statstutor

community project

encouraging academics to share statistics support resources

stcp-karadimitriou-correlationR

The following resources are associated:
Scatterplots in R, Regression in R, dataset Birthweight reduced.csv' and the correlation in R script file

Correlation in R

Variables: Two continuous (scale) variables.

Common Applications: Exploring the (linear) relationship between two variables; e.g. as variable X increases does variable Y increase or decrease?

Pearson's correlation measures the existence (given by a p-value), strength and direction (given by the coefficient *r* between -1 and +1) of a linear relationship between two variables. The exact size of the coefficient is a measure of the strength of the correlation (with 1 being a perfect positive correlation). The further away r is from 0, the stronger the relationship. If the outcome is significant, conclude that a correlation exists but use the correlation coefficient to describe the relationship.

Guidelines for interpretation of a correlation coefficient

Correlation coefficient			Association
-0.3 < r < 0.3			Weak
-0.5 to -0.3	or	0.3 to 0.5	Moderate
-0.9 to -0.5	or	0.5 to 0.9	Strong
-1 to -0.9	or	0.9 to 1	Very strong

Data: The data set '*Birthweight reduced.csv'* contains details of 42 babies and their parents at birth. The question of interest here is whether Birth weight (lbs) and gestational age of the baby at birth (in weeks) are related.

Birthweight	Gestation
5.80	33
4.20	33
6.40	34

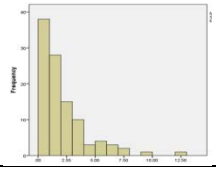
Open the birthweight reduced dataset from a csv file, call it birthweightR and attach the data so just the variable name is needed in commands.

```
birthweightR<-read.csv("D:\\Birthweight reduced.csv",header=T)
attach(birthweightR)
```

Research question: Is there a relationship between gestational age and birthweight?
Null Hypothesis: (H_0): There is no correlation between gestational age and birthweight (equivalent to saying r = 0)
Alternative hypothesis (H_1): There is a correlation gestational age and birthweight (equivalent to saying $r \neq 0$).

© Sofia Maria Karadimitriou and Ellen Marshall
University of Sheffield
Reviewer: Jim Bull
University of Swansea
Based on material by Peter Samuels and Mollie Gilchrist of Birmingham City University

Assumptions for correlation

Assumptions	How to check	What to do if assumption is not met
Linearly related continuous variables	Scatter plot of two variables. This is an example of a non-linear relationship 	Transform data
Both variables are normally distributed	Histograms of variables/ Shapiro Wilk test of normality. This is an example of very skewed data 	Use rank correlation for ordinal or skewed variables: Spearman's or Kendall tau

Steps in R

Step 1: Draw a scatterplot of the data to see any underlying trend in the relationship. One of the assumptions for Pearson's correlation is that the variables are linearly related.

To plot a scatterplot of the two variables, use the `plot()` command

```
plot(Gestation,Birthweight,main='Scatterplot of gestational age and birthweight',xlab='Gestation (weeks)',ylab='Birthweight(lbs)')
```

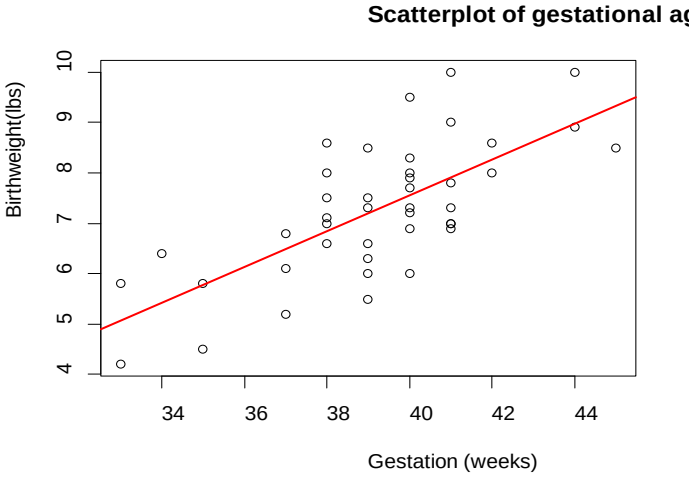
The **Scatterplots in R** resource gives more details on customising the scatterplot.

To fit a red line of best fit through the scatter use

```
abline(lm(Birthweight~Gestation),col='red',lwd=2)
```

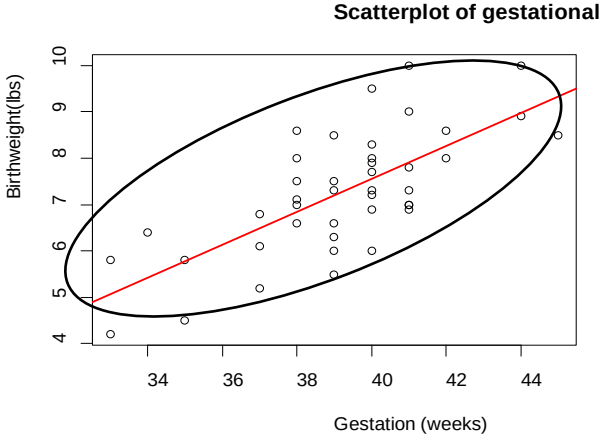
In this example there is perhaps an underlying assumption that birth weight is affected by gestational age. Therefore birthweight has been placed on the vertical (Y) axis and the gestational age has been placed on the horizontal (X) axis.

