Mathematics Learning Support Centre

Statistics: Correlation

Richard Buxton. 2008.

1 Introduction

We are often interested in the relationship between two variables.

- Do people with more years of full-time education earn higher salaries?
- Do factories with more safety officers have fewer accidents?

Questions like this only make sense if the possible values of our variables have a natural order. The techniques that we look at in this handout assume that variables are measured on a scale that is at least *ordinal*. In discussing Pearson's correlation coefficient, we shall need to go further and assume that we have interval scale data - i.e. that equal intervals correspond to equal changes in the characteristic that we are trying to measure.

2 Plotting the data

The first step in looking for a correlation is to draw a scatterplot of the data. Figure 1 shows four examples[1].

2.1 Interpreting a scatterplot

These are some of the points to look for.

- How strong is the relationship?
  - In Figure 1, the relationship between gas consumption and outside temperature is very strong, while the relationship between Educational level and Crime rate is much weaker.
- Is the relationship increasing or decreasing?
  - In the 'Gas' example, higher outside temperatures are associated with *lower* gas consumption, but in the 'Ice cream' example, higher mean temperatures go with higher levels of ice cream consumption.

[1] Source of data: Hand(1994)

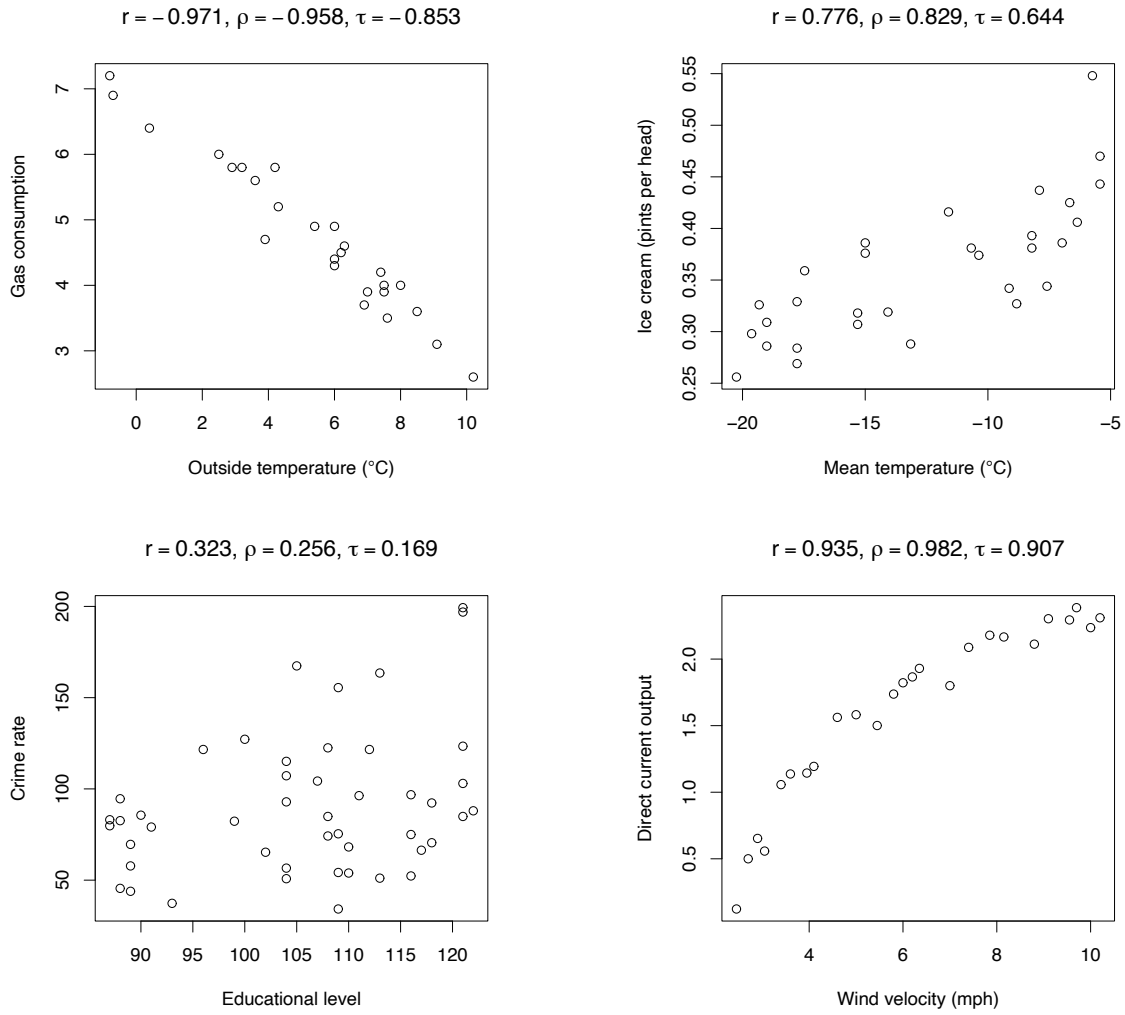


Figure 1: Scatterplots showing strong and weak relationships

- Is the relationship roughly linear?
  - The plot in the top left of Figure 1 shows a clear linear pattern, while the plot in the bottom right suggests a non-linear relationship with the initial steep slope leveling off as the wind speed increases.
- What is the *slope* of the relationship?
  - Is an increase in one variable associated with a small, or a large, increase in the other one? For example, factories with more safety officers may have fewer accidents, but is the reduction in accidents large enough to justify the cost of the additional safety officers?
- Are there any *outliers*?
  - Figure 2 shows a plot of Police expenditure per head against Population size

for 47 US states[2]. At first glance, there seems to be an increasing relationship, with larger states spending more per head on policing. But if you cover up the two points at the top right of the plot, the correlation seems to disappear. The evidence for a correlation comes almost entirely from these two points, so we'll need to check our data source to make sure that the points really are correct.

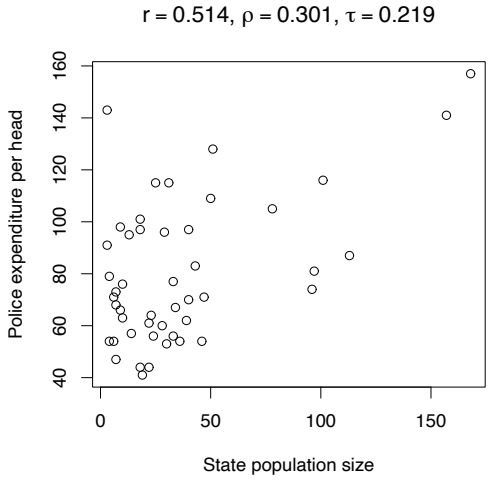


Figure 2: Effect of outliers

2.2 Scatterplots in SPSS

The simplest way to draw a scatterplot in SPSS is to use the *Chart Builder*.

—**Graphs**
—**Chart Builder**
—choose **Scatter/Dot**
—drag the **Simple Scatter** plot into the plotting region
—drag the variables that you want to plot into the **X-Axis** and **Y-Axis** boxes
—Click **OK**

If your data can be split into distinct groups - for example, by gender, you may find it helpful to use a **Grouped Scatter** plot, instead of a **Simple Scatter** plot. Put the two main variables on the x and y axes, as above, but then drag the grouping variable (e.g. gender) into the **Set Colour** box.

If you want to look at all pairwise correlations among a group of variables, use a scatterplot matrix. Drag the **Scatterplot Matrix** into the plotting region and drag all your variables into the **Scattermatrix** box.

[2]Source of data: Hand(1994)

3 Correlation coefficients

A correlation coefficient gives a numerical summary of the degree of association between two variables - e,g, to what degree do high values of one variable go with high values of the other one?

Correlation coefficients vary from -1 to +1, with positive values indicating an increasing relationship and negative values indicating a decreasing relationship.

We focus on two widely used measures of correlation - Pearson's r and Kendall's τ.

- Pearson's coefficient
  - measures degree to which a relationship conforms to a straight line
- Kendall's coefficient
  - measures degree to which a relationship is always increasing or always decreasing

Spearman's rank correlation coefficient, ρ behaves in much the same way as Kendall's τ, but has a less direct interpretation.

3.1 Which coefficient should I use?

- Interval scale data and interested in linear relationships - e.g. wish to build linear model
  - Use Pearson's coefficient
- Interested in *any* increasing/decreasing relationship
  - Use Kendall's coefficient

3.2 Pearson's coefficient

Suppose we have n data pairs (x_i, y_i)

Pearson's correlation coefficient is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

... where $\overline{x}$ and $\overline{y}$ are the means of the x and y values.

In practice, we always use statistical software to do the calculations.

Looking back at Figure 1, notice how the absolute size of the coefficient drops towards zero as we get more and more scatter. While Pearson's r is good at measuring the strength of

a *linear* association, it can be quite misleading in the presence of curvature. Look at the wind turbine data in the bottom right plot of Figure 1. Pearson's r is 0.935, suggesting a strong linear association, but a linear model would clearly not be sensible here.

Because Pearson's r is based on the idea of linearity, it only makes sense for data that is measured on at least an *interval* scale. For ordinal data, use Kendall's τ or Spearman's ρ.

3.3 Kendall's coefficient

Kendall's τ can be used with any variables that are at least ordinal.

Each pair of data points is classified as concordant, discordant or tied.

- Concordant
  - Both variables increase or both variables decrease
- Discordant
  - One variable increases while the other one decreases
- Tied
  - One or both variables stays constant

Writing C, D and T for the number of concordant, discordant and tied pairs, Kendall's coefficient is given by...

$$\tau = \frac{C - D}{N}$$

...where $N = C + D + T$ (the total number of pairs).

The idea is that *concordant* pairs suggest an increasing relationship, while *discordant* pairs suggest a decreasing relationship. Kendall's τ is just the proportion of concordant pairs minus the proportion of discordant pairs.

The value of τ gives a measure of the degree to which a relationship is always increasing, or always decreasing - see Figure 1.

3.4 Modification of Kendall's coefficient for tied data

If some of the pairs of observations are tied, Kendall's coefficient cannot reach the limiting values of ± 1 even if all untied pairs are concordant (discordant). This is a particular problem in the analysis of contingency tables, where there will usually be a large number of ties. Kendall proposed the following as an alternative to the simpler coefficient defined above.

$$\tau_b = \frac{C - D}{\sqrt{(n(n-1)/2 - t_x)(n(n-1)/2 - t_y)}}$$

... where t_x is the number of tied x values and t_y is the number of tied y values.

This version of Kendall's τ is the one used by SPSS.

3.5 Interpreting a correlation coefficient

It's easy to misinterpret a correlation coefficient. These are some of the points to watch.

- A correlation coefficient can be badly affected by one or two outlying observations. For the 'police expenditure' data in Figure 2, the value of Pearson's r is 0.514, but if the two points at the top right of this plot are removed, the correlation drops to 0.237. Always look at a scatter plot before calculating a correlation coefficient!

- Correlation is not the same as causality. For example, factories with more safety officers may have fewer accidents, but this doesn't prove that the variation in accident levels is attributable to the provision of safety officers. The correlation may be a spurious one induced by another factor such as the age of the factory.

  One possible approach here is to use *partial correlation*. We 'adjust' our two variables to remove any variation that can be accounted for by our third variable (age of factory) and then look at the correlation between the two adjusted variables.

- Even if a relationship is genuine, a strong correlation doesn't necessarily imply that a change in one variable will produce a large change in the other one. The two sets of data shown in Figure 3 give the same correlation coefficient, but say quite different things about the effect of engine capacity on fuel economy.

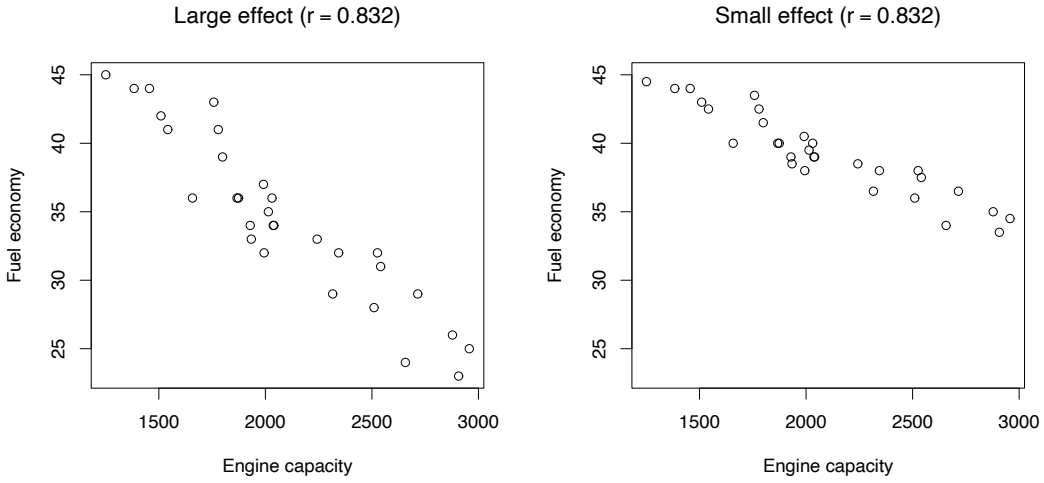


Figure 3: Correlation and size of effect

- Correlation coefficients are subject to sampling variation and may give a misleading picture of the correlation in the population we're sampling. We can quantify the uncertainty in an estimate of a correlation by quoting a confidence interval, or

range of plausible values. For the 'Ice cream' data in Figure 1, the 95% confidence interval for Pearson's r is 0.576 to 0.888, so we can be fairly sure that the population coefficient lies in this range.

For details of how to calculate confidence intervals for correlation coefficients, see Howells(1994) and Hollander(1999).

3.6 Testing for zero correlation

Most statistical software packages allow us to check whether a sample correlation is compatible with zero correlation in the population we're sampling. The test that is carried out here first assumes that the population correlation is zero and calculates the chance of obtaining a sample correlation as large or larger in absolute size than our observed value - this chance is given as the *p value*. If the *p value* is very small, we conclude that our sample correlation is probably incompatible with zero correlation in the population.

