KNOWLEDGE ENGINEERING LAB (CSE 4.1.7)

3. Introduction to regression using R.

Air Velocity (cm/sec)	20,60,100,140,180,220,260,300,340,380
Evaporation Coefficient(mm2/sec)	0.18, 0.37, 0.35, 0.78, 0.56, 0.75, 1.18, 1.36, 1.17, 1.65

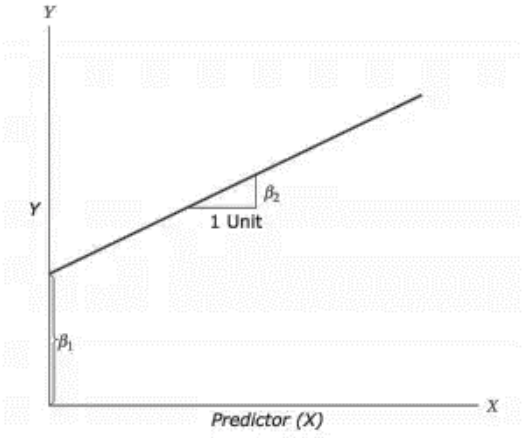
Use R to perform linear regression on the given the data. Analyze the significance of residual standard-error value, R-squared value, F-statistic. Find the correlation coefficient for this data and analyze the significance of the correlation value.

Introduction to Linear Regression

Linear regression is one of the most commonly used predictive modelling techniques. The aim of linear regression is to find a mathematical equation for a continuous response variable Y as a function of one or more X variable(s). So that you can use this regression model to predict the Y when only the X is known. It is expressed in the equation 1.

$$Y = \beta 1 + \beta 2X + \epsilon \qquad (1)$$

Where $\beta 1$ is intercept, and $\beta 2$ is slope, and ϵ is the error term.



Problem Specification

In the given problem 'Air velocity', and 'Evaporation Coefficient' are the variables with 10 observations.

The goal here is to establish a mathematical equation for 'Evaporation Coefficient' as a function of 'Air velocity', so you can use it to predict 'Evaporation Coefficient' when only the 'Air velocity' of the car is known. So, it is desirable to build a linear regression model with the response variable as 'Evaporation Coefficient' and the predictor as 'Air velocity'. Before we begin building the regression model, it is a good practice to analyse and understand the variables.

```
> airvelocity<-c(20,60,100,140,180,220,260,300,340,380)
> evaporationcoefficient<-c(0.18, 0.37, 0.35, 0.78, 0.56, 0.75, 1.18, 1.36, 1
.17, 1.65)
> airvelocity
 [1]  20  60 100 140 180 220 260 300 340 380
> evaporationcoefficient
 [1] 0.18 0.37 0.35 0.78 0.56 0.75 1.18 1.36 1.17 1.65
```

Graphical analysis

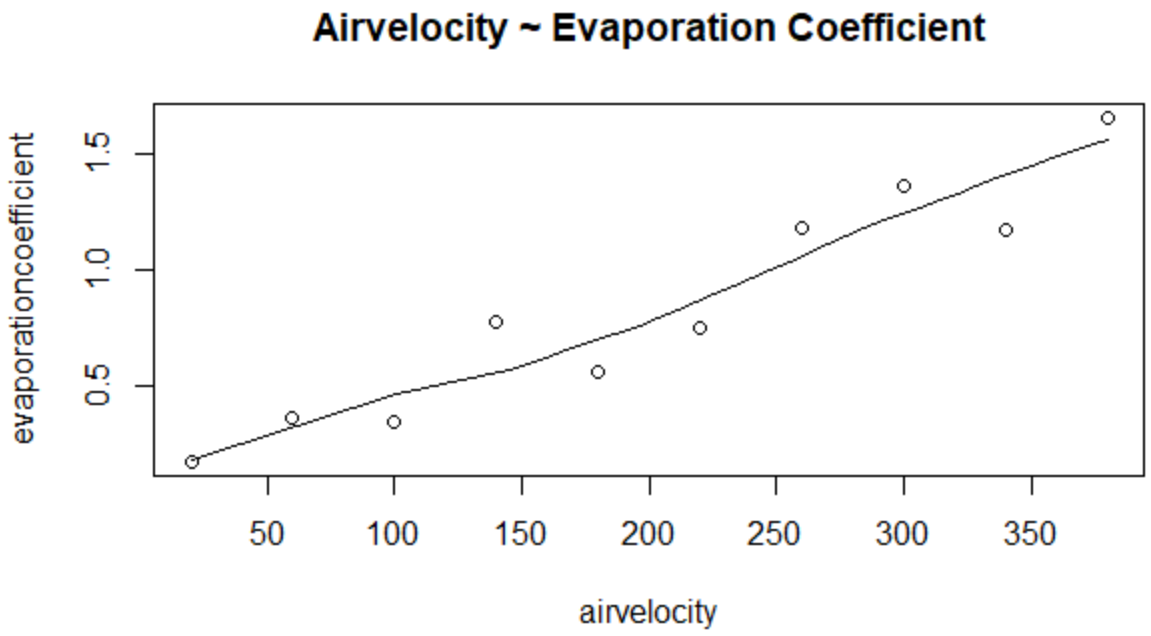
The aim of this exercise is to build a simple regression model that you can use to predict 'Evaporation Coefficient'. But before jumping in to the syntax, let's try to understand these variables graphically.

