

## Chapter 7

# Data Analytics 5: Simple Linear Correlation & Regression

*Prepared by Dr Yuslina Zakaria*

In this practical session, you will be learning on how to examine the correlation between two variables using both parametric and nonparametric statistical techniques. We will then examine the statistical significance of the relationship and proceed to the simple linear regression to create a linear regression model and predict the dependent variable from the given independent variable.

### R Packages:

For this lesson, we only use `stats` package provided by the R base functions. No additional package is needed.

### Datasets

A study was conducted among 100 women between age of 20 to 30 years old to investigate the relationship between androgen sex trait (height) and reproductive ambition score (ideal number of children, ideal own age at first child). Maternal personality score (importance of having children, self-rated maternal/broodiness) and career orientation score (importance of having career) were also recorded as additional features.

+ ``ambition.csv`` - ``height`` is measure in cm.


The data set can be downloaded from [http://tiny.cc/phc410\\_da5](http://tiny.cc/phc410_da5).

### Magnitude of Correlation Coefficient

The strength of correlation can be interpreted as follows:

- 0.00 - 0.39 (Weak correlation)
- 0.40 - 0.79 (Moderate correlation)
- 0.80 - 0.99 (Strong correlation)
- 1.00 (Perfect correlation)

### Instructions:

-  explains the steps for activity you need to follow.

### Preparing the data



Import `ambition.csv` from `input` folder as `traits` using `read.csv()`.

```
traits <- read.csv("input/ambition.csv")

#observe reprod
str(traits)
## 'data.frame':    100 obs. of  5 variables:
## $ ID           : int  1 2 3 4 5 6 7 8 9 10 ...
## $ height       : int 170 172 156 152 153 165 145 153 153 160 ...
## $ reproductive: int  60 55 75 70 65 70 70 75 70 75 ...
## $ career       : int  80 85 60 65 55 60 60 65 65 70 ...
## $ maternal     : int  65 60 80 75 70 70 75 80 75 80 ...
```

## 7.1 Pearson Correlation

### Assumptions

The correlation analysis has two general assumptions:

1. Both independent and dependent variables are normally distributed.
2. The relationship between the two variables is linear.

We can use a histogram, a Q-Q plot and also statistical test i.e. Shapiro-Wilk test to examine if both variables are normally distributed. These steps for visual inspections have been covered in previous practical session.

Scatter plots can help us to visualise linear relationships between the independent and dependent variables. As covered in the practical session of Exploratory Data Analysis, a scatter plot can be drawn using `plot()` function.

To calculate a Pearson correlation coefficient ( $r$ ), we use:

```
# Pearson correlation coefficient
cor.test(x, y, method="pearson") # x = independent variable & y = dependent variable
```

`cor.test()` function returns both the correlation coefficient and the significance level (or p-value) of the correlation.

### Question

Suppose we wanted to examine the relationship between height and reproductive ambition score.

### Hypothesis

$H_0$ : There is no relationship between the height and reproductive ambition score ( $r = 0$ )

$H_a$ : There is a relationship between the height and reproductive ambition score ( $r \neq 0$ )

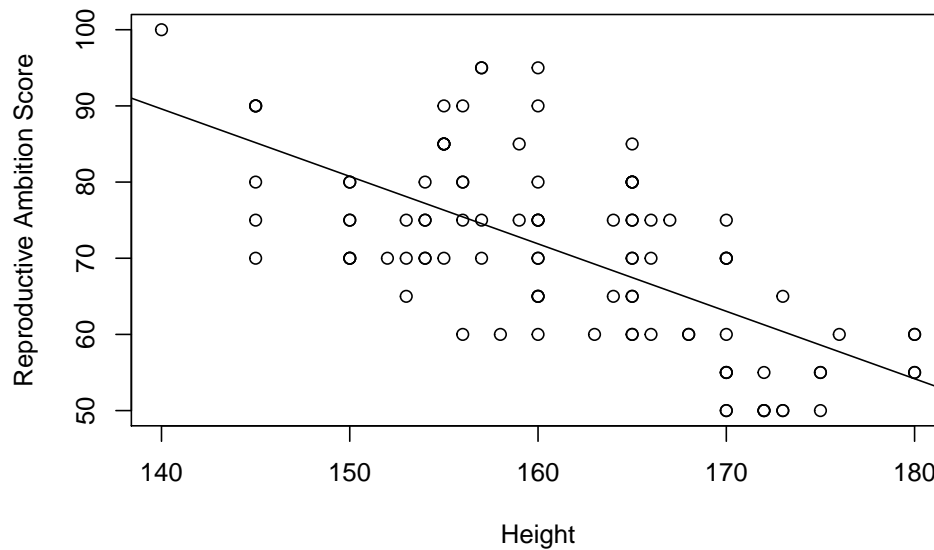
### Performing the test

Assuming both variables are normally distributed, let us draw the scatter plot between the two variables.