The limitation of a test for zero correlation is that it doesn't tell us anything about the *size* of the correlation. A correlation can be nonzero, but too small to be of any practical interest. For example, if we test for zero correlation with the data in the plot in the bottom left of Figure 1, we obtain p value of 0.027, which gives strong evidence for a nonzero correlation. But would a relationship as weak as this be of any practical interest?

3.7 Correlation coefficients in SPSS

—**Analyze**
—**Correlate**
—**Bivariate**
—Drag the two variables that you want to correlate into the **Variables** box
—Select the required correlation coefficients
—Click **OK**

Table 1 shows the SPSS output for the Ice cream data shown in Figure 1. This table relates to Pearson's coefficient - the output for Kendall's τ and Spearman's ρ is similar.

Correlations

		Consumption	Temperature
Consumption	Pearson Correlation	1.000	.776
	Sig. (2-tailed)		.000
	N	30	30
Temperature	Pearson Correlation	.776	1.000
	Sig. (2-tailed)	.000	
	N	30	30

Table 1: SPSS Correlation output

Each box of the table contains the information on the correlation between the corresponding row and column variables. Looking at the top right box, Pearson's r is 0.776, suggesting a moderately strong increasing relationship. The second figure is the p-value for a test of the hypothesis that the population correlation is zero. The figure here has been rounded to 3 decimal places, so the figure of 0.000 tells us that the p-value is less than 0.0005. This is a very small probability, so we can be almost certain that the population correlation is not zero. The third figure tells us the number of observations in our sample.

Unfortunately, SPSS does not provide confidence intervals for correlation coefficients. One package that does offer confidence intervals for both Pearson's and Kendall's coefficients is the package *StatsDirect* - see StatsDirect (2008).

4 References

For a simple introduction to correlation, see Moore (2004). For a more comprehensive treatment, see Howell (2002).

Hand, D.J. (1994). A Handbook of Small Data Sets Chapman and Hall, London.

Hollander, M. and Wolfe, D.A. (1999). Nonparametric Statistical Methods, Wiley, New York.

Howell, D.C. (2002). Statistical methods for psychology, Wiley, New York.

Moore, D.S. (2004). The basic practice of statistics, W.H.Freeman, New York.

StatsDirect (2008) See website at **www.statsdirect.com**

Statistical Analysis 8: Two-way analysis of variance (ANOVA)

Research question type: Explaining a continuous variable with 2 categorical variables

What kind of variables? Continuous (**scale/interval/ratio**) and 2 independent categorical variables (**factors**)

Common Applications: Comparing means of a single variable at different **levels** of two conditions (**factors**) in scientific experiments.

Example:

The effective life (in hours) of batteries is compared by material type (1, 2 or 3) and operating temperature: Low (-10˚C), Medium (20˚C) or High (45˚C). Twelve batteries are randomly selected from each material type and are then randomly allocated to each temperature level. The resulting life of all 36 batteries is shown below:

Table 1: Life (in hours) of batteries by material type and temperature

		Temperature (˚C)		
		Low (-10˚C)	Medium (20˚C)	High (45˚C)
Material type	1	130, 155, 74, 180	34, 40, 80, 75	20, 70, 82, 58
	2	150, 188, 159, 126	136, 122, 106, 115	25, 70, 58, 45
	3	138, 110, 168, 160	174, 120, 150, 139	96, 104, 82, 60

Source: Montgomery (2001)

Research question: Is there difference in mean life of the batteries for differing material type and operating temperature levels?

In analysis of variance we compare the variability **between** the groups (how far apart are the means?) to the variability **within** the groups (how much natural variation is there in our measurements?). This is why it is called analysis of variance, abbreviated to **ANOVA**.

This example has two **factors** (material type and temperature), each with 3 **levels**.

Hypotheses:

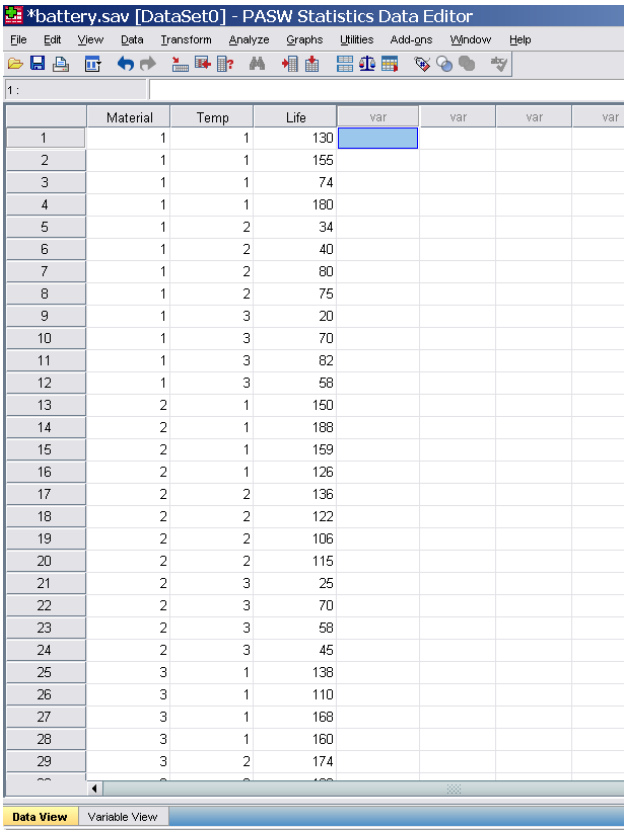
The 'null hypothesis' might be:

> H_0: There is **no difference in mean** battery life for different combinations of material type and temperature level

And an 'alternative hypothesis' might be:

> H_1: There **is a difference in mean** battery life for different combinations of material type and temperature level

If the alternative hypothesis is accepted, further analysis is performed to explore where the individual differences are.



Steps in SPSS (PASW):

Data need to be arranged in SPSS in a particular way to perform a two-way ANOVA. The dependent variable (battery life) values need to be in one column, and each factor needs a column containing a code to represent the different levels.

In this example *Material* has codes 1 to 3 for material type in the first column and *Temp* has codes 1 for Low, 2 for Medium and 3 for High operating temperatures.

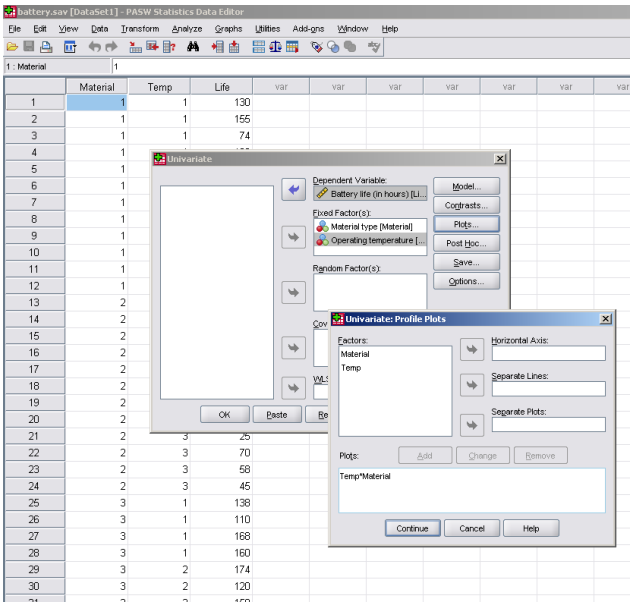
The battery life (*Life*) is entered in the third column – see screen to the left.

Note carefully how the data are entered.

The raw data file for this example is available on W:\EC\STUDENT\ MATHS SUPPORT CENTRE STATS WORKSHEETS\battery.sav

Then choose:
Analyze > General Linear Model > Two-Way ANOVA…

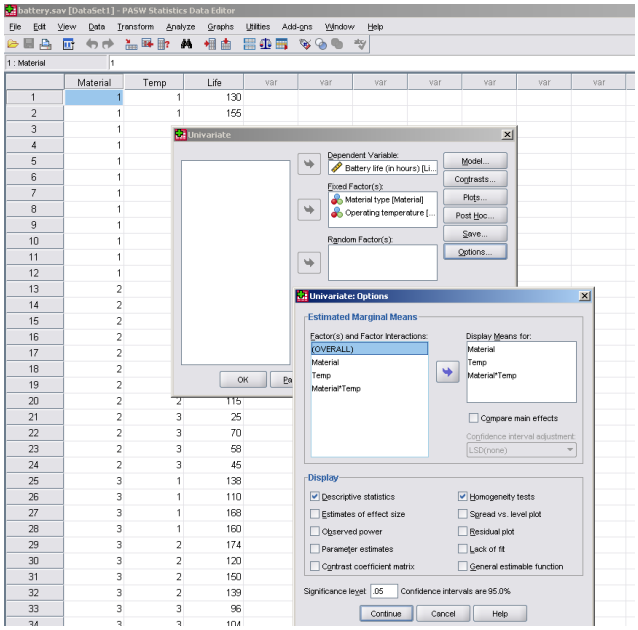
- Transfer the outcome variable (*Life* in this example) into the Dependent Variable box, and the factor variables (*Material* and *Temp* in this case) as the Fixed Factor(s)
- Click on **Model…** and select Full factorial to get the 'main effects' from each of the two factors and the 'interaction effect' of the two factors. [It is possible to build a Custom model, if you prefer]
- Continue



- Click on **Plots…**, and choose Temp for Horizontal Axis and Material in Separate Lines (see right)
- Click Add and Continue

- Click on **Post Hoc…** and select *Material* and *Temp*
- Check Tukey (or post hoc test of choice)
- Continue

- Click on **Options…** and choose to Display Means for *Material*, *Temp* and *Material*Temp*
- Check Descriptive statistics and Homogeneity tests (see right)
- Continue and OK



In the SPSS output there is a table showing the **descriptive statistics** for the main variable (battery life) at each of the levels for each factor (9 in this example), plus Totals.

Check the result of **Levene**'s test for a p-value (Sig.) > 0.05, so that similar variances for each group of measurements can be assumed (otherwise the ANOVA is probably invalid).

In the example, p = 0.529, so the two-way ANOVA can proceed.

The **Tests of Between Subjects Effects** table gives the results of the ANOVA. Table 2 below shows the output for the battery example with the important numbers emboldened.

Table 2: Tests of Between-Subjects Effects

Dependent Variable: Battery life (in hours)

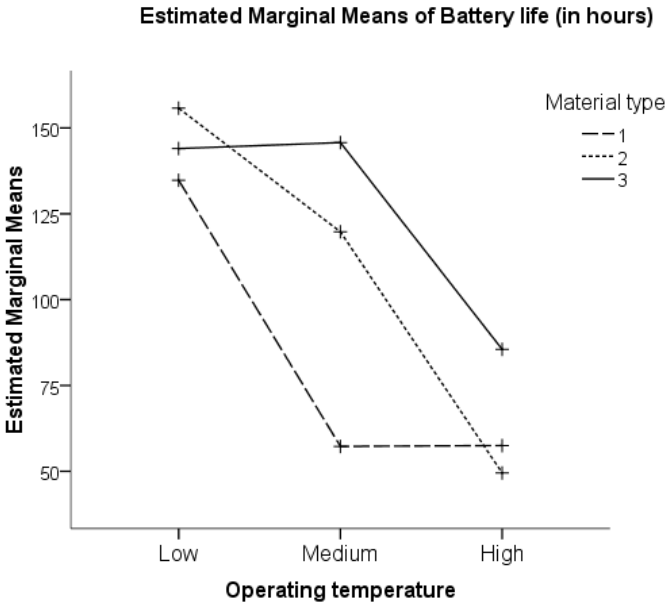
Source	Type III Sum of Squares	df	Mean Square	F	**Sig.**
Corrected Model	59416.222[a]	8	7427.028	11.000	.000
Intercept	400900.028	1	400900.028	593.739	.000
Material	10683.722	2	5341.861	**7.911**	**.002**
Temp	39118.722	2	19559.361	**28.968**	**.000**
Material * Temp	9613.778	4	2403.444	**3.560**	**.019**
Error	18230.750	27	675.213		
Total	478547.000	36			
Corrected Total	77646.972	35			

***p*-values**

Results:

From the Descriptive Statistics table, it can be seen that, overall, battery life decreases with higher operating temperature, although battery life remains high for material 3 at medium temperature. This pattern is more obvious when looking at the **plot** – see right.

Since the lines representing the three materials in the plot are not parallel, this implies there is an interaction effect between material and operating temperature. [The lines would be approximately parallel if there were no interaction.] So, how battery life changes with temperature depends on the material, and vice versa.



The ANOVA table gives F statistics = 7.91, p=0.002; 28.97, p<0.001 and 3.56, p=0.019, for material, operating temperature and material*temperature, respectively [**NEVER write p = 0.000**]. So**, both** material and temperature are needed, as well as their **interaction**, to explain battery life.

The nature of these differences can be explored further by looking at the SPSS output from the '**post hoc**' tests. These suggest that mean battery life overall is statistically significantly longer for material 3 than 1 (p=0.001), and at lower compared to higher temperature levels. However, at low operating temperatures material 2 appeared to give a longer life than for materials 2 and 3, but lasted least at high temperatures.

Conclusion:
From the results it can be said that there is strong evidence that the mean battery life varies with material used and operating temperature (p=0.002 and p<0.001). The presence of interaction between material and temperature means that the way battery life changes for different materials depends on the temperature. Similarly, the way battery life changes for different temperatures depends on material. Overall, material 3 performs best.

The three tables of estimated marginal means give details of mean battery lives by factor, plus 95% CIs, giving more detail concerning the accuracy of these battery life estimates from the sample to the overall population. In this example, the CIs are all fairly 'wide', so results should be used with caution.

Validity of two-way ANOVA:

ANOVA is based on two assumptions:

- the observations are random samples from normal distributions
- the populations have the same variance [variance = (standard deviation)^2]
- observations are independent of each other

So, before carrying out any tests the data must be examined in more detail to determine whether these assumptions are satisfied. See one-way ANOVA sheet for more information relating to this aspect.

Comments:

- Multiple t-tests should not be performed
- It is possible to perform two-way ANOVA with different sample sizes per group. Select Type IV Sum of squares in the Univariate: Model dialog box.