The circles on the scatterplot are reasonably closely scattered about an underlying straight line (as opposed to a curve or a random scattering), a linear relationship between the two variables is confirmed. The scatterplot implies that as the gestational age increases, so does birthweight so the Pearson correlation coefficient will be positive. An example of negative correlation is the amount spent on heating and daily temperature: as the temperature increases the amount spent on heating decreases.



It is also important that there are no large outliers and that the values of the birthweight variable do not vary more as the values of the gestational age variable change. This means that most of the points lie within an ellipse or cigar shape orientated in the direction of the linear model (see diagram on right). Outliers can have a significant impact on the strength of a relationship. There are none here.

If no underlying straight line then there is no point going on to the next calculation.



Step 2: Calculate the correlation coefficient

Pearson's correlation coefficient can be produced using the `cor()` command, however it does not conduct any significance tests. Use the `cor.test()` command to carry out a test of significance. If there is missing data, use `cor(...,use="complete.obs")` in both commands.

The command `cor(Birthweight, Gestation)` produces this output

```
> cor(Birthweight,Gestation)
[1] 0.7062919
```

The correlation coefficient of 0.71 suggests a strong positive correlation between birthweight and gestation.

The command `cor.test(Birthweight, Gestation)` tests the hypothesis r = 0.

```
        Pearson's product-moment correlation

data:  Birthweight and Gestation
t = 6.31, df = 40, p-value = 1.733e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5123421 0.8316893
sample estimates:
      cor
0.7062919
```

As the p value for the test is much smaller than 0.05 ($p < 0.001$), the null hypothesis ($r = 0$) is rejected. There is strong evidence to suggest that the correlation coefficient is different to 0.

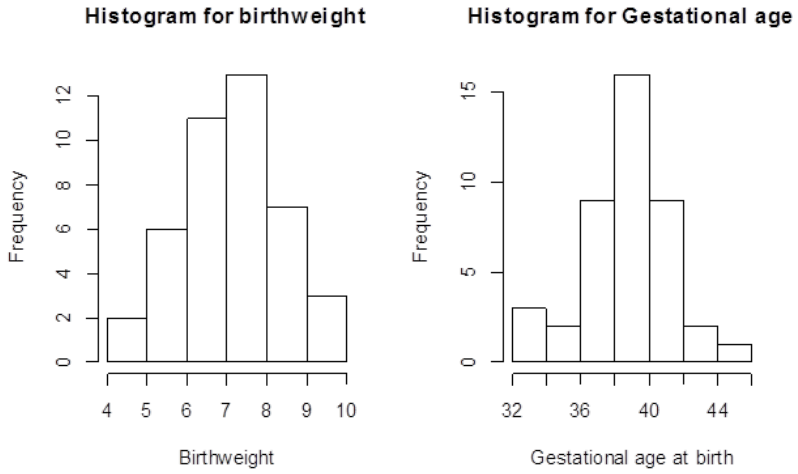
Check that the variables are normally distributed using histograms. See the **Checking normality in R** resource for more details. Plot the histograms for the birthweight of babies and gestational age next to each other.

```
par(mfrow=c(1,2))

hist(Birthweight,main='Histogram for
birthweight',xlab='Birthweight')

hist(Gestation,main='Histogram
for Gestational
age',xlab='Gestational age at
birth')
```

Both variables are normally distributed.



Reporting correlation

The relationship between gestational age at birth and birth weight was investigated using Pearson's correlation. There was significant evidence ($p < 0.001$) to suggest an association between gestational age and birthweight. Pearson's correlation coefficient was 0.71 indicating a strong positive relationship.

The correlations between several variables can be displayed in a table using `cbind()`.
The round command rounds numbers to a specified number of decimals, for example rounded to 2 decimal places.

```
round(cor(cbind(Birthweight,Gestation,mppwt,mheight)),2)
```

	Birthweight	Gestation	mppwt	mheight
Birthweight	1.00	0.71	0.39	0.37
Gestation	0.71	1.00	0.25	0.23
mppwt	0.39	0.25	1.00	0.67
mheight	0.37	0.23	0.67	1.00

Notes:

1. The p-value for a Pearson correlation test and the Pearson correlation coefficient are not the same thing. The larger the sample size, the lower the value of *r* at which a significant result occurs. For small samples it is possible to have a high correlation coefficient which is not significant and for large samples it is possible to have a small correlation coefficient which is significant. Thus it is **important to look at the value of *r* as well as the p-value.**
2. We cannot conclude that knowledge about gestational age causes an increase in birthweight. Perhaps a third (mediating) variable is involved? Causality can only be established by a randomised control trial.

Comments

- Conclusions are only valid within the range of data collected.
- Pearson correlation also assumes the data values are independent. If the assumptions of Pearson correlation are not met or the data is ordinal, other coefficients can be calculated:
    - Kendall's τ ('tau') measures the degree to which a relationship is always positive or always negative. Use `cor(dependent, independent, method = "kendall")`
    - Spearman's coefficient of rank correlation, ρ ('rho'), behaves in a similar way to Kendall's τ but has a less direct interpretation. Use `cor(dependent, independent, method = "spearman")`