MH3510 Assignment 1

Lye En Lih (U2121387B)

04 October 2024

Goal of Assignment: To explore the relationship between the amount of β-erythroidine in an aqueous solution and the colorimeter reading of the turbidity

Initialising Data Structures

X: Concentration (mg/mL)
Y: Colorimeter Reading

Storing the data into two seperate vectors x and y for concentration and colorimeter reading respectively

```
x <- c(40, 50, 60, 70, 80, 90, 40, 60, 80, 50) #Concentration Values (in mg/mL)
y <- c(69, 175, 272, 335, 490, 415, 72, 265, 492, 180) # Colorimeter Reading Values
```

a. Fit a simple regression to the data

Simple Linear Regression Model (SLR)

Here, we will use the lm() function to derive the model.

```
slr <- lm(y ~ x)

# Display the coefficients of this model
slr
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##    -252.297        8.529
```

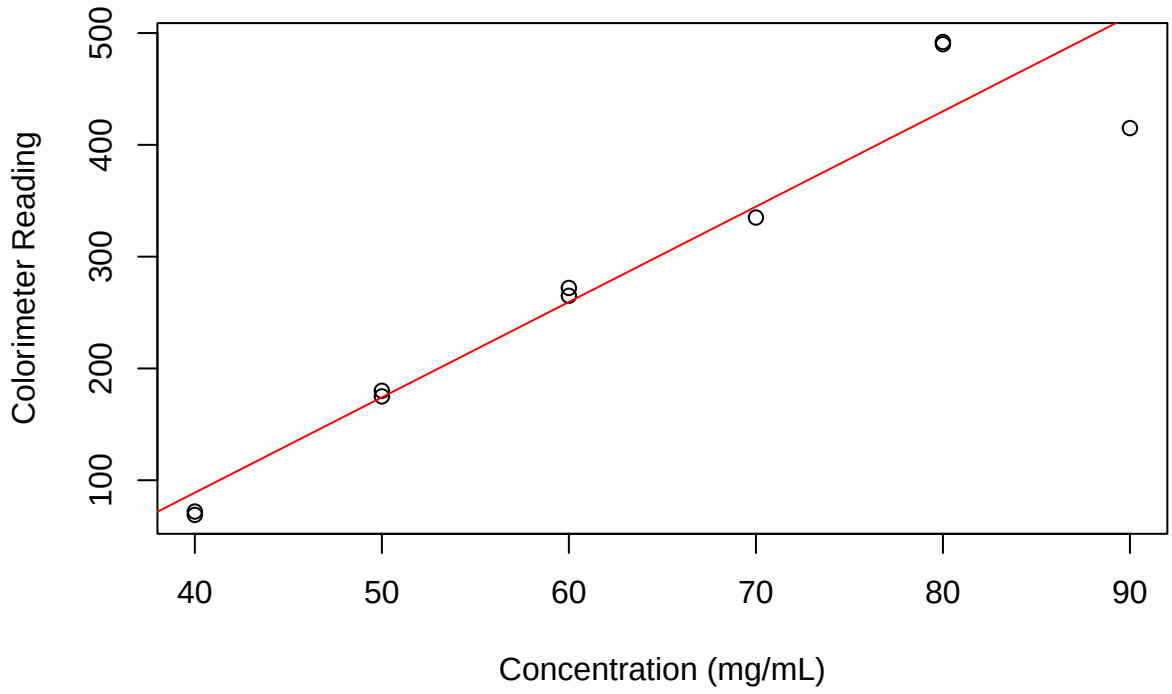
From the plot above, we can derive the values for β_0 (intercept) = -252.297 and β_1 (slope) = 8.529. Hence, this gives us the best fit regression line of the Colorimeter Reading against Concentration:

$$\hat{y} = -252.297 + 8.529x$$

Plot Colorimeter Reading Against Concentration with Best Fit Line

```r
# Plot ScatterPlot
plot(x, y, xlab = "Concentration (mg/mL)", ylab = "Colorimeter Reading", pch = 1,
     col = "black")

# Plot Best Fit Regression Line
abline(slr, col = "red")
```



As shown in the scatter plot, we can observe that the linear model of `Colorimeter Reading (Y)` against `Concentration (X)` fits fairly well to the data as there is a **positive linear relationship** and the points lie closely to the Best Fit Regression Line.

However, we can also observe that there are large deviations of `Actual Colorimeter Reading` and `Predicted Colorimeter Reading` at concentrations of `80 mg/mL` and `90 mg/mL`. The model's adequacy will be further discussed in part (b).

b. Statistical Analysis of the Model

For this assignment, we will use the level of significance $\alpha = 0.05$.

R^2 Statistic

```
# Getting R_squared value from model summary
r_squared <- summary(slr)$r.squared

# Displaying R^2 valye
r_squared
```

```
## [1] 0.9156376
```

Examining the R^2 value based on the output above

From the SLR Model result, we obtained a R^2 value of 0.916 (rounded to 3 Decimal Places).
Hence, there is 91.6% variation of `Colorimeter Reading` which could be explained using the `Concentration`.
This indicates a **strong linear relationship** between the `Concentration` and `Colorimeter Reading`.

F-Test

```
# Getting F-test from model summary
f_stat <- summary(slr)$fstatistic[1]
p_value <- pf(f_stat, df1 = summary(slr)$fstatistic[2], df2 =summary(slr)$fstatistic[3],
              lower.tail = FALSE)

# Displaying F-Test Value
f_stat
```

```
##  value
## 86.829
```

```
# Displaying p-value
p_value
```

```
##        value
## 1.434372e-05
```

Examining the F-statistic and p-value based on the output above

We use F test to test the significance of the Simple Linear Regression using the following hypothesis:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

From the SLR Model result, we obtained a F-statistic value of 86.829 with degree of freedom (1,8) with a corresponding p-value of 1.434e-05. Since there is a high F-statistic value and the p-value being much smaller than 0.05 (significance level), we can deduce that the SLR Model explains a significant amount of the variability in `Colorimeter Reading` based on `Concentration`.

Therefore, we conclude that the SLR model is **statistically significant**, indicating that there is a **significant linear relationship** between `Colorimeter Reading` and `Concentration`.

Comments on the Model Adequacy

Observations:

The Simple Linear Regression Model yielded the following results:
R^2 **value**: 0.916
F-statistic: 82.829
P-value: 1.434e-05

1. **Model Fit**: Using the scatter plot in (a) and the R^2 value = 0.916, we can conclude the SLR Model provides a good fit for `Colorimeter Reading` against `Concentration` with a **strong linear relationship**. However, the deviations and outliers in the scatter plot suggests that a higher order polynomial such as x^2 should be fitted at concentrations of `80 mg/mL` and `90 mg/mL`.
2. **F-statistic Significance**: The high value of F-statistic = 82.829 in conjunction with the p-value being significantly smaller than 0.05 shows that the model is statistically significant, hence there is a **significant linear relationship** between `Colorimeter Reading` and `Concentration`.
3. **Overall Model Adequacy**: Since the R^2 value is high and the F-test is significant, we can conclude that **the model is both a good fit and statistically significant** for the relationship between `Colorimeter Reading` and `Concentration`.
4. **Further Improvements**: Even though the R^2 value and F-test are good ways to check for overall significance of the model, it is important to check for other diagnostics such as residual plots to ensure there are no violations of regression assumptions. Hence, to further improve the check for the model's adequacy, a residual plot and Q-Q plots should be used to check for patterns in residuals and normality of residuals.