Typically, for each of the predictors, the following plots help visualize the patterns:

Using Scatter Plot to Visualize the Relationship

Scatter plots can help visualize linear relationships between the response and predictor variables. Ideally, if you have many predictor variables, a scatter plot is drawn for each one of them against the response, along with the line of best fit as seen below.

```
> scatter.smooth(airvelocity, evaporationcoefficient, main="Airvelocity ~ Eva
poration Coefficient")
```



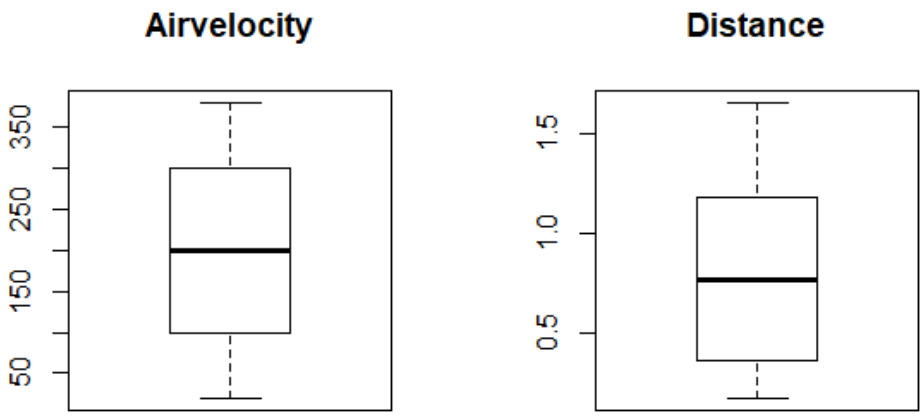
The scatter plot along with the smoothing line above suggests a linear and positive relationship between the 'Air Velocity' and 'Evaporation Coefficient'.

This is a good thing. Because, one of the underlying assumptions of linear regression is, the relationship between the response and predictor variables is **linear and additive**.

Using BoxPlot to Check for Outliers

Generally, an outlier is any datapoint that lies outside the 1.5 * inter quartile range (IQR). IQR is calculated as the distance between the 25th percentile and 75th percentile values for that variable (1).

```
> par(mfrow=c(1, 2))
> boxplot(airvelocity, main="Airvelocity", sub=paste("Outlier rows: ", boxplo
t.stats(airvelocity)$out))  # box plot for 'speed'
> boxplot(evaporationcoefficient, main="Distance", sub=paste("Outlier rows: "
, boxplot.stats(evaporationcoefficient)$out))  # box plot for 'distance'
```



Outlier rows: Outlier rows:

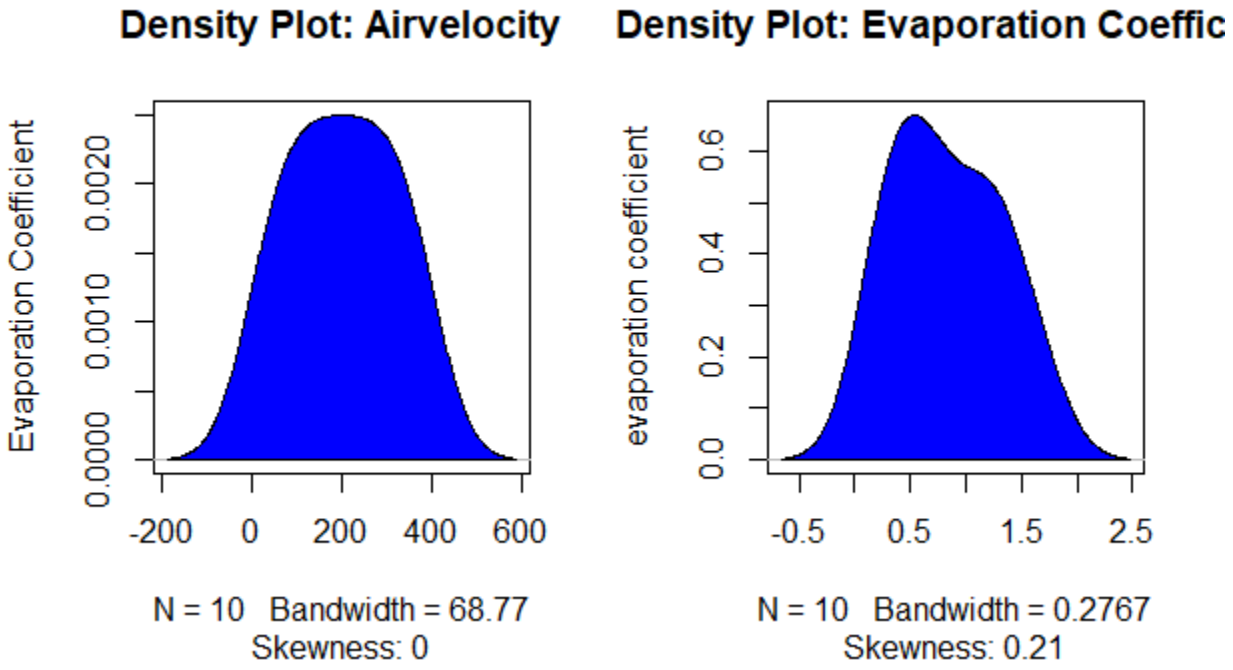
There are **no outliers** in the given data.

Using density plot to check if response variable is close to normal

Density Plot visualizes the distribution of data over a continuous interval or time period. This cha rt is a variation of a Histogram that uses kernel smoothing to plot values, allowing for smoother d istributions by smoothing out the noise. The peaks of a Density Plot help display where values ar e concentrated over the interval.

Skewness is the degree to which returns are asymmetric around the mean. Since a normal distrib ution is symmetric around the mean, skewness can be taken as one measure of how returns are n ot distributed normally. Left-skewed distributions are also called negatively-skewed distributions . That's because there is a long tail in the negative direction on the number line. The mean is also to the left of the peak. Right-skewed distributions are also called positive-skew distributions. Tha t's because there is a long tail in the positive direction on the number line. The mean is also to th e right of the peak.

```
> install.packages("e1071")
> par(mfrow=c(1, 2))
> plot(density(airvelocity), main="Density Plot: Airvelocity", ylab="Evaporat
ion Coefficient", sub=paste("Skewness:", round(e1071::skewness(airvelocity),
2)))  # density plot for 'Air Velocity'
> polygon(density(airvelocity), col="blue")
> plot(density(evaporationcoefficient), main="Density Plot: Evaporation Coeff
icient", ylab="evaporation coefficient", sub=paste("Skewness:", round(e1071::
skewness(evaporationcoefficient), 2)))  # density plot for 'dist'
> polygon(density(evaporationcoefficient), col="blue")
```



Correlation value and it's analysis

Correlation analysis studies the strength of relationship between two continuous variables. It involves computing the correlation coefficient between the two variables. Correlation is a statistical measure that shows the degree of linear dependence between two variables. In order to compute correlation, the two variables must occur in pairs, just like what we have here with 'Air velocity' and 'Evaporation Coefficient'. Correlation can take values between -1 to +1.

```
> cor(airvelocity, evaporationcoefficient)
```

```
[1] 0.9514814
```

Building the Linear Regression Model

The function used for building linear models is lm()

```
> linearMod <- lm(evaporationcoefficient ~ airvelocity)
> print(linearMod)

Call:
lm(formula = evaporationcoefficient ~ airvelocity)
Coefficients:
(Intercept)  airvelocity
   0.069242     0.003829
```

The results show the intercept and the beta coefficient for Evaporation Coefficient.

From the output above:

- The estimated regression line equation can be written as follow:

$$Evaporation\ coefficient\ =\ 0.069242 +\ 0.003829 * Air\ velocity$$

- The intercept ($\beta 1$) is 0.069242. It can be interpreted as the predicted Evaporation coefficient at zero Air velocity.
- The regression beta coefficient for the variable Air velocity ($\beta 1$), also known as the slope, is 0.048.