Montgomery DC (2001) *Design and Analysis of Experiments* (5th Ed.) John Wiley & Sons Inc., New York

Mathematics Learning Support Centre

Statistics: 2.3 The Mann-Whitney U Test

Rosie Shier. 2004.

1 Introduction

The Mann-Whitney U test is a non-parametric test that can be used in place of an unpaired t-test. It is used to test the null hypothesis that two samples come from the same population (i.e. have the same median) or, alternatively, whether observations in one sample tend to be larger than observations in the other. Although it is a non-parametric test it does assume that the two distributions are similar in shape.

2 Carrying out the Mann-Whitney U test

Suppose we have a sample of n_x observations $\{x_1, x_2, \ldots x_n\}$ in one group (i.e. from one population) and a sample of n_y observations $\{y_1, y_2, \ldots y_n\}$ in another group (i.e. from another population).

The Mann-Whitney test is based on a comparison of every observation x_i in the first sample with every observation y_j in the other sample. The total number of pairwise comparisons that can be made is $n_x n_y$.

If the samples have the same median then each x_i has an equal chance (i.e. probability $\frac{1}{2}$) of being greater or smaller than each y_j.

So, under the null hypothesis	$H_0 : P(x_i > y_j) = \frac{1}{2}$
and under the alternative hypothesis	$H_1 : P(x_i > y_j) \neq \frac{1}{2}$

We count the number of times an x_i from sample 1 is greater than a y_j from sample 2. This number is denoted by U_x. Similarly, the number of times an x_i from sample 1 is smaller than a y_j from sample 2 is denoted by U_y. Under the null hypothesis we would expect U_x and U_y to be approximately equal.

Procedure for carrying out the test:

1. Arrange all the observations in order of magnitude.
2. Under each observation, write down X or Y (or some other relevant symbol) to indicate which sample they are from.
3. Under each x write down the number of ys which are to the left of it (i.e. smaller than it); this indicates $x_i > y_j$. Under each y write down the number of xs which are to the left of it (i.e. smaller than it); this indicates $y_j > x_i$

4. Add up the total number of times $x_i > y_j$ — denote by U_x. Add up the total number of times $y_j > x_i$ — denote by U_y. Check that $U_x + U_y = n_x n_y$.

5. Calculate $U = min(U_x, U_y)$

6. Use statistical tables for the Mann-Whitney U test to find the probability of observing a value of U or lower. If the test is one-sided, this is your p-value; if the test is a two-sided test, double this probabililty to obtain the p-value.

NOTE: If the number of observations is such that $n_x n_y$ is large enough (> 20), a normal approximation can be used with $\mu_U = \frac{n_x n_y}{2}, \sigma_U = \sqrt{\frac{n_x n_y (N+1)}{12}}$, where $N = n_x + n_y$.

Dealing with ties: It is possible that two or more observations nay be the same. If this is the case we can still calculate U by allocating half the tie to the X value and half the tie to the Y value. However, if this is the case then the normal approximation must be used with an adjustment to the standard deviation. This becomes:

$$\sigma_U = \sqrt{\frac{n_x n_y}{N(N-1)} \times \left[\frac{N^3 - N}{12} - \sum_{j=1}^{g} \frac{t_j^3 - t_j}{12}\right]}$$

where $N = n_x + n_y$
$g =$ the number of groups of ties
$t_j =$ the number of tied ranks in group j

Note that the Mann-Whitney U test is statistically equivalent to the Wilcoxon rank sum test (not to be confused with the Wilcoxon **signed** rank sum test, which is for paired data).

Example:
The following data shows the age at diagnosis of type II diabetes in young adults. Is the age at diagnosis different for males and females?

Males:	19 22 16 29 24
Females:	20 11 17 12

Solution:

1. Arrange in order of magnitude

Age	11	12	16	17	19	20	22	24	29
M/F	F	F	M	F	M	F	M	M	M
$M > F$			2		3		4	4	4
$F > M$	0	0		1		2			

2. Affix M or F to each observation (see above).

3. Under each M write the number of Fs to the left of it; under each F write the number of Ms to the left of it (see above).

4. $U_M = 2 + 3 + 4 + 4 + 4 = 17 \qquad U_F = 0 + 0 + 1 + 2 = 3$

5. $U = min(U_M, U_F) = 3$

6. Using tables for the Mann-Whitney U test we get a two-sided p-value of $p = 0.11$

7. If we use a normal approximation we get:
$$z = \frac{U - \frac{n_x n_y}{2}}{\sqrt{\frac{n_x n_y (N+1)}{12}}} = \frac{3-10}{\sqrt{50/3}} = -1.715$$ This gives a two-sided p-value of $p = 0.09$.

The exact test and the normal approximation give similar results. We would conclude that there is no real evidence that the age at diagnosis is different for males and females, although the results are borderline and the lack of statistical significance in this case may just be due to the very small sample. The actual median age at diagnosis is 14.5 years for females and 22 for males, which is quite a substantial difference. In this case it would be advisable to conduct a larger study.

3 Carrying out the Mann-Whitney U test in SPSS

— Choose **Analyze**
— Select **Nonparametric Tests**
— Select **2 Independent Samples**
— Highlight your test variable (in our example this would be age) and click on the arrow to move this into the **Test Variable List** box
— Highlight the grouping variable and click on the arrow to move this into the **Grouping Variable** box.
— Click on **Define Groups** and type in the codes that indicate which group an observation belongs to (in our example, the codes which indicate whether a subject is male or female). Click on **Continue**
— Under **Test Type** make sure that **Mann-Whitney U** is selected
— If you want exact probabilities, click on **Exact**, choose **Exact**, then **Continue**
— Click on **OK**

The output will look like this:

Ranks

	Sex	N	Mean Rank	Sum of Ranks
Age	Male	5	6.40	32
	Female	4	3.25	13

Test Statistics

	Age
Mann-Whitney U	3.000
Wilcoxon W	13.000
z	-1.715
Asymp. Sig. (2-tailed)	0.086
Exact Sig. [2*(1-tailed Sig.)]	0.111
Exact Sig. (2-tailed)	0.111
Exact Sig. (1-tailed)	0.056
Point Probability	0.024

We are interested in the exact p-value ("Exact Sig (2-tailed)") and the p-value based on the normal approximation ("Asymp Sig (2-tailed"). If the normal approximation is appropriate then these should be roughly similar.

Pearson's correlation

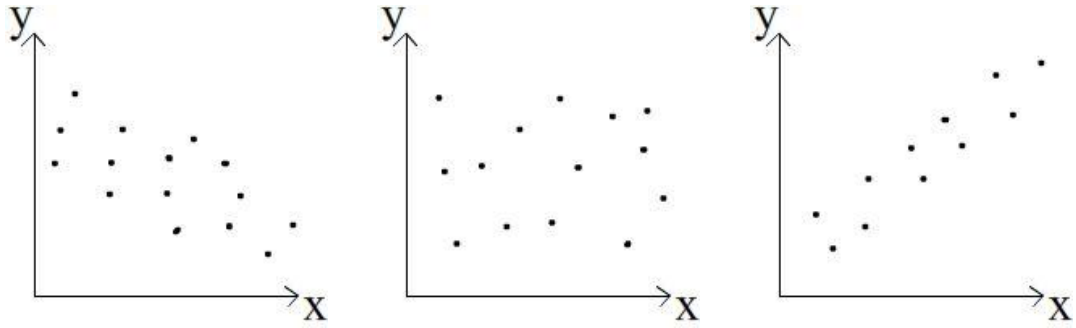
Introduction

Often several quantitative variables are measured on each member of a sample. If we consider a pair of such variables, it is frequently of interest to establish if there is a relationship between the two; i.e. to see if they are *correlated*.

We can categorise the type of correlation by considering as one variable increases what happens to the other variable:

- *Positive correlation* – the other variable has a tendency to also increase;
- *Negative correlation* – the other variable has a tendency to decrease;
- *No correlatio*n – the other variable does not tend to either increase or decrease.

The starting point of any such analysis should thus be the construction and subsequent examination of a *scatterplot*. Examples of negative, no and positive correlation are as follows.

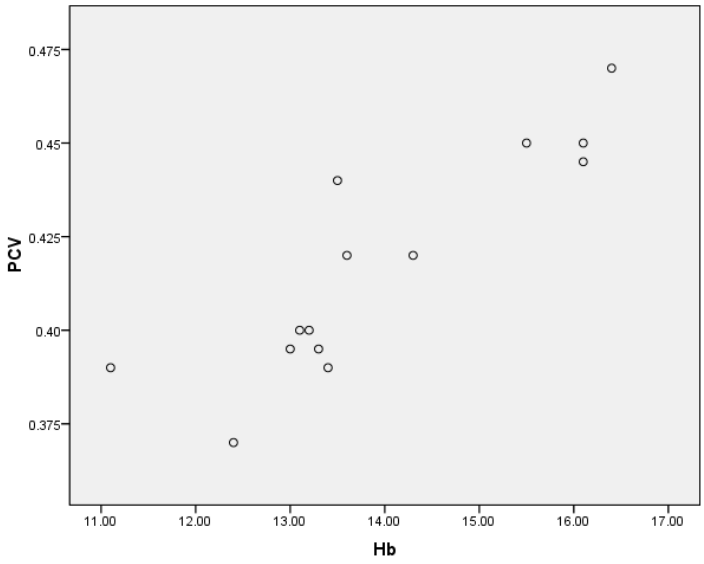


Negative correlation **No correlation** **Positive correlation**

Example

Let us now consider a specific example. The following data concerns the blood haemoglobin (Hb) levels and packed cell volumes (PCV) of 14 female blood bank donors. It is of interest to know if there is a relationship between the two variables Hb and PCV when considered in the female population.

Hb	PCV
15.5	0.450
13.6	0.420
13.5	0.440
13.0	0.395
13.3	0.395
12.4	0.370
11.1	0.390
13.1	0.400
16.1	0.445
16.4	0.470
13.4	0.390
13.2	0.400
14.3	0.420
16.1	0.450



The scatterplot suggests a definite relationship between PVC and Hb, with larger values of Hb tending to be associated with larger values of PCV.

There appears to be a positive correlation between the two variables.

We also note that there appears to be a *linear* relationship between the two variables.

Correlation coefficient

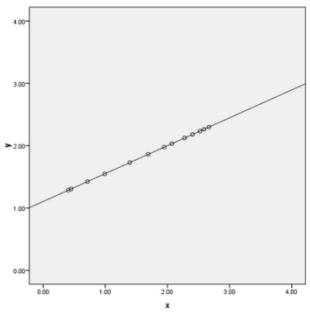
Pearson's correlation coefficient is a statistical measure of the strength of a *linear* relationship between paired data. In a sample it is denoted by r and is by design constrained as follows

$$-1 \leq r \leq 1$$

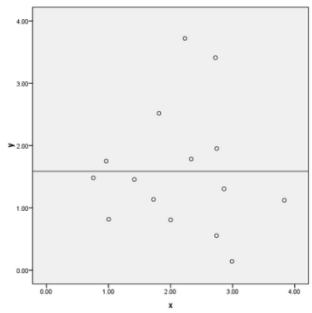
Furthermore:

- Positive values denote positive linear correlation;
- Negative values denote negative linear correlation;
- A value of 0 denotes no linear correlation;
- The closer the value is to 1 or –1, the stronger the linear correlation.

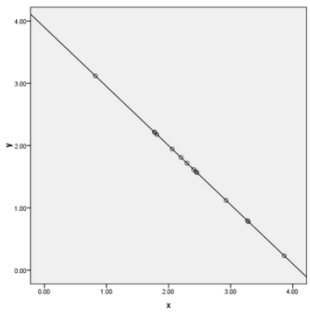
In the figures various samples and their corresponding sample correlation coefficient values are presented. The first three represent the "extreme" correlation values of -1, 0 and 1:



$r = -1$
perfect -ve correlation



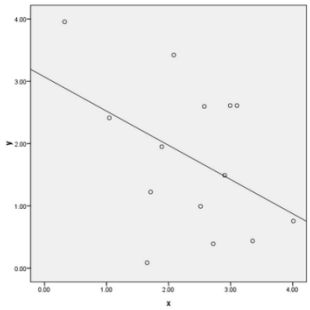
$r = 0$
no correlation



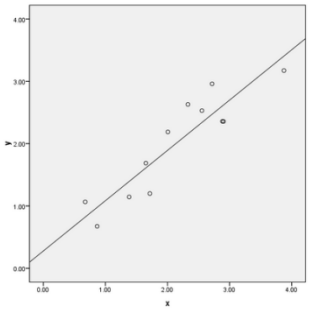
$r = 1$
perfect +ve correlation

When $r = \pm 1$ we say we have *perfect* correlation with the points being in a perfect straight line.

Invariably what we observe in a sample are values as follows:



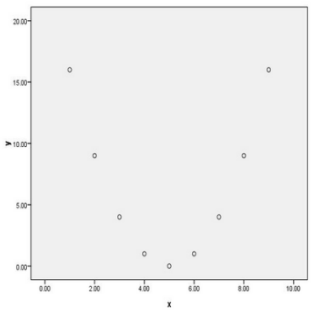
$r = -.45$
moderate -ve correlation



$r = .92$
very strong +ve correlation

Note:

1) the correlation coefficient does not relate to the gradient beyond sharing its +ve or –ve sign!
2) The correlation coefficient is a measure of linear relationship and thus a value of $r = 0$ does not imply there is no relationship between the variables. For example in the following scatterplot $r = 0$ which implies no (linear) correlation however there is a perfect quadratic relationship:



$r = 0$
perfect quadratic relationship

Correlation is an effect size and so we can verbally describe the strength of the correlation using the guide that Evans (1996) suggests for the absolute value of *r*:

- .00-.19 "very weak"
- .20-.39 "weak"
- .40-.59 "moderate"
- .60-.79 "strong"
- .80-1.0 "very strong"

For example a correlation value of $r = .42$ would be a "moderate positive correlation".

