Testing for Correlation

The following table provides the weight W (in pounds) and length l (in inches) for 15 alligators in central Florida. Because length is easier observe than weight, we want to construct a model relating weight to length, which could be used to predict the weights of alligators of specified lengths. In particular, we want to fit the model

$$\ln W = \ln \alpha_0 + \alpha_1 \ln l + \epsilon = \beta_0 + \beta_1 x + \epsilon.$$

Let's test if there is a correlation between $X = \ln l$ and $Y = \ln W$.

Alligator	$x = \ln l$	$y = \ln W$
1	3.87	4.87
2	3.61	3.93
3	4.33	6.46
4	3.43	3.33
5	3.81	4.38
6	3.83	4.70
7	3.46	3.50
8	3.76	4.50
9	3.50	3.58
10	3.58	3.64
11	4.19	5.90
12	3.78	4.43
13	3.71	4.38
14	3.73	4.42
15	3.78	4.25

Calculating the Test Statistic

We first create vectors to store the observed values of X and Y.

```
x = c(3.87, 3.61, 4.33, 3.43, 3.81, 3.83, 3.46, 3.76, 3.5, 3.58, 4.19,
      3.78, 3.71, 3.73, 3.78)
y = c(4.87, 3.93, 6.46, 3.33, 4.38, 4.70, 3.5, 4.5, 3.58, 3.64, 5.9, 4.43, 4.38, 4.42, 4.25)
```

We can calculate S_{xx}, S_{xy}, and S_{yy}, which will be used to calculate sample correlation coefficient r, using the following code.

```
# Number of observations.
n <- length(x)

# Sums.
sumx <- sum(x = x)
sumy <- sum(x = y)
sumxy <- sum(x = x*y)
sumxx <- sum(x = x^2)
sumyy <- sum(x = y^2)

# Calculate Sxy and Sxx
```

```
Sxy <- sumxy - 1/n*sumx*sumy
Sxx <- sumxx - 1/n*sumx^2
Syy <- sumyy - 1/n*sumy^2
```

This yields

$$\begin{array}{lll} \sum x_i = 56.37 & \sum x_i^2 = 212.6933 & S_{xx} = 0.85484 \\ \sum y_i = 66.27 & \sum y_i^2 = 303.0409 & S_{yy} = 10.26004 \\ & \sum x_i y_i = 251.9757 & S_{xy} = 2.93304 \end{array}$$

We can calculate the sample correlation coefficient using the formula

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

and the following code.

```
r <- Sxy/sqrt(Sxx*Syy)
```

This gives $r = 0.9903781$.

Performing the test

The sample correlation coefficient r has value close to 1, which suggests a strong positive correlation between the logarithms of length and width. Assuming that (X, Y) has a bivariate normal distribution, we can test whether X and Y are independent using the hypotheses $H_0 : \rho = 0$ and $H_a : \rho \neq 0$. We can calculate the value of the test statistic using the formula

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

and the following code.

```
tval <- r*sqrt(n-2)/sqrt(1-r^2)
```

This gives $t = 25.8032943$. Testing at $\alpha = 0.05$ level of significance, we have rejection region given by $|t| > t_{0.025} = 2.1603687$. Since the observed value of t lies in the rejection region, we can reject H_0 and conclude that we have evidence that X and Y are not independent. The following code calculates the attained significance level.

```
2*pt(q = tval, df = n - 2, lower.tail = FALSE)
```

```
## [1] 1.494611e-12
```

Coefficient of Determination

The coefficient of determination r^2 is given by following code.

```
r^2
```

```
## [1] 0.9808488
```