lightblue")



# Heatmap: Correlation between Variables  
correlation\_matrix <- cor(data[, c("growth\_rate", "sunlight\_exposure", "soil\_ph", "precipitation", "average\_temperature")])  
heatmap(correlation\_matrix, col = colorRampPalette(c("#4575B4", "#FFFFFF", "#D45050"))(100),  
 main = "Correlation between Variables", xlab = "", ylab = "")

