Kumar - SDS 291 Homework 1

HOMEWORK 1: FUNDAMENTALS OF MODELING AND INFERENCE

SET UP

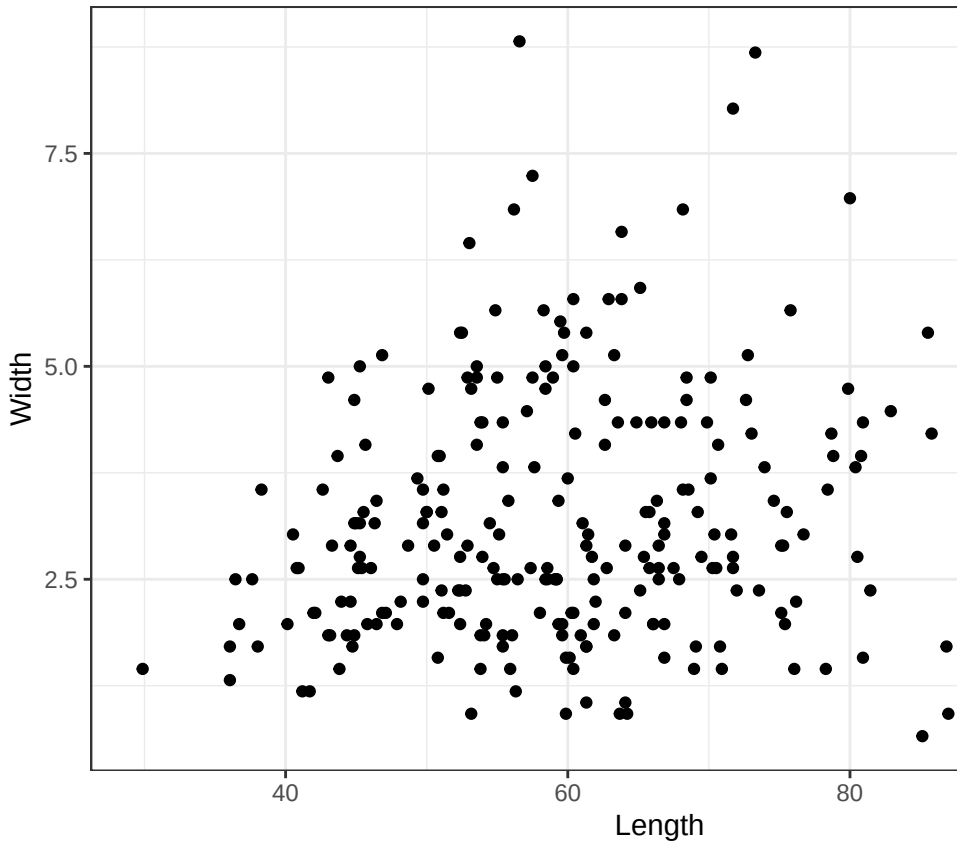
Load libraries and data

```
library(Stat2Data)
library(ggplot2)

data("LeafWidth")
head(LeafWidth)
```

```
##       Width   Length   LWRatio     Area Year
## 1 0.6578947 85.13158 129.40000 70.32548 1974
## 2 0.9210526 53.15789  57.71429 48.96122 1918
## 3 0.9210526 59.86842  65.00000 57.72161 1993
## 4 0.9210526 63.68421  69.14286 37.36150 1946
## 5 0.9210526 64.21053  69.71429 62.55194 1956
## 6 0.9210526 86.97368  94.42857 67.52078 1964
```

EXERCISE 1: Leaf Width

```
ggplot(data = LeafWidth, mapping = aes(x = Length,
                                       y = Width)) +
  geom_point() +
  theme_bw()
```



EXERCISE 1A: Visualize the data

This is a scatter plot showing the relationship between the length and width of leaves from the dataset `LeafWidth`. There isn't a clear linear relationship between these two variables. It seems that for leaves with smaller lengths, there are smaller widths, but as length increases width doesn't necessarily increase proportionally.

