Chapter 15

Relationship between two quantitative variables; Scatterplots and Correlations.

In this chapter we will discuss the following topics:

- How to make scatterplots with the R-function **plot()**. The scatterplot is used to investigate the relationship between two quantitative variables.
- How to calculate the correlation between two variables using the R-function **cor()**.

Scatterplots

The relationship between two quantitative variables can be displayed graphically by a scatterplot. The values of the explanatory variable appear along the horizontal axis and the values of the response variable appear along the vertical axis. Each subject in the data is represented by a point. If there is no clear distinction between which variable is the explanatory variable or the response variable, we can place either one of them on the horizontal axis.

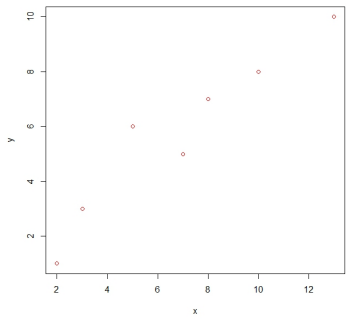
Problem. Let

```
 x  y
 2  1
 3  3
 5  6
 7  5
 8  7
10  8
13 10
```

Make a scatterplot of y versus x.

Solution. We have

```
> x=c(2,3,5,7,8,10,13)
> y=c(1,3,6,5,7,8,10)
> plot(x,y,col="red")
```



Problem. The following data [1] shows the gold medal performance in men's long jump in the Olympic Games from 1900-1984. The distance of the long jump is measured in inches.

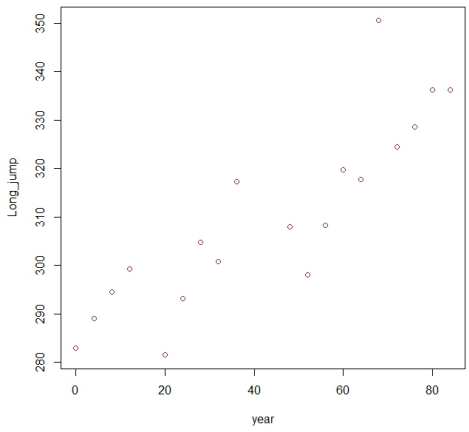
Long_jump	year
282.8750	0
289.0000	4
294.5000	8
299.2500	12
281.5000	20
293.1250	24
304.7500	28
300.7500	32
317.3125	36
308.0000	48
298.0000	52
308.2500	56
319.7500	60
317.7500	64
350.5000	68
324.5000	72
328.5000	76
336.2500	80
336.2500	84

Make a scatterplot to examine the relationship between the year of the Olympic and the gold medal performance in men's long jump. It has been claimed that the Olympic Games in Mexico City in 1968 had an unusual performance in track and field, particularly in long jump possible caused by the high altitude. What do you find?

Solution. The explanatory variable is the *year* and the response variable is the *distance of the long jump* in inches. We do the following:

```
> table=read.delim("Long_Jump.txt")
> attach(table)
> plot(year,Long_jump,col="brown")
```

There is a clear *direction* in the plot; the data points moves from lower left to upper right. That is, recent Olympics Games tend to give greater distances in long jump. The *form* of the relationship is approximately a straight line. The *strength* of the relationship is strong in particular if we omit the outliers; There are lower performances in long jump in the Olympic Games after the two Word Wars while there is an extreme high performance in 1968. We will examine these data closer in chapter 24.



Explanation. The code can be explained as follows:

- When we use the command **attach(table)**, we attach the table or data frame to the search path such that we can refer to the variables in the table or data frame by their name.

Including Categorical Variables in Scatterplots

We can add a categorical variable to a scatterplot by representing each category by a different plot symbol or color.

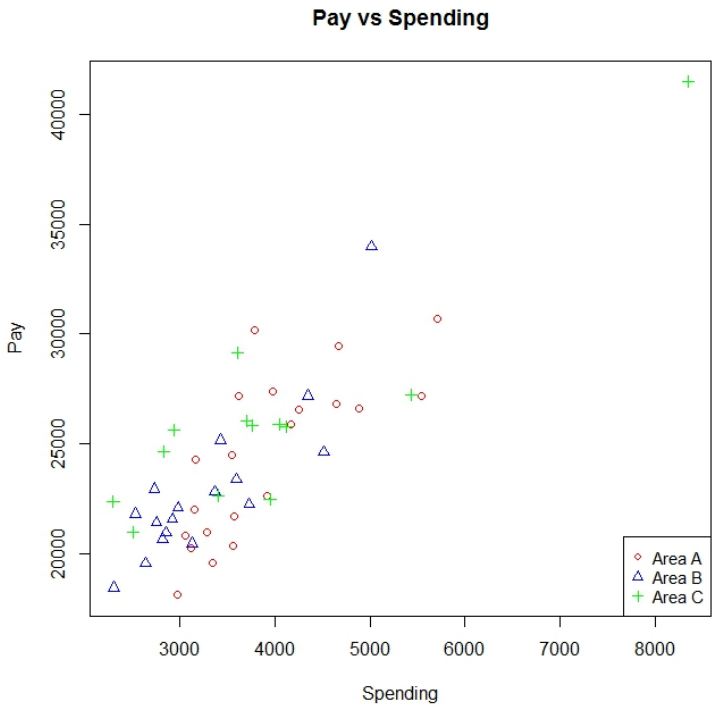
Problem. The following data [2] (somewhat modified) shows the average annual teacher pay in public schools versus spending on public schools per pupil in dollar in 1985 for 50 states and the District of Columbia. Region A is the Northeast and North Central, region B is the South, and region C is the West.