Assumptions

The calculation of Pearson's correlation coefficient and subsequent significance testing of it requires the following data assumptions to hold:

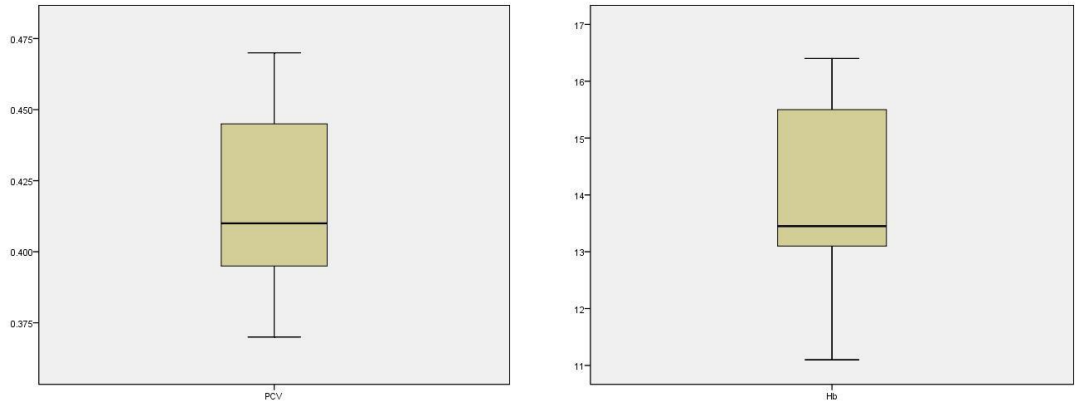
- interval or ratio level;
- linearly related;
- bivariate normally distributed.

In practice the last assumption is checked by requiring both variables to be individually normally distributed (which is a by-product consequence of bivariate normality). Pragmatically Pearson's correlation coefficient is sensitive to skewed distributions and outliers, thus if we do not have these conditions we are content.

If your data does not meet the above assumptions then use Spearman's rank correlation!

Example (revisited)

We have no concerns over the first two data assumptions, but we need to check the normality of our variables. One simple way of doing is to examine boxplots of the data. These are given below.



The boxplot for PCV is fairly consistent with one from a normal distribution; the median is fairly close to the centre of the box and the whiskers are of approximate equal length.

The boxplot for Hb is slightly disturbing in that the median is close to the lower quartile which would be suggesting positive skewness. Although countering this is the argument that with positively skewed data the lower whisker should be shorter than the upper whisker; this is not the case here.

Since we have some doubts over normality, we shall examine the skewness coefficients to see if they suggest whether either of the variables is skewed.

Descriptive Statistics

	N	Skewness	
	Statistic	Statistic	Std. Error
Hb	14	.262	.597
PCV	14	.299	.597
Valid N (listwise)	14		

Both have skewness coefficients that are indeed positive, but a quick check to see if these are not sufficiently large to warrant concern is to see if the absolute values of the skewness coefficients are less than two times their standard errors. In both cases they are which is consistent with the data being normal. Hence we do not have any concerns over the normality of our data and can continue with the correlation analysis.

For the Haemoglobin/PCV data, SPSS produces the following correlation output:

Correlations

		Hb	PCV
Hb	Pearson Correlation	1	.877**
	Sig. (2-tailed)		.000
	N	14	14
PCV	Pearson Correlation	.877**	1
	Sig. (2-tailed)	.000	
	N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

The Pearson correlation coefficient value of 0.877 confirms what was apparent from the graph, i.e. there appears to be a positive correlation between the two variables.

However, we need to perform a significance test to decide whether based upon this sample there is any or no evidence to suggest that linear correlation is present in the population.

To do this we test the null hypothesis, H_0, that there is no correlation in the population against the alternative hypothesis, H_1, that there is correlation; our data will indicate which of these opposing hypotheses is most likely to be true. We can thus express this test as:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

i.e. the null hypothesis of no linear correlation present in population against the alternative that there is linear correlation present.

SPSS reports the p-value for this test as being .000 and thus we can say that we have very strong evidence to believe H_1, i.e. we have some evidence to believe that Hb and PCV are linearly correlated in the female population.

The significant Pearson correlation coefficient value of 0.877 confirms what was apparent from the graph; there appears to be a very strong positive correlation between the two variables. Thus large values of Hb are associated with large PCV values.

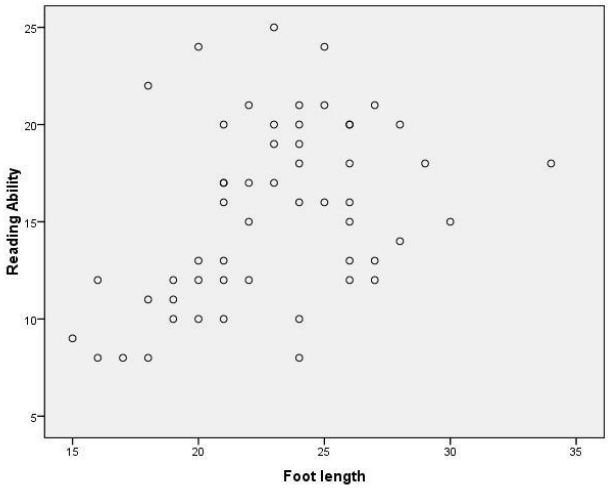
This could be formally reported as follows:

"A Pearson's correlation was run to determine the relationship between 14 females' Hb and PCV values. There was a very strong, positive correlation between Hb and PCV ($r = .88$, $N=14$, $p < .001$)."

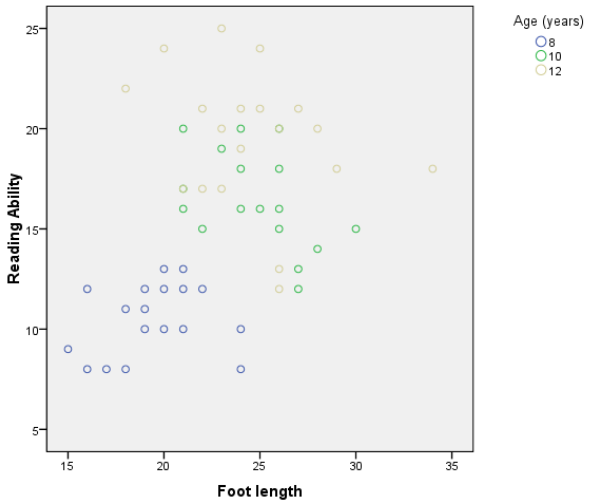
Caution

The existence of a strong correlation does not imply a causal link between the variables. For example we can not imply that Hb causes PCV or vice versa.

Also you should be aware of the possibility of hidden or intervening variables. For instance suppose we consider the relationship between reading ability and foot length for children. A scatter plot and correlation analysis of the data indicates that there is a very strong correlation between reading ability and foot length (r = .88, N=54, p =.003):

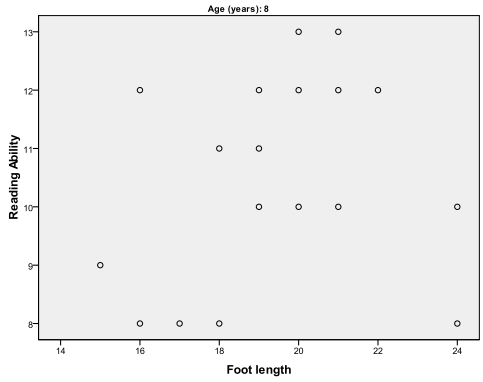


However, if we consider taking into account the children's age, we can see that this apparent correlation may be spurious.



If we now reanalyse the data by age group we indeed find that in each case there appears to be no correlation between the two variables:

Age (years) = 8

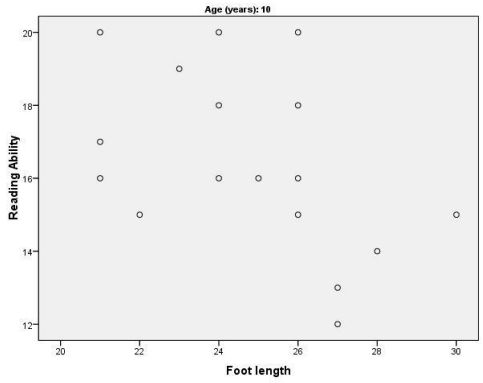


Correlations[a]

		Reading Ability	Foot length
Reading Ability	Pearson Correlation	1	.210
	Sig. (2-tailed)		.403
	N	18	18
Foot length	Pearson Correlation	.210	1
	Sig. (2-tailed)	.403	
	N	18	18

a. Age (years) = 8

Age (years) = 10

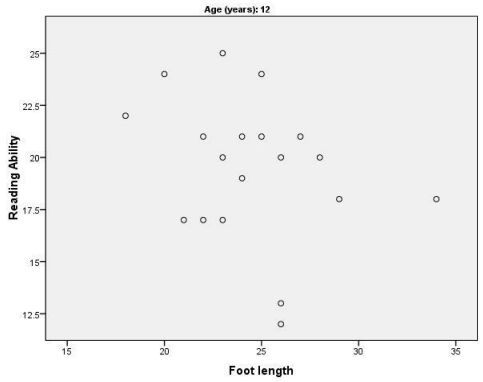


Correlations[a]

		Reading Ability	Foot length
Reading Ability	Pearson Correlation	1	-.465
	Sig. (2-tailed)		.060
	N	17	17
Foot length	Pearson Correlation	-.465	1
	Sig. (2-tailed)	.060	
	N	17	17

a. Age (years) = 10

Age (years) = 12



Correlations[a]

		Reading Ability	Foot length
Reading Ability	Pearson Correlation	1	-.290
	Sig. (2-tailed)		.228
	N	19	19
Foot length	Pearson Correlation	-.290	1
	Sig. (2-tailed)	.228	
	N	19	19

a. Age (years) = 12

Mathematics Learning Support Centre

Statistics: 3.2 Principal Components Analysis

Rosie Cornish. 2007.

1 Introduction

This handout is designed to provide only a brief introduction to principal components analysis and how it is done. Books giving further details are listed at the end.

Principal components analysis is a multivariate method used for data reduction purposes. The basic idea is to represent a set of variables by a smaller number of variables called **principal components**. These are chosen in such a way that they are uncorrelated (and are therefore measuring different, unrelated aspects, or dimensions, of the data).

2 Assumptions

Principal components analysis, like factor analysis, is designed for interval data, although it can also be used for ordinal data (e.g. scores assigned to Likert scales). The variables should be linearly related to each other. This can be checked by looking at scatterplots of pairs of variables. Obviously the variables must also be at least moderately correlated to each other, otherwise the number of principal components will be almost the same as the number of original variables, which means that carrying out a principal components analysis would be pointless.

3 What principal components analysis does

If you have p variables $X_1, X_2, \ldots, X_p$ measured on a sample of n subjects, then the i^{th} principal component, Z_i can be written as a linear combination of the original variables. Thus,

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$$

The principal components are chosen such that the first one, $Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$ accounts for as much of the variation in the data (i.e. in the original variables) as possible subject to the constraint that

$$a_{11}^2 + a_{12}^2 + \ldots + a_{1p}^2 = 1$$

Then the second principal component $Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$ is chosen such that its variance is as high as possible. A similar constraint applies — namely, that

$$a_{21}^2 + a_{22}^2 + \ldots + a_{2p}^2 = 1$$

Another constraint is that the second component is chosen such that it is uncorrelated with the first component. The remaining principal components are chosen in the same way.

Simple linear regression

Introduction

Simple linear regression is a statistical method for obtaining a formula to predict values of one variable from another where there is a causal relationship between the two variables.

Straight line formula

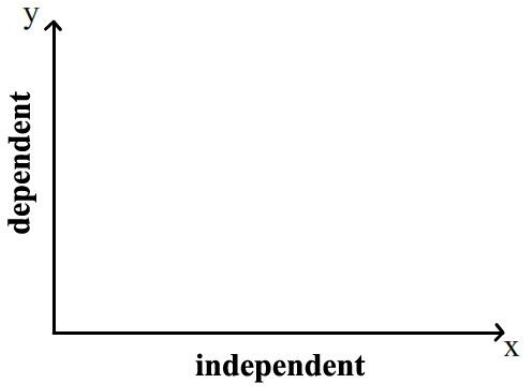
Central to simple linear regression is the formula for a straight line that is most commonly represented as $y = mx + c$ or $y = a + bx$. Statisticians however generally prefer to use the following form involving betas:

$$y = \beta_0 + \beta_1 x$$

The variables y and x are those whose relationship we are studying. We give them the following names:

- y: dependent (or response) variable;
- x: independent (or predictor or explanatory) variable.

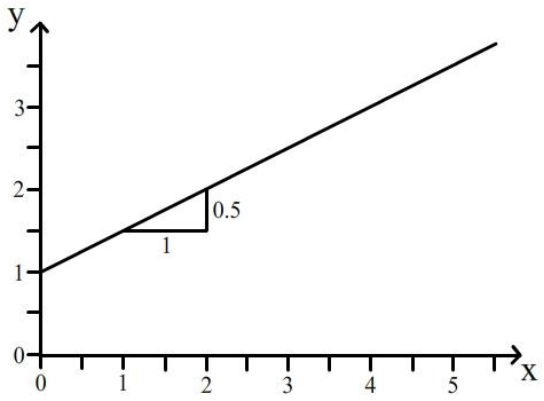
It is convention when plotting data to put the dependent and independent data on the y and x axis respectively;



β_0 and β_1 are constants and are parameters (or coefficients) that need to be estimated from data. Their roles in the straight line formula are as follows:

- β_0 : intercept;
- β_1 : gradient.

For instance the line $y = 1 + 0.5x$ has an intercept of 1 and a gradient of 0.5. Its graph is as follows:



Model assumptions

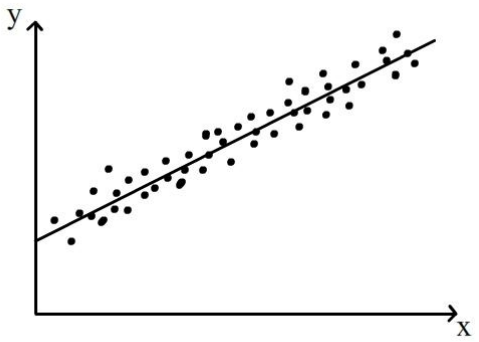
In simple linear regression we aim to predict the response for the *i*th individual, Y_i, using the individual's score of a single predictor variable, X_i. The form of the model is given by:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

which comprises a deterministic component involving the two *regression coefficients* (β_0 and β_1) and a random component involving the *residual* (error) term (ε_i).