EXERCISE 1B: Fit a regression model Fit a linear model

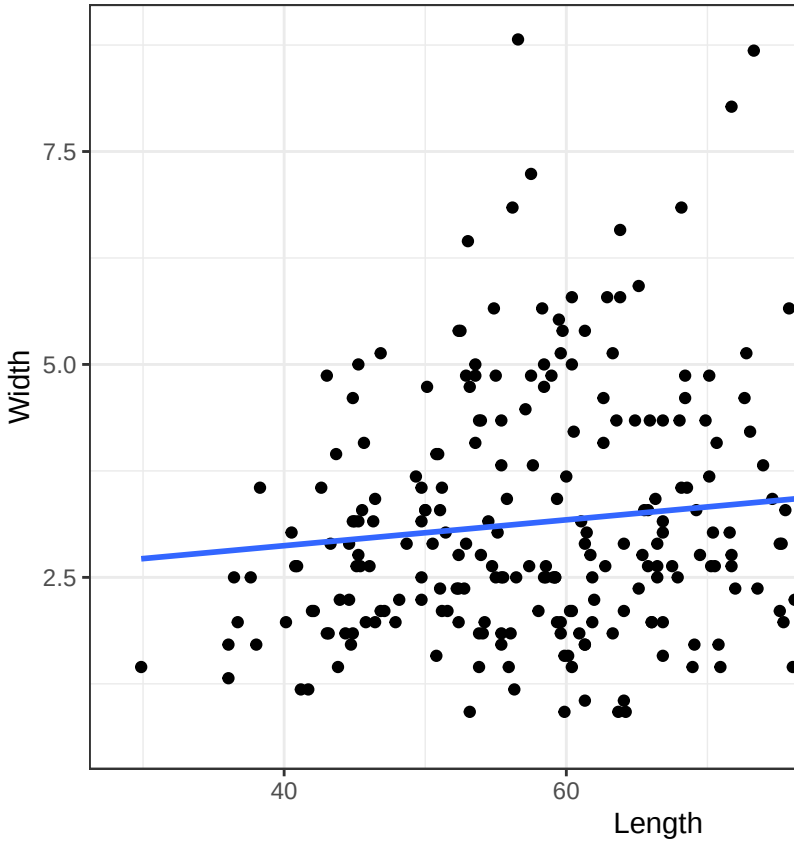
```
## fit a model
leaf_model = lm(Width ~ Length, data = LeafWidth)

## report the regression table
summary(leaf_model)
```

```
##
## Call:
## lm(formula = Width ~ Length, data = LeafWidth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8990 -1.0488 -0.3194  0.8751  5.6922
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.264938   0.446616   5.071  7.7e-07 ***
## Length      0.015176   0.007304   2.078   0.0388 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.457 on 250 degrees of freedom
## Multiple R-squared:  0.01697,    Adjusted R-squared:  0.01304
## F-statistic: 4.316 on 1 and 250 DF,  p-value: 0.03877
```

Interpret the model: This summary table tells us the intercept (2.265) and slope (0.015) values for this regression model. Practically, this means, based on the dataset, that the model finds for each increase in average leaf length by one mm, average leaf width increases by 0.015 mm. It also means that for leaves with average length 0 mm, they have an average width of 2.265 mm.

```
ggplot(data = LeafWidth, mapping = aes(x = Length, y = Width)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE, formula = y ~ x) +
  theme_bw()
```



EXERCISE 1C: Visualize the regression model

EXERCISE 1D: Residual error We are given that the average leaf width recorded in 1985 is 5.0mm, and its length was 45.26316mm.

Given values

```
actual_length = 45.26316
actual_width = 5.00
model_slope = 0.015176
model_intercept = 2.264938
```

Predicted leaf width based on length of 45.26316mm.

```
predicted_width = model_slope*actual_length + model_intercept
predicted_width
```

```
## [1] 2.951852
```

Calculate residual

```
residual_1985 = actual_width - predicted_width
residual_1985
```

```
## [1] 2.048148
```

EXERCISE 2: Coffee Quality

EXERCISE 2A: Hypotheses This test is a randomization distribution, where Eleki has placed the `altitude` as the explanatory variable and `aroma` as the outcome variable. Eleki's null hypothesis is that there is no relationship between altitude and aroma rating. Their alternative hypothesis is that there is some relationship between altitude and aroma rating.

In symbols, this looks like:

$H_0 : \beta_0 = 0$ $H_a : \beta_0 \neq 0$

EXERCISE 2B: The null distribution This distribution is a histogram of slope values – it's the visualization of a random distribution. The random distribution is a dataset of all possible permutations of a given dataset, or all possible iterations of X and Y values given that the Y values are randomly matched with X values. Then R calculates the slope (correlation value) of all the random distributions and displays them in the Null Distribution: a histogram of all slope values.

Practically, this Null Distribution shows the variety of slope values available for this random distribution and how common they are.

EXERCISE 2C: The red shading The red shaded area of Figure 1 indicates original sample's slope. The thick red line represents the original sample's slope and the shaded areas on either side, all slope values equal to or greater than (magnitude) the slope value from the original sample.

EXERCISE 2D: Interpreting the p-value The p-value of this hypothesis test was 0.518. A p-value, in this case, is the probability that, given the null hypothesis is true (there is no relationship between aroma and altitude), a slope value equal to or greater than the one from your original sample could have been generated randomly. Therefore, given there's no relationship between altitude and aroma, there was a 51.8% chance this dataset could have occurred randomly.

When the p-value is very low, say 0.001, then it's highly unlikely the null hypothesis is true because that means that, given aroma and altitude have no relationship, there is a %0.1 chance this sample occurred randomly, which means it's likely altitude and aroma have a relationship.

EXERCISE 2E: Drawing conclusions There is not sufficient evidence to reject the null hypothesis, which means we cannot definitively state if there is a relationship between `altitude` and `aroma`.