State	Pay	Spending	Area
ME	19583	3346	A
NH	20263	3114	A
VT	20325	3554	A
MA	26800	4642	A
..	..	..	.
KA	22644	3914	A
DE	24624	4517	B
MD	27186	4349	B
..	..	..	.
TX	25160	3429	B
MT	22482	3947	C
..	..	..	.
HA	25845	3766	C

Make a scatterplot showing the relationship between spending and pay for all 50 states and the District of Columbia. Use separate symbols to distinguish between the three areas.

Solution. *Area* is a categorical variable with three levels, A,B, and C. We will represent *Area* A by an open circle of red color, *Area* B with a triangle of blue color, and *Area* C with a plus symbol of green color. We let *spending* be the explanatory variable and *pay* the response variable. We do the following:

```
> data=read.delim("School_Spending.txt")
> attach(data)
> Area=as.numeric(Area)
> Area
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[39] 3 3 3 3 3 3 3 3 3 3 3 3 3
> plot(Spending,Pay,pch=c(1,2,3)[Area],col=c("red","blue","green")[Area],
 + main="Pay vs Spending")
> legend("bottomright", c("Area A", "Area B", "Area C"),cex=0.9,pch=c(1,2,3),
 + col=c("red","blue","green"))
```



From the plot, we can see that there is a positive association between spending and pay for all areas. Area C has higher average teacher pay in states on the lower end of spending on public schools compared to areas A and B, where B has the lowest value. There is an outlier in terms of spending on public school for the state of Alaska.

Explanation. The code can be explained as follows:

- The command **Area=as.numeric(Area)** assigns numerical values to the levels in the categorical variable **Area**. It assigns the values in alphabetic order such that $A = 1$, $B = 2$, and $C = 3$.

- The entry **pch** specifies the type of point to use (i.e. circle, plus, triangle, etc.). The range of possible values are 0-25 for symbols and 32-255 for characters. The default is 1.
- The entry **pch=c(1,2,3)[Area]** assigns different types of symbols to the points for the levels **A=1**, **B=2**, and **C=3** in the vector variable **Area**.
- The entry **col=c("red","blue","green")[Area]** assigns different colors to the points for the levels **A=1**, **B=2**, and **C=3** in the vector variable **Area**.

Correlation

The correlation measures the association between two quantitative variables. Its range of values is from -1 to 1 for which -1 and 1 indicate a perfect linear relationship while 0 indicates no linear relationship. A positive sign indicates that both variables simultaneously take on large or small values. A negative sign indicates that one variable takes on large values when the other one takes on small values. The correlation, r, between x and y in a sample of n individuals is defined by

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right),$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y, with standard deviations s_x and s_y, respectively. This is called the Pearson Correlation.

Problem. Calculate the correlation between x and y given in the first problem.

Solution. We do the following:

```
> x=c(2,3,5,7,8,10,13)
> y=c(1,3,6,5,7,8,10)
> cor(x,y)
[1] 0.9541879
```

The correlation of $r = 0.954$ indicates a strong positive linear relationship between x and y.

Problem. Calculate the correlation between the year of the Olympic Games and the distance in long jump in problem 2 above.

```
> cor(year,Long_jump)
[1] 0.8703381
```

The correlation of $r = 0.870$ indicates a strong positive linear relationship between the year of the Olympic Games and the distance in long jump.

References

[1] This data is distributed with the software package, Data Desk. Data Description, Inc. Data Desk. Ithaca, NY: Data Description, Inc., 1993. The data and story are also found on DASL: http://lib.stat.cmu.edu/DASL/DataArchive.html

[2] National Education Association, as reported by the Albuquerque Tribune, 11/7/86. The data and story are also found on DASL: http://lib.stat.cmu.edu/DASL/DataArchive.html

00