The deterministic component is in the form of a straight line which provides the predicted (mean/expected) response for a given predictor variable value.

The residual terms represent the difference between the predicted value and the observed value of an individual. They are assumed to be independently and identically distributed normally with zero mean and variance σ^2, and account for natural variability as well as maybe measurement error. Our data should thus appear to be a collection of points that are randomly scattered around a straight line with constant variability along the line:



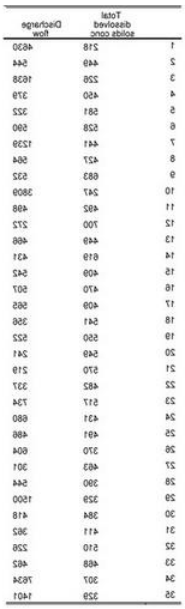
The deterministic component is a linear function of the unknown regression coefficients which need to be estimated so that the model 'best' describes the data. This is achieved mathematically by minimising the sum of the squared residual terms (*least squares*). The fitting also produces an estimate of the error variance which is necessary for things like significance test regarding the regression coefficients and for producing confidence/prediction intervals.

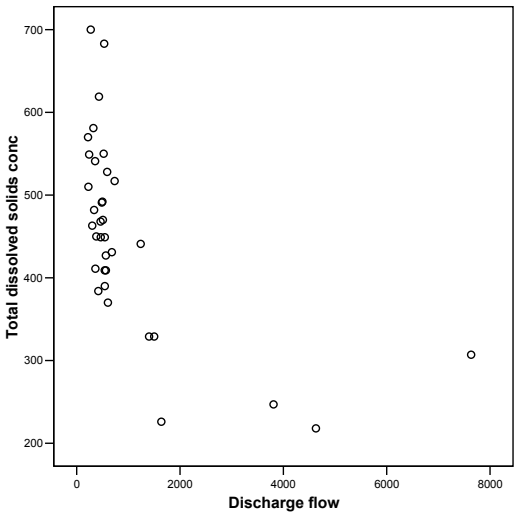
Example

Suppose we are interested in predicting the total dissolved solids (TDS) concentrations (mg/L) in a particular river as a function of the discharge flow (m^3/s). We have collected data that comprise a sample of 35 observations that were collected over the previous year.

The first step is to look carefully at the data:

- Is there an upwards/downwards trend in the data or could a horizontal line be fit though the data?
- Is the trend linear or curvilinear?
- Is there constant variance along the regression line or does it systematically change as the predictor variable changes?





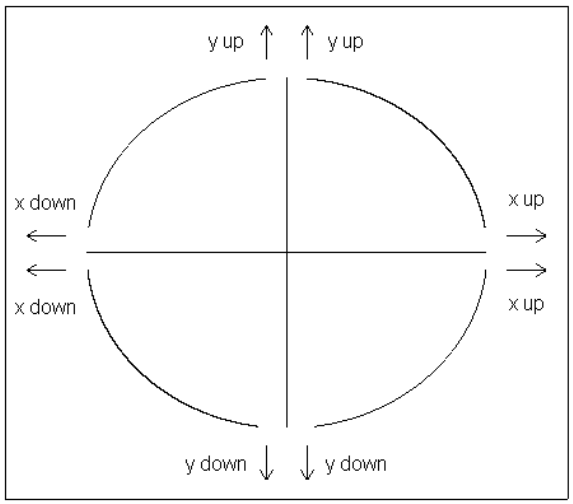
The scatterplot above suggests that there is a downwards trend in the data, however there is a curvilinear relationship. The variance about a hypothetical curve appears fairly constant.

Transformations

Simple linear regression is appropriate for modelling linear trends where the data is uniformly spread around the line. If this is not the case then we should be using other modelling techniques and/or transforming our data to meet the requirements. When considering transformations the following is a guide:

- If the trend is curvilinear consider a transformation of the predictor variable, *x*.
- If constant variance is a problem (and maybe curvilinear as well) consider either a transformation of the response variable, *y*, or a transformation of both the response and the predictor variable, *x* and *y*.

Tukey's "bulging rule" can act as a guide to selecting power transformations.



Compare your data to the above and if it has the shape in any of the quadrants then consider the transformations where:

- up – use powers of the variable greater than 1 (e.g. x^2, etc);
- down - powers of the variable less than 1 (e.g. log(x), 1/x, $\sqrt{x}$ etc).

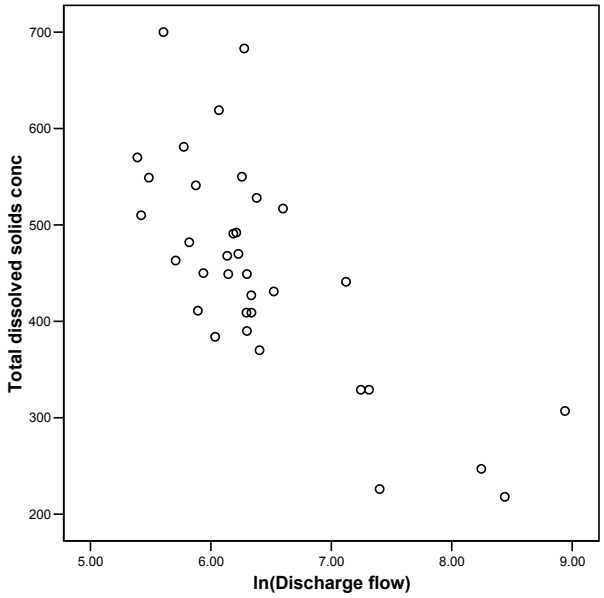
Note, sometimes a second application of Tukey's bulging rule is necessary to gain linearity with constant variability.

Example (revisited)

Returning to our example, the scatterplot reveals the data to belong to the bottom left quadrant of Tukey's bulging rule. Since the variance about a hypothetical curve appears fairly constant, thus we shall try transforming just the predictor variable.
Tukey's bulging rule suggests a "down" power; we shall try the log natural transformation first

The resulting scatterplot of TDS against ln(Discharge) is now far more satisfactory:

	Total dissolved solids conc (mg/L)	Discharge flow (cu m/s)	Ln Discharge
1	218	4630	8.44
2	449	544	6.30
3	226	1638	7.40
4	450	379	5.94
5	581	322	5.77
6	528	590	6.38
7	441	1239	7.12
8	427	564	6.34
9	683	532	6.28
10	247	3809	8.25
11	492	498	6.21
12	700	272	5.61
13	449	466	6.14
14	619	431	6.07
15	409	542	6.30
16	470	507	6.23
17	409	565	6.34
18	541	356	5.87
19	550	522	6.26
20	549	241	5.48
21	570	219	5.39
22	482	337	5.82
23	517	734	6.60
24	431	680	6.52
25	491	486	6.19
26	370	604	6.40
27	463	301	5.71
28	390	544	6.30
29	329	1500	7.31
30	384	418	6.04
31	411	362	5.89
32	510	226	5.42
33	468	462	6.14
34	307	7634	8.94
35	329	1401	7.24



The data now appears to be suitable for simple linear regression and we shall now consider selected output from the statistics package SPSS.

Correlations

		Total dissolved solids conc	ln(Discharge flow)
Pearson Correlation	Total dissolved solids conc	1.000	-.735
	ln(Discharge flow)	-.735	1.000
Sig. (1-tailed)	Total dissolved solids conc	.	.000
	ln(Discharge flow)	.000	.
N	Total dissolved solids conc	35	35
	ln(Discharge flow)	35	35

The correlations table displays Pearson correlation coefficients, significance values, and the number of cases with non-missing values. As expected we see that we have a strong negative correlation (-.735) between the two variables. From the significance test p-value we can see that we have very strong evidence ($p<0.001$) to suggest that there is a linear correlation between the two variables.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.735[a]	.540	.526	78.261

a. Predictors: (Constant), ln(Discharge flow)

b. Dependent Variable: Total dissolved solids conc

The model summary table displays:

- R, the multiple correlation coefficient, is a measure of the strength of the linear relationship between the response variable and the set of explanatory variables. It is the highest possible simple correlation between the response variable and any linear combination of the explanatory variables. For simple linear regression where we have just two variables, this is the same as the absolute value of the Pearson's correlation coefficient we have already seen above. However, in multiple regression this allows us to measure the correlation involving the response variable and more than one explanatory variable.
- R squared is the proportion of variation in the response variable explained by the regression model. The values of R squared range from 0 to 1; small values indicate that the model does not fit the data well. From the above we can see that the model fits the data reasonably well; 54% of the variation in the *TDS* values can be explained by the fitted line together with the *lnDischarge* values. R squared is also known as the *coefficient of determination*.
- The R squared value can be over optimistic in its estimate of how well a model fits the population; the adjusted R square value is attempts to correct for this. Here we can see it has slightly reduced the estimated proportion. If you have a small data set it may be worth reporting the adjusted R squared value.
- The standard error of the estimate is the estimate of the standard deviation of the error term of the model, σ. This gives us an idea of the expected variability of predictions and is used in calculation of confidence intervals and significance tests.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1103.967	105.320		10.482	.000	889.693	1318.242
	ln(Discharge flow)	-101.275	16.281	-.735	-6.221	.000	-134.399	-68.152

a. Dependent Variable: Total dissolved solids conc

The unstandardized coefficients are the coefficients of the estimated regression model. Thus the expected *TDS* value is given by:

$$TDS = 1103.967 - 101.275\ln(Discharge).$$

Thus we can see that for each one unit increase in ln(*Discharge*), the *TDS* value is expected to decrease by 101.275 units. The intercept for this example could be interpreted as the *TDS* value (1103.967) when the ln(*Discharge*) flow is zero (i.e. Discharge = 1 m^3/s.

The standardized coefficients are appropriate in multiple regression when we have explanatory variables that are measured on different units. These coefficients are obtained from regression after the explanatory variables are all standardized. The idea is that the coefficients of explanatory variables can be more easily compared with each other as they are then on the same scale. In simple linear regression they are of little concern.

The standard errors give us estimates of the variability of the (unstandardised) coefficients and are used for significance tests for the coefficients and for the displayed 95% confidence intervals. The *t* values and corresponding significance vales are tests assessing the worth of the (unstandardised) coefficients. It is usually of importance to be assessing the worth of our predictor variable and hence evaluating the significance of the coefficient β_1 in our model formulation. That is we are assessing for evidence of a significant non-zero slope. If the coefficient is not significantly different to zero then this implies the predictor variable does not influence our response variable.

Here we have both test are highly significant ($p<0.001$), indicating that we have very strong evidence of need both the coefficients in our model. The resulting confidence intervals expand our understanding of the problem. For example, with 95% confidence we believe that the interval between -134.399 and -68.152 covers the true unknown *TDS* value change per ln(*Discharge*) unit.

The remaining output is concerned with checking the model assumptions of normality, linearity, homoscedasticity and independence of the residuals. Residuals are the differences between the observed and predicted responses. The residual scatterplots allow you to check:

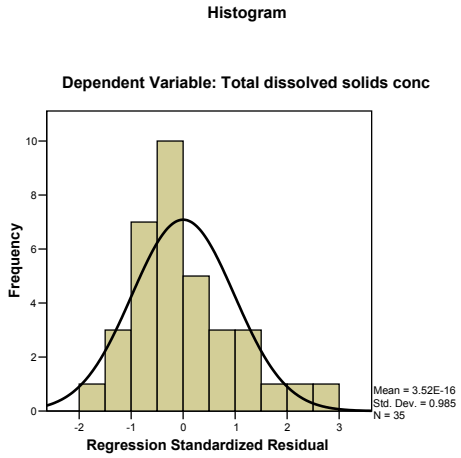
- *Normality*: the residuals should be normally distributed about the predicted responses;
- *Linearity*: the residuals should have a straight line relationship with the predicted responses;
- *Homoscedasticity*: the variance of the residuals about predicted responses should be the same for all predicted responses.

Residuals Statistics[a]

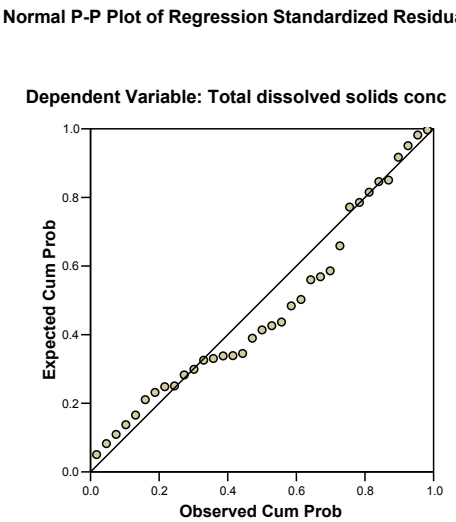
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	198.53	558.19	454.00	83.491	35
Residual	-128.404	214.702	.000	77.101	35
Std. Predicted Value	-3.060	1.248	.000	1.000	35
Std. Residual	-1.641	2.743	.000	.985	35

a. Dependent Variable: Total dissolved solids conc

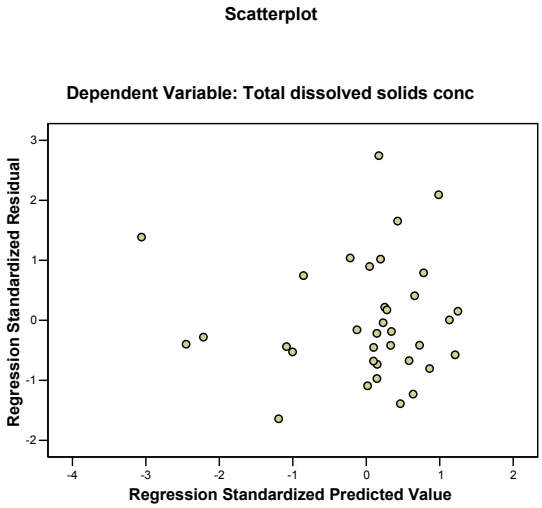
The above table summarises the predicted values and residuals in unstandarised and standardised forms. It is usual practice to consider standardised residuals due to their ease of interpretation. For instance outliers (observations that do not appear to fit the model that well) can be identified as those observations with standardised residual values above 3.3 (or less than -3.3). From the above we can see that we do not appear to have any outliers.