Model Assessment

Before using this formula to predict Evaporation coefficient, we should make sure that this model is statistically significant.

We start by displaying the statistical summary of the model using the R function summary():

```
> summary(linearMod)

Call:
lm(formula = evaporationcoefficient ~ airvelocity)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20103 -0.14671  0.05261  0.12318  0.17473

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0692424  0.1009737   0.686    0.512
airvelocity 0.0038288  0.0004378   8.746 2.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1591 on 8 degrees of freedom
Multiple R-squared:  0.9053,  Adjusted R-squared:  0.8935
F-statistic: 76.49 on 1 and 8 DF,  p-value: 2.286e-05
```

The summary outputs show 6 components, including:

Call

Shows the function call used to compute the regression model.

Residuals

Normally it gives a basic idea about difference between the observed value of the dependent variable (Y) and the predicted value (X), it gives specific detail i.e. minimum, first quarter, median, third quarter and max value, normally it does not used in our analysis.

Coefficients

The coefficients table, in the model statistical summary, shows:

- The estimates of the beta coefficients
- The standard errors (SE), which defines the accuracy of beta coefficients. For a given beta coefficient, the SE reflects how the coefficient varies under repeated sampling. It can be used to compute the confidence intervals and the t-statistic (2).

  We have standard error 0.1009737, 0.0004378 for $\beta 1$ and $\beta 2$ respectively that are close to o
- The t-value is calculated by taking the coefficient divided by the Std. Error. It is then used to test whether or not the coefficient is significantly different from zero. If it isn’t significant, then the coefficient really isn’t adding anything to the model and could be dropped or investigated further. Pr(>|t|) is the significance level.

  We have p value as 2.286e-05 the predefined statistical significance value is 0.05

Residual standard error

The RSE is an estimate of the standard deviation of ϵ. Roughly speaking, it is the average amount that the response will deviate from the true regression line (3). It is computed by

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R-squared value

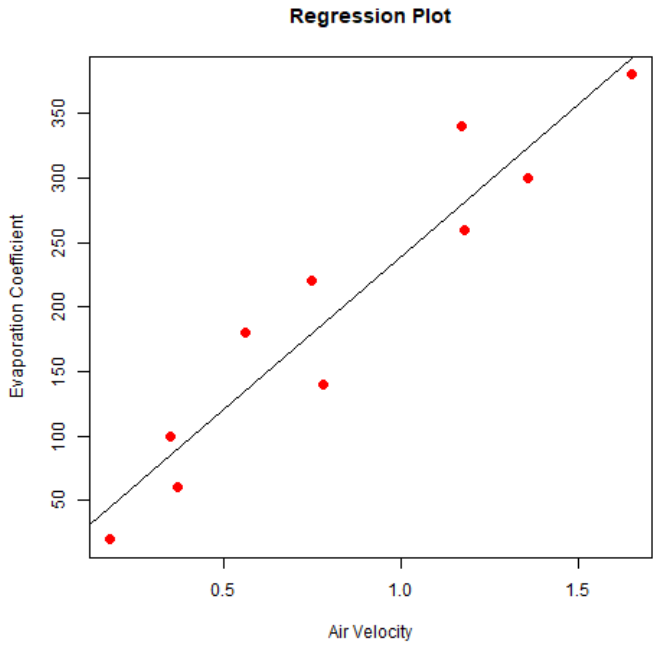
The R^2 statistic provides an alternative measure of fit. It represents the proportion of variance explained and so it always takes on a value between 0 and 1, and is independent of the scale of Y. R^2 is simply a function of residual sum of squares (RSS) and total sum of squares (TSS):

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Plotting Regression

Now, we plot out prediction results with the help of the plot() function. This function takes parameter x and y as an input vector and many more arguments.(4)

```
> png(file ="linear_regression.png")
> plot(evaporationcoefficient,airvelocity, col = "red",main = "Regression Plo
t",abline(lm(airvelocity ~ evaporationcoefficient)),cex = 1.3,pch = 16,xlab =
"Air Velocity",ylab = "Evaporation Coefficient")
> dev.off()
```



Reference

1. Complete Introduction to Linear Regression in R [Internet]. Available from: https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/

2. Simple Linear Regression in R [Internet]. Available from: http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/

3. UC Business Analytics R Programming Guide. Available from: https://uc-r.github.io/linear_regression

4. Linear Regression [Internet]. Available from: https://www.javatpoint.com/r-linear-regression