Draw scatter plot between height and reproductive

```
plot(traits$height, traits$reproductive, xlab="Height",
      ylab="Reproductive Ambition Score")
abline(lm(traits$reproductive ~ traits$height))
```



For the scatter plot, we can predict that there is a negative linear relationship between the two variables. Now, we perform the Pearson correlation test to get the correlation coefficient of the relationship.



Perform Pearson's correlation test between `height` and `reproductive`

```
# Pearson correlation coefficient
cor.test(trait$height, trait$reproductive, method="pearson")
##
## Pearson's product-moment correlation
##
## data: trait$height and trait$reproductive
## t = -8.8028, df = 98, p-value = 0.00000000000004778
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.7615238 -0.5383463
## sample estimates:
## cor
## -0.6644995
```

### Conclusion and Interpretation

Based on the results,  $p\text{-value} (0.000) < 0.05$  and thus we reject the null hypothesis. There is a significant linear correlation between height and reproductive ambition score. We can conclude that there is a moderate negative relationship between height and reproductive ambition score ( $r = -0.66$ ). The taller the woman was, the weaker was her reproductive ambition score.

## 7.2 Spearman's Rank Correlation

Spearman's rank correlation is a nonparametric counterpart of the Pearson correlation. It is based on rank of observations and more appropriate to be used for ordinal data. It is used when Pearson's assumptions are violated.

To perform the Spearman's rank correlation, we use:

```
# Spearman's rank correlation coefficient
cor.test(x, y, method="spearman") # x = independent & y = dependent
```

### Question

Suppose we wanted to examine the relationship between height and career orientation score. Assume that the data violate the assumptions of Pearson's  $r$ .

### Hypothesis

$H_0$ : There is no relationship between the height and career orientation score ( $\rho = 0$ )  $H_a$ : There is a relationship between the height and career orientation score ( $\rho \neq 0$ )

### Performing the test

Now, we perform the Spearman's rank correlation to examine the relationship of both variables.



Evaluate the relationship between `height` and `career` using Spearman's rho. Use `exact=FALSE` to consider ties.

```
# Spearman's rank correlation coefficient
cor.test(traits$height, traits$career, method="spearman", exact = FALSE)
##
## Spearman's rank correlation rho
##
## data: traits$height and traits$career
## S = 60551, p-value = 0.000000000001083
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.6366589
```

### Conclusion and Interpretation

The p-value ( $0.000$ )  $< 0.05$  and thus we reject the null hypothesis. There is a significant linear correlation between height and career score. We can conclude that there is a moderate positive relationship between height and career score ( $\rho = 0.64$ ). The taller the woman was, the stronger was her career score.

## 7.3 Simple Linear Regression

Simple linear regression can be used to predict the dependent variable given the value of the independent variable. It assumes that there is a linear relationship between both variables and the relationship

is statistically significant.

In R, we use `lm()` function to build a linear regression model.

Suppose we wanted to build a linear model between `height` and `career`, where `height` is the independent variable and `career` is the dependent variable.



Build a linear model between `height` and `career`.

```
# Build a simple linear model - height (independent) & career (dependent)
linearMod <- lm(career ~ height, data=traits)
linearMod
##
## Call:
## lm(formula = career ~ height, data = traits)
##
## Coefficients:
## (Intercept)      height
##    -42.9280      0.6929
```

From the output, we can see that the y-Intercept value is -42.93 and the `height` or the slope is 0.69. Therefore, the formula of the linear model can be given as:

$$Y = 0.69X - 42.93$$

## 7.4 Regression Model Predictions

Before we use the regression model, we need to ensure that the model is statistically significant by examining the `summary()` of generated `linearMod`.



Examine the significance of linear regression model.

```
# Print the summary statistics of linearMod
summary(linearMod)
##
## Call:
## lm(formula = career ~ height, data = traits)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3943  -5.1585   0.3379   5.4128  20.1414
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept) -42.92799    13.63087  -3.149    0.00217 **
## height       0.69286     0.08432   8.217 0.000000000000872 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.627 on 98 degrees of freedom
## Multiple R-squared:  0.4079, Adjusted R-squared:  0.4019
## F-statistic: 67.52 on 1 and 98 DF, p-value: 0.0000000000008721
```

We can consider a linear model to be statistically significant if both p-values (for the model and the predictor variable) are less than significance level of 0.05.

Based on the output, both model p-value (0.00217) and the predictor p-value (0.000) are less than 0.05. Therefore, the linear regression model is significant and can be used to make predictions.

### Question

What is the career orientation score if the height of Ms X is 163cm?



Predict `career` score if `height` is 163.

```
# predict career score using linearMod
predict(linearMod,data.frame(height = 163))
##          1
## 70.00858
```

### Interpretation

Using the linear regression model of  $y = 0.6929x - 42.93$ , the career orientation score is 70 given the value of height is 163.