The above plot is a check on normality; the histogram should appear normal; a fitted normal distribution aids us in our consideration. Serious departures would suggest that normality assumption is not met. Here we have a slight suggestion of positive skewness but considering we have only 35 data points we have no real cause for concern.



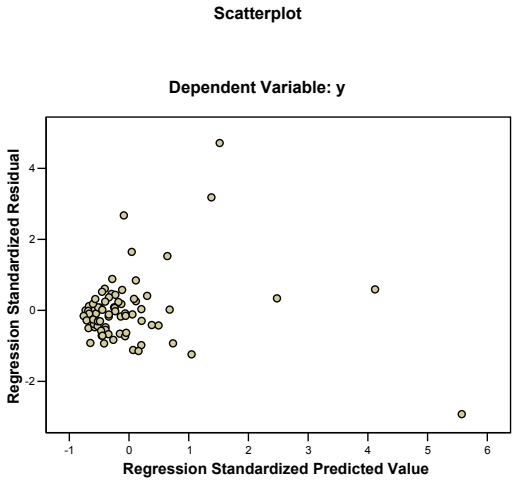
The above plot is a check on normality; the plotted points should follow the straight line. Serious departures would suggest that normality assumption is not met. Here we have no major cause for concern.



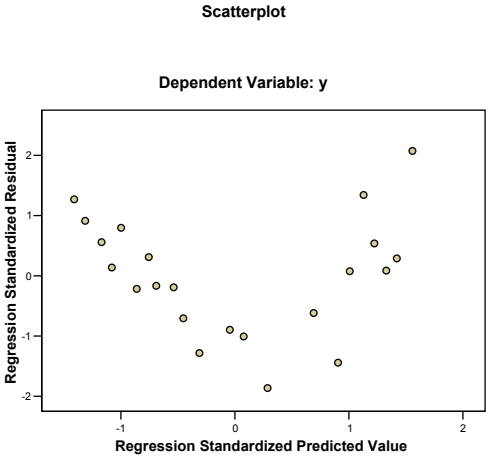
The above scatterplot of standardised residuals against predicted values should be a random pattern centred around the line of zero standard residual value. The points should have the same dispersion about this line over the predicted value range. From the above we can see no clear relationship between the residuals and the predicted values which is consistent with the assumption of linearity. The dispersion of residuals over the predicted value range between -1 and 1 looks constant, for predicted values below -1 there is too few points to provide evidence against a change in variability.

Model violations

So what do residual scatterplots of models that violate the model look like? Here are two common examples together with suggested remedies for the next regression to try.



In the plot above there is clear evidence of heteroscedasticty; change of variance with predicted value. Try log natural or square root transformation of y to stabilise variance.



In the plot above there is a clear curved pattern in the residuals. Try transforming x to obtain a linear relationship between it and the response variable.

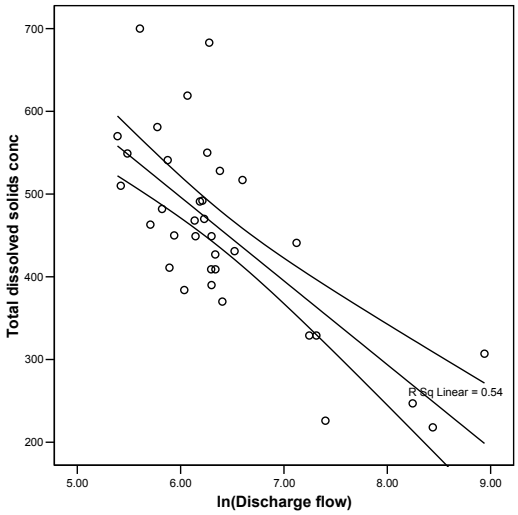
Example (revisited)

In order to get *TDS* predictions for particular *Discharge* values we can use the fitted line, say for a Discharge of 2000 m^3/s:

$$\begin{aligned} TDS &= 1103.967 - 101.275\ln(2000) \\ &= 334.186 \end{aligned}$$

Alternatively, we could let a statistics like SPSS to do the work and calculate confidence or prediction intervals at the same time. We shall now consider some of the other output that SPSS gives us.

The following gives the fitted line together with 95% confidence interval for the expected *TDS* response.



When requesting a predicted value we can also obtain the following:

- the predicted values for the various Discharges together with the associated standard errors of the predictions;
- 95% CI for the expected response;
- 95% CI for individual predicted responses;

For example for a *Discharge* of 2000 m^3/s:

- the expected TDS is 334.18 mg/L (s.e. = 23.366);
- we are 95% certain that interval from 286.64 to 381.72 mg/L covers the unknown expected TDS value;
- we are 95% certain that interval from 168.01 to 500.35 mg/L covers the range of predicted individual TDS observations.

Caution: beware of extrapolation! It would be unwise to predict the TDS for a Discharge value of 12,000 m^3/s as this is far beyond the observed data range.

Mathematics Learning Support Centre

Simple Linear Regression: Reliability of predictions

Richard Buxton. 2008.

1 Introduction

We often use regression models to make predictions.

In Figure 1 (a), we've fitted a model relating a household's weekly gas consumption to the average outside temperature[1]. We can now use the model to predict the gas consumption in a week when the outside temperature is say 6 deg C.

Similarly, in Figure 1 (b), we've fitted a model relating the lung capacity (FEV1) of a child to their age[2]. We can use this model to predict the lung capacity of an 8 year old.

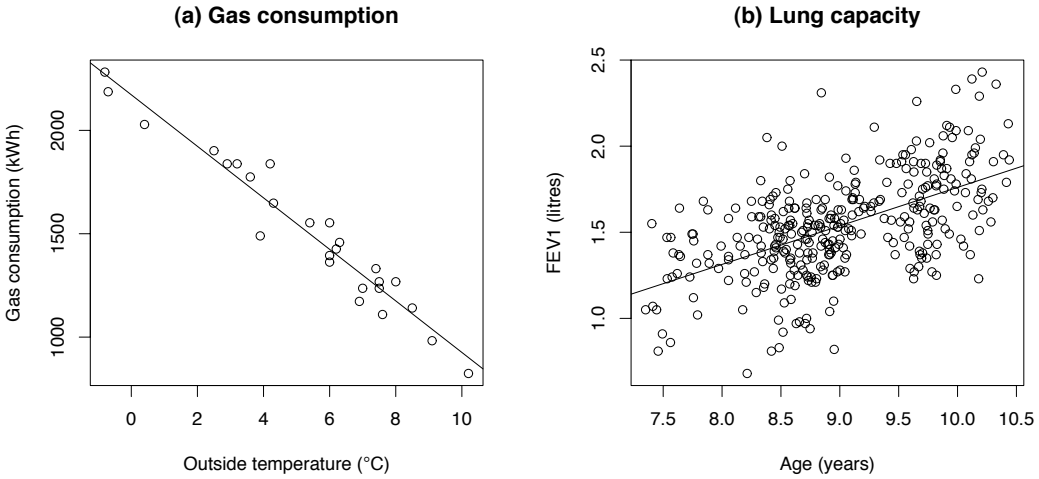


Figure 1: Models for gas consumption and lung capacity

Our predictions are nearly always subject to some uncertainty. This uncertainty arises because not all the variation in the response can be explained by the fitted model. By making some assumptions about the unexplained variation, we can quantify the uncertainty and calculate a confidence interval, or range of plausible values for a prediction.

This handout explains how to check the assumptions of simple linear regression and how to obtain confidence intervals for predictions.

[1]Source of data: Hand (1994)
[2]Source of data: Kirkwood and Sterne (2003)

If you're new to regression analysis, you'll probably find it useful to read the leaflet 'Simple Linear Regression: Introduction' before continuing with this one.

2 Assumptions of simple linear regression

We make the following assumptions...

- Mean response varies linearly with predictor
- Unexplained variation is Normally and independently distributed with constant variance

To check these assumptions, we look at plots of the *residuals* and *fitted values*. The fitted values are the values of the response predicted by the model. The residuals are obtained by taking the observed values of the response and subtracting the fitted values. The two most useful plots are...

- Plot of Residuals vs Fitted values
  - We can use this plot to check the assumptions of linearity and constant variance. For example, Figure 2 shows some plots for a regression model relating stopping distance to speed[3]. The plot on the left shows the data, with a fitted linear model. The plot on the right shows the residuals plotted against the fitted values - a smooth curve has been added to highlight the pattern of the plot.

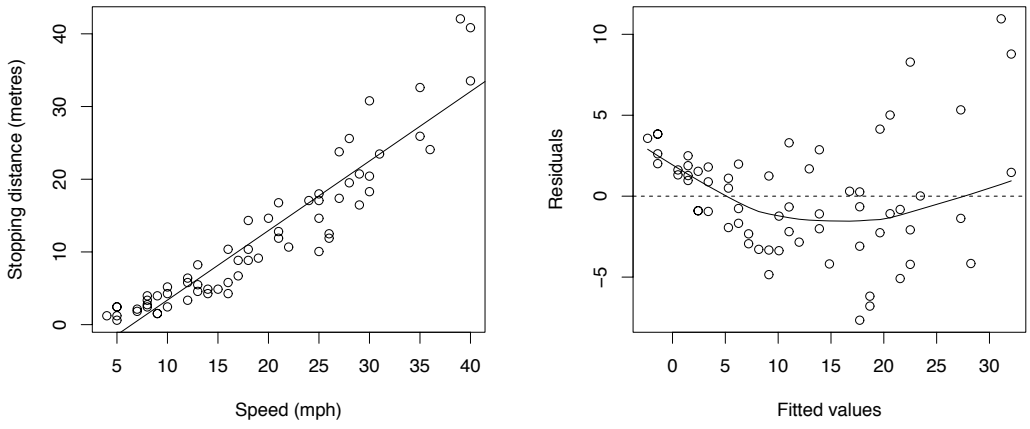


Figure 2: Stopping distance vs Speed

[3]Source of data: Hand (1994)

Ideally, the residual plot should show a horizontal band of roughly equal width. In this case, we have a strong 'U' shape, suggesting that the residuals go from positive to negative to positive. This suggests that we're fitting a line to a non-linear relationship - see plot of original data. In addition, the width of the band of data increases from the left to the right, suggesting that the variance is increasing. There are various courses of action that we can take to deal with these problems - for details, consult a Statistician.

Figure 3 shows some diagnostic plots for the regression of lung capacity on age that we looked at in Section 1. Looking at the plot of residuals vs fitted values, we have a horizontal band of data of roughly constant width. So the assumptions of linearity and constant variance do seem to hold here.

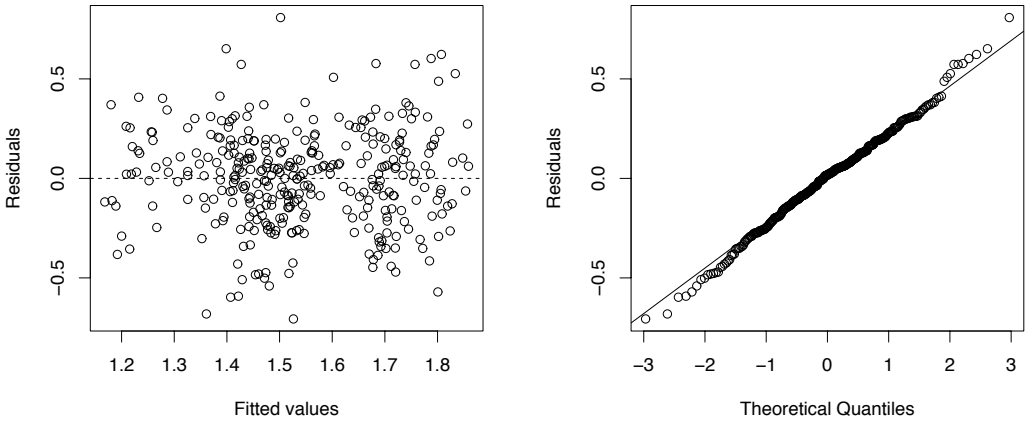


Figure 3: Diagnostic plots for Lung data

- Normal probability plot of residuals
  - This plot is used to check the assumption that the unexplained variation follows a Normal distribution. Normality is indicated by a roughly linear plot. Any strong systematic curvature suggests some degree of non-Normality. The Normal plot in Figure 3 is roughly linear, confirming that the unexplained variation is roughly Normal.

There are several ways of checking the assumption that the random variation is statistically independent. For details, see Koop (2008). The assumption of independence is not usually a problem except for data that has been collected at successive points in time - e.g. monthly unemployment figures.

3 Checking assumptions in SPSS

—**Analyse**
—**Regression**
—**Linear**
—drag the response variable into the **Dependent** box
—drag the predictor variable into the **Independent(s)** box
—**Plots**
—drag the residuals (***ZRESID**) into the **Y** box
—drag the fitted values (***ZPRED**) into the **X** box
—Select **Normal Probability Plot**
—Click **Continue**
—Click **OK**

Use the plot of Residuals against Fitted values to check for any evidence of non-linearity or non-constant variance. Use the Normal probability plot to check for evidence of non-Normality.

4 Confidence intervals for predictions

Provided the assumptions in Section 2 are satisfied, we can obtain confidence intervals for any predictions that we make.

We illustrate with the example on lung capacity (FEV1) vs age. The fitted model is...

$$F = -0.475 + 0.224\ A$$

... where F is FEV1 and A is Age.

Suppose we want to predict FEV1 for a child of 9. We can obtain a point prediction by simply substituting 9 in place of A in the fitted model. But calculating a confidence interval is more difficult, so in practice, we use statistical software to make our predictions.

There are two types of confidence interval...

- Confidence interval for individual case
  - Range of plausible values for a single case - e.g. for the FEV1 of a single child
- Confidence interval for mean
  - Range of plausible values for the mean - e.g. for the mean FEV1 over a large number of children, all of the same age.

Table 1 shows the SPSS output for age 9.

Age	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
9.00	1.53695	1.51102	1.56288	1.06161	2.01229

Table 1: Prediction and confidence intervals

If we're predicting the FEV1 for a single child, we use the columns headed LICI_1 (Lower Individual Confidence Interval) and UICI_1 (Upper Individual Confidence Interval).

95% confidence interval 1.06 to 2.01 litres

We can be fairly sure that the FEV1 value will lie within this range.

If we wished to predict the *mean* value of FEV1 for a large group of children, all of age 9, we would use the columns LMCI_1 and UMCI_1.

95% confidence interval for mean 1.51 to 1.56 litres

This interval is much narrower. We're much less sure about the lung function of a single child than we are about the mean lung function for a large group of children.

5 Making predictions in SPSS

Go to the SPSS Data Editor and add the new predictor values (i.e. the values at which you wish to make predictions) to the bottom of the column containing the predictor.

—**Analyse**
—**Regression**
—**Linear**
—drag the response variable into the **Dependent** box
—drag the predictor variable into the **Independent(s)** box
—**Save**
—under **Predicted Values**, select **Unstandardized**
—under **Prediction Intervals**, select **Mean** or **Individual**
—Click **Continue**
—Click **OK**

SPSS calculates the predicted values and confidence intervals and puts them in five new columns in the Data Editor window.

6 Confidence interval for slope of regression model

We're sometimes interested in the change in the response corresponding to a given change in the predictor. For example, how much will our stopping distance increase if we travel 10mph faster? We can answer this kind of question by looking at the *slope* of the regression line.

Table 2 shows some SPSS output giving the coefficients of the Lung model, together with confidence intervals for both the slope and intercept.

Coefficients

Model		Unstandardized Coefficients		95% Conf Int for B	
		B	Std.Error	Lower Bound	Upper Bound
1	(Constant)	-.475	.164	-.798	-.153
	A	.224	.018	.188	.259

Table 2: Confidence intervals for coefficients

The coefficient of A is 0.224. This tells us that an increase of one year in age is associated with an increase in FEV1 of around 0.224 litres. The columns on the right of the table give a confidence interval for this figure.

95% confidence interval 0.188 to 0.259 litres

This gives us a range of plausible values for the increase in FEV1 corresponding to a unit increase in age.

If we're interested in the change in FEV1 corresponding to say a *two* year increase in age, we can obtain a confidence interval by simply multiplying the lower and upper ends of our confidence interval by 2 to give...

95% confidence interval 0.376 to 0.518 litres.

7 Estimating the slope in SPSS

—**Analyse**
—**Regression**
—**Linear**
—drag the response variable into the **Dependent** box
—drag the predictor variable into the **Independent(s)** box
—**Statistics**
—under **Regression Coefficient**, select **Confidence Intervals**
—Click **Continue**
—Click **OK**

The table of coefficients will now include confidence intervals for the intercept and slope.

8 References

For a simple *introduction* to regression, see Moore and McCabe (2004). For a more comprehensive treatment, see Freund and Wilson (1998).

Freund, R.J. and Wilson, W.J. (1998). Regression Analysis: statistical Modeling of a Response Variable, Academic Press.

Hand, D.J. (1994). A Handbook of Small Data Sets, Chapman and Hall.

Kirkwood, B.R. and Sterne, J.A.C. (2003). Essential Medical Statistics, Blackell Science.

Koop, G. (2008). Introduction to Econometrics, Wiley.

Moore, D.S. and McCabe, G.P. (2004). Introduction to the practice of statistics, 5th edition, W.H.Freeman.

Spearman's correlation

Introduction

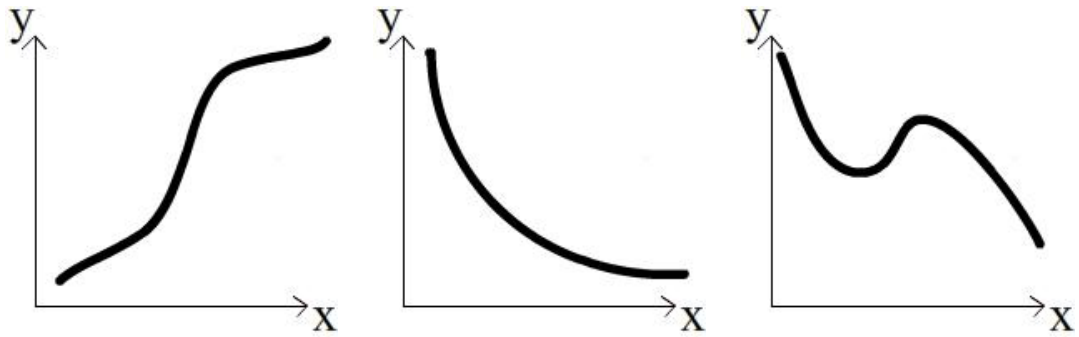
Before learning about Spearman's correllation it is important to understand Pearson's correlation which is a statistical measure of the strength of a *linear* relationship between paired data. Its calculation and subsequent significance testing of it requires the following data assumptions to hold:

- interval or ratio level;
- linearly related;
- bivariate normally distributed.

If your data does not meet the above assumptions then use Spearman's rank correlation!

Monotonic function

To understand Spearman's correlation it is necessary to know what a monotonic function is. A monotonic function is one that either never increases or never decreases as its independent variable increases. The following graphs illustrate monotonic functions:



Monotonically increasing Monotonically decreasing Not monotonic

- Monotonically increasing - as the x variable increases the y variable never decreases;
- Monotonically decreasing - as the x variable increases the y variable never increases;
- Not monotonic - as the x variable increases the y variable sometimes decreases and sometimes increases.

Spearman's correlation coefficient

Spearman's correlation coefficient is a statistical measure of the strength of a *monotonic* relationship between paired data. In a sample it is denoted by r_s and is by design constrained as follows

$$-1 \leq r_s \leq 1$$

And its interpretation is similar to that of Pearsons, e.g. the closer r_s is to ± 1 the stronger the monotonic relationship. Correlation is an effect size and so we can verbally describe the strength of the correlation using the following guide for the absolute value of r_s:

- .00-.19 "very weak"
- .20-.39 "weak"
- .40-.59 "moderate"
- .60-.79 "strong"
- .80-1.0 "very strong"

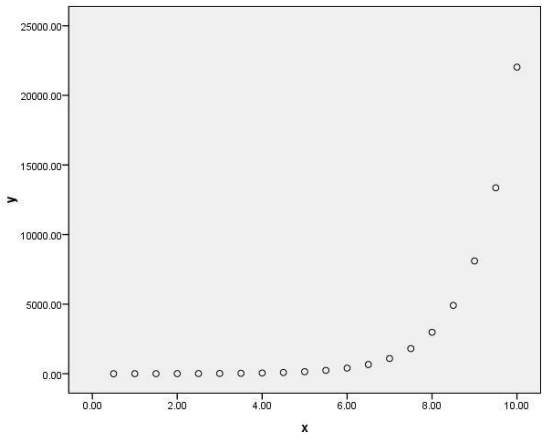
The calculation of Spearman's correlation coefficient and subsequent significance testing of it requires the following data assumptions to hold:

- interval or ratio level or ordinal;
- monotonically related.

Note, unlike Pearson's correlation, there is no requirement of normality and hence it is a nonparametric statistic.

Let us consider some examples to illustrate it. The following table gives x and y values for the relationship $y = \exp(x)$. From the graph we can see that this is a perfectly increasing monotonic relationship.

	x	y
1	.5	1.6
2	1.0	2.7
3	1.5	4.5
4	2.0	7.4
5	2.5	12.2
6	3.0	20.1
7	3.5	33.1
8	4.0	54.6
9	4.5	90.0
10	5.0	148.4
11	5.5	244.7
12	6.0	403.4
13	6.5	665.1
14	7.0	1096.6
15	7.5	1808.0
16	8.0	2981.0
17	8.5	4914.8
18	9.0	8103.1
19	9.5	13359.7
20	10.0	22026.5

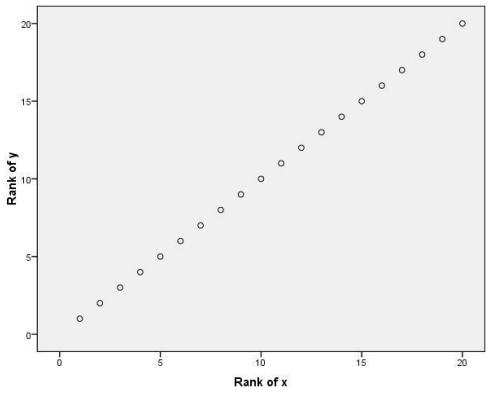


The calculation of Pearson's correlation for this data gives a value of .699 which does not reflect that there is indeed a perfect relationship between the data. Spearman's correlation for this data however is 1, reflecting the perfect monotonic relationship.

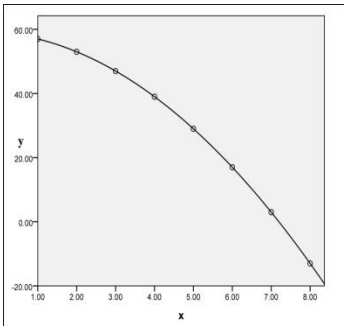
Spearman's correlation works by calculating Pearson's correlation on the ranked values of this data. Ranking (from low to high) is obtained by assigning a rank of 1 to the lowest value, 2 to the next lowest and so on.

If we look at the plot of the ranked data, then we see that they are perfectly linearly related.

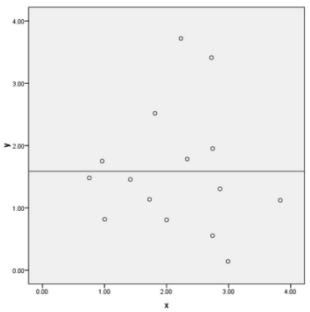
	x	Rank of x	y	Rank of y
1	.5	1	1.6	1
2	1.0	2	2.7	2
3	1.5	3	4.5	3
4	2.0	4	7.4	4
5	2.5	5	12.2	5
6	3.0	6	20.1	6
7	3.5	7	33.1	7
8	4.0	8	54.6	8
9	4.5	9	90.0	9
10	5.0	10	148.4	10
11	5.5	11	244.7	11
12	6.0	12	403.4	12
13	6.5	13	665.1	13
14	7.0	14	1096.6	14
15	7.5	15	1808.0	15
16	8.0	16	2981.0	16
17	8.5	17	4914.8	17
18	9.0	18	8103.1	18
19	9.5	19	13359.7	19
20	10.0	20	22026.5	20



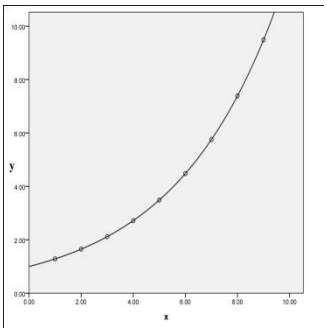
In the figures below various samples and their corresponding sample correlation coefficient values are presented. The first three represent the "extreme" monotonic correlation values of -1, 0 and 1:



$r_s = -1$
perfect –ve
monotonic correlation

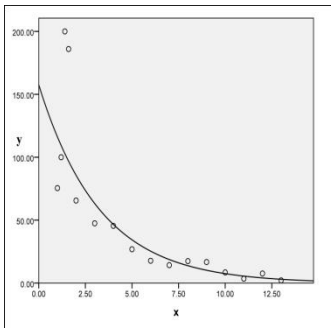


$r_s = 0$
no correlation

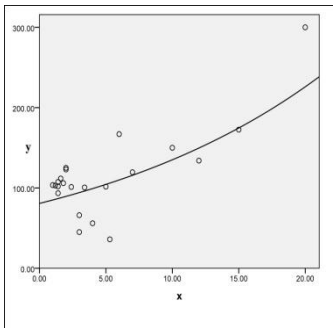


$r_s = 1$
perfect +ve
monotonic correlation

Invariably what we observe in a sample are values as follows:

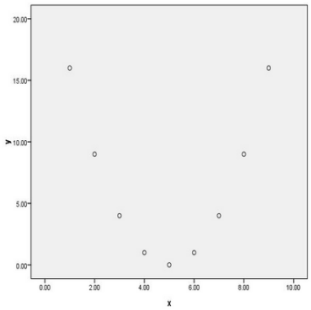


$r_s = -.941$
very strong -ve
monotonic correlation



$r_s = .372$
weak +ve
monotonic correlation

Note: Spearman's correlation coefficient is a measure of a monotonic relationship and thus a value of $r_s = 0$ does not imply there is no relationship between the variables. For example in the following scatterplot $r_s = 0$ which implies no (monotonic) correlation however there is a perfect quadratic relationship:



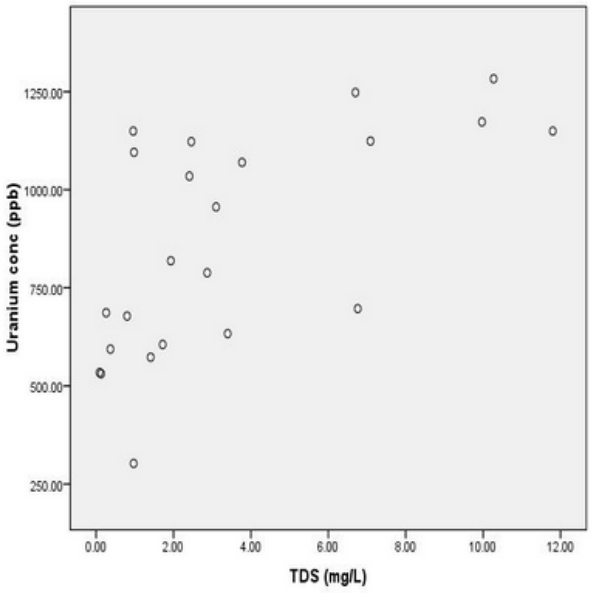
$r_s = 0$
perfect quadratic relationship

Example

The following data comprises 23 groundwater samples that were collected recording the Uranium concentration (ppb) and the total dissolved solids (mg/L). It is of interest to know if the two variables are correlated?

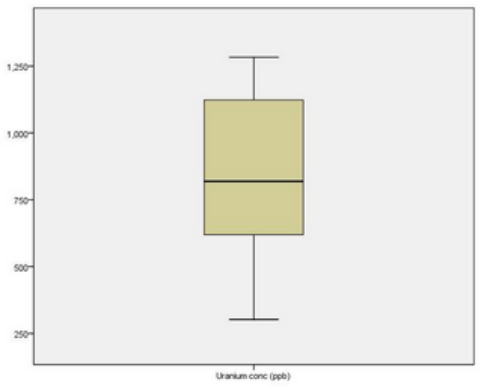
We should initial consider if Pearson's correlation is appropriate or whether we should resort to Spearman's if there are assumption violations.

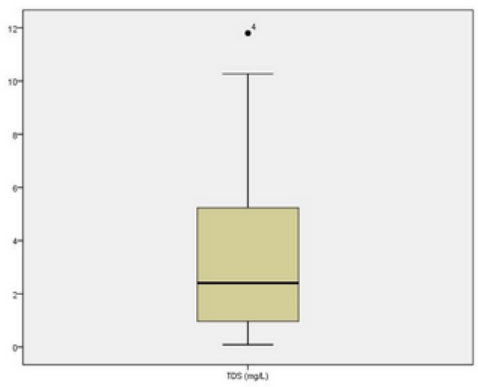
	Uranium conc (ppb)	TDS (mg/L)
1	678.10	.80
2	818.93	1.93
3	302.38	.97
4	1149.60	11.80
5	573.14	1.41
6	1034.55	2.41
7	633.25	3.40
8	1095.42	.98
9	1122.58	2.46
10	686.51	.26
11	1172.84	9.97
12	593.70	.37
13	1247.95	6.70
14	533.99	.09
15	605.51	1.72
16	696.96	6.76
17	1282.95	10.27
18	531.16	.13
19	788.36	2.87
20	956.06	3.10
21	1149.38	.96
22	1069.82	3.77
23	1124.17	7.09



The scatterplot suggests a definite positive correlation between Uranium and TDS. However, there is possibly slight evidence of non-linearity for TDS values close to zero. However, this is debateable and so we shall move on and consider the other normality assumption.

We need to perform some normality checks for the two variables. One simple way of doing this is to examine boxplots of the data. These are given below.





The boxplot for Uranium is fairly consistent with one from a normal distribution; the median is fairly close to the centre of the box and the whiskers are of approximate equal length.

The boxplot for TDS is slightly disturbing in that the median is close to the lower quartile and the lower whisker is shorter than the upper one, which would be suggesting positive skewness. Also there is an outlier and Pearson's correlation is sensitive to these as well as skewness.

Since we have some doubts over normality, we shall examine the skewness coefficients to see if there is further evidence to suggest whether either of the variables is skewed.

Descriptive Statistics

	N	Skewness	
	Statistic	Statistic	Std. Error
Uranium conc (ppb)	23	-.148	.481
Valid N (listwise)	23		

Descriptive Statistics

	N	Skewness	
	Statistic	Statistic	Std. Error
TDS (mg/L)	23	1.189	.481
Valid N (listwise)	23		

A quick check to see if the skewness coefficients are not sufficiently large to warrant concern is to see if the absolute values of the skewness coefficients are less than two times their standard errors. Using this guide, the Uranium data's skewness is consistent with the data being normal. However the TDS skewness coefficient appears to be large enough to warrant concern that ther is positive skewness present (1.189 > 2 x .481).

Hence we do have concerns over the normality of our data and should continue with a Spearman's correlation analysis. SPSS produces the following Spearman's correlation output:

Correlations

			Uranium conc (ppb)	TDS (mg/L)
Spearman's rho	Uranium conc (ppb)	Correlation Coefficient	1.000	.708**
		Sig. (2-tailed)	.	.000
		N	23	23
	TDS (mg/L)	Correlation Coefficient	.708**	1.000
		Sig. (2-tailed)	.000	.
		N	23	23

**. Correlation is significant at the 0.01 level (2-tailed).

The significant Spearman correlation coefficient value of 0.708 confirms what was apparent from the graph; there appears to be a strong positive correlation between the two variables. Thus large values of uranium are associated with large TDS values

However, we need to perform a significance test to decide whether based upon this sample there is any or no evidence to suggest that linear correlation is present in the population. To do this we test the null hypothesis, H_0, that there is no monotonic

correlation in the population against the alternative hypothesis, H_1, that there is monotonic correlation; our data will indicate which of these opposing hypotheses is most likely to be true. Let ρ_s be the Spearman's population correlation coefficient then we can thus express this test as:

$$H_0 : \rho_s = 0$$
$$H_1 : \rho_s \neq 0$$

i.e. the null hypothesis of no monotonic correlation present in population against the alternative that there is monotonic correlation present.

Since SPSS reports the p-value for this test as being .000 we can say that we have very strong evidence to believe H_1, i.e. we have some evidence to believe that groundwater uranium and TDS values are monotonically correlated in the population.

This could be formally reported as follows:

"A Spearman's correlation was run to determine the relationship between 23 groundwater uranium and TDS values. There was a strong, positive monotonic correlation between Uranium and TDS (r_s = .71, n = 23, p < .001)."

Mathematics Learning Support Centre

Statistics: 2.2 The Wilcoxon signed rank sum test

Rosie Shier. 2004.

1 Introduction

The Wilcoxon signed rank sum test is another example of a non-parametric or distribution free test (see 2.1 The Sign Test). As for the sign test, the Wilcoxon signed rank sum test is used is used to test the null hypothesis that the median of a distribution is equal to some value. It can be used a) in place of a one-sample t-test b) in place of a paired t-test or c) for ordered categorial data where a numerical scale is inappropriate but where it is possible to rank the observations.

2 Carrying out the Wilcoxon signed rank sum test

Case 1: Paired data

1. State the null hypothesis - in this case it is that the median difference, M, is equal to zero.
2. Calculate each paired difference, $d_i = x_i - y_i$, where x_i, y_i are the pairs of observations.
3. Rank the d_is, ignoring the signs (i.e. assign rank 1 to the smallest $|d_i|$, rank 2 to the next etc.)
4. Label each rank with its sign, according to the sign of d_i.
5. Calculate W^+, the sum of the ranks of the positive d_is, and W^-, the sum of the ranks of the negative d_is. (As a check the total, $W^+ + W^-$, should be equal to $\frac{n(n+1)}{2}$, where n is the number of pairs of observations in the sample).

Case 2: Single set of observations

1. State the null hypothesis - the median value is equal to some value M.
2. Calculate the difference between each observation and the hypothesised median, $d_i = x_i - M$.
3. Apply Steps 3-5 as above.

Under the null hypothesis, we would expect the distribution of the differences to be approximately symmetric around zero and the the distribution of positives and negatives to be distributed at random among the ranks. Under this assumption, it is possible to

work out the exact probability of every possible outcome for W. To carry out the test, we therefore proceed as follows:

6. Choose $W = \min(W^-, W^+)$.
7. Use tables of critical values for the Wilcoxon signed rank sum test to find the probability of observing a value of W or more extreme. Most tables give both one-sided and two-sided p-values. If not, double the one-sided p-value to obtain the two-sided p-value. This is an exact test.

Normal approximation
If the number of observations/pairs is such that $\frac{n(n+1)}{2}$ is large enough (> 20), a normal approximation can be used with $\mu_W = \frac{n(n+1)}{4}, \quad \sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}}$

Dealing with ties:
There are two types of tied observations that may arise when using the Wilcoxon signed rank test:

- Observations in the sample may be exactly equal to M (i.e. 0 in the case of paired differences). Ignore such observations and adjust n accordingly.
- Two or more observations/differences may be equal. If so, average the ranks across the tied observations and reduce the variance by $\frac{t^3-t}{48}$ for each group of t tied ranks.

Example:
The table below shows the hours of relief provided by two analgesic drugs in 12 patients suffering from arthritis. Is there any evidence that one drug provides longer relief than the other?

Case	Drug A	Drug B	Case	Drug A	Drug B
1	2.0	3.5	7	14.9	16.7
2	3.6	5.7	8	6.6	6.0
3	2.6	2.9	9	2.3	3.8
4	2.6	2.4	10	2.0	4.0
5	7.3	9.9	11	6.8	9.1
6	3.4	3.3	12	8.5	20.9

Solution:

1. In this case our null hypothesis is that the median difference is zero.
2. Our actual differences (Drug B - Drug A) are:

   $$+1.5, +2.1, +0.3, -0.2, +2.6, -0.1, +1.8, -0.6, +1.5, +2.0, +2.3, +12.4$$

   Our actual median difference is 1.65 hours.
3. Ranking the differences and affixing a sign to each rank (steps 3 and 4 above):

Diff.	0.1	0.2	0.3	0.6	1.5	1.5	1.8	2.0	2.1	2.3	2.6	12.4
Rank	1	2	3	4	5.5	5.5	7	8	9	10	11	12
Sign	-	-	+	-	+	+	+	+	+	+	+	+

Calculating W^+ and W^- gives:
$W^- = 1 + 2 + 4 = 7 \qquad W^+ = 3 + 5.5 + 5.5 + 7 + 8 + 9 + 10 + 11 + 12 = 71$
Therefore, we have $n = \frac{12\times13}{2} = 78 W = \max(W^-, W^+) = 71$.
We can use a normal approximation in this case. We have one group of 2 tied ranks, so we must reduce the variance by $\frac{8-2}{48} = 0.125$. We get:

$$z = \frac{71 - \frac{12\times13}{4}}{\sqrt{\frac{12\times13\times25}{24} - 0.125}} = \frac{71 - 39}{\sqrt{162.5 - 0.125}} = 2.511$$

This gives a two-sided p-value of $p = 0.012$. There is strong evidence that Drug B provides more relief than Drug A.

3 Carrying out the Wilcoxon signed rank sum test in SPSS

— Choose **Analyze**
— Select **Nonparametric Tests**
— Select **2 Related Samples**
— Specify which two variables comprise your pairs of observation by clicking on them both then clicking on the arrow to put them under **Test Pair(s) List**.
— Under **Test Type** select **Wilcoxon**
— If you want exact probabilities (i.e. based on the binomial distribution), click on **Exact**, choose **Exact**, then **Continue**
— Click on **OK**

The output will look like this:

Ranks

		N	Mean Rank	Sum of Ranks
Drug B - Drug A	Negative Ranks	3[a]	2.33	7.00
	Positive Ranks	9[b]	7.89	71.00
	Ties	0[c]		
	Total	12		

a. DRUGB < DRUGA
b. DRUGB > DRUGA
c. DRUGA = DRUGB

Test Statistics[b]

	DRUGB - DRUGA
z	−2.511[a]
Asymp. Sig. (2-tailed)	0.012
Exact Sig. (2-tailed	0.009
Exact Sig. (1-tailed	0.004
Point probability	0.001

a. Based on negative ranks
b. Wilcoxon Signed Ranks Test

When you do a principal components analysis you get what are called eigenvalues. It is not necessary to understand what eigenvalues are in order to understand the principles behind principal components analysis. However, it is useful to know that the eigenvalues are the variances of the principal components. In other words, the first eigenvalue is the variance of the first principal component, the second eigenvalue is the variance of the second principal component, and so on. Thus, because of the way the principal components are selected, the first eigenvalue will be the largest, the second the next largest, etc. There will be p eigenvalues altogether but some may be equal to zero.

Once the principal components have been calculated you will need to decide how many to keep. Essentially any principal components that account for only a small proportion of the variation in the data (i.e. those with small eigenvalues) are discarded. Different methods are used to decide which principal components to retain:

- Choose sufficient principal components to account for a particular percentage (e.g. 75%) of the total variability in the data.
- Choose only those principal components with eigenvalues over 1 (if using the correlation matrix).
- Use the scree plot of the eigenvalues. This will indicate whether there is an obvious cut-off between large and small eigenvalues.

4 Carrying out principal components analysis in SPSS

Note that SPSS will not give you the actual principal components. However, these can be calculated from the output provided.

– **Analyze**
– **Data Reduction**
– **Factor**
– Select the variables you want the factor analysis to be based on and move them into the **Variable(s)** box.
– In the **Extraction** window, select **Principal components**. Under **Analyze** ensure that **Correlation Matrix** is selected (this is the default). The default is also to extract eigenvalues over 1. You can either keep it like this or specify the number of factors to be equal to the number of original variables (later on you can decide which principal components to keep and which to discard). Click on **Continue**.
– In the **Rotation** window, select **None** under **Method**. Click on **Continue**.
– In the **Scores** window you can specify whether you want SPSS to save the values of the 'principal components' (as mentioned above, these are not actually the principal components but can be used to calculate them) for each observation (this will save them as new variables in the data set). Under **Method** choose **Regression**. Click on **Continue**.
– **OK**

IMPORTANT NOTE
Note that what SPSS gives you are NOT the principal components. However, these can be calculated quite easily:

1. To get the coefficients of the principal components (the as):
SPSS will give you a table entitled the **Component Matrix**. The components are listed as columns in this table; the variables are listed as rows. To get the values of the as you must **DIVIDE** the values in the table by the SQUARE ROOT of the corresponding eigenvalue. For example, to get $a_{11}, a_{12}, a_{13}, \ldots$ you should divide EACH number in the first column by the square root of the first (largest) eigenvalue (i.e. the eigenvalue corresponding to component 1). Similarly, the second column should be divided by the square root of the eigenvalue corresponding to component 2, and so on.

2. To get the values of the principal components as new variables in the data set:
SPSS will have saved variables called **FAC1_1, FAC2_1**, and so on. These are NOT the values of the principal components. To get the principal components you have to **MULTIPLY** these factor scores by the square root of the corresponding eigenvalue. For example, if the eigenvalue for the first principal component was 3.65, you would compute the first principal component in SPSS as follows:

– **Transform**
– **Compute**
– Under **Target variable** write PC1 (or something similar — to stand for first principal component) and under **Numeric Expression** type in FAC1_1 * sqrt(3.65).
– Click on **OK**

Use a similar process to compute the values of the second principal component (calling this one PC2, using FAC2_1 and replacing the 3.65 by whatever the second eigenvalue is).

5 References

- Manly, B.F.J. (2005), **Multivariate Statistical Methods: A primer**, Third edition, Chapman and Hall.
- Rencher, A.C. (2002), **Methods of Multivariate Analysis**, Second edition, Wiley.