10. When assumptions are violated - data transformation and non-parametric methods

Log-normally distributed data

Log-normal distribution is very common in any kind of real data, i.e. random variables logarithm of which follows normal distribution. As a result, log-normal variables may range from zero limit (excluding zero itself) to plus infinity – that is pretty realistic e.g. for dimensions, mass, time etc. In contrast to normal distribution, log-normal variables are positively skewed (i.e. are not distributed symmetrically around the mean) and display a positive correlation between mean and variance (Fig. 10.1). A straightforward suggestion for such data is to apply log-transformation of the values to obtain normally distributed variables (Figs 10.1, 10.2, Table 10.1). ANOVA applied on non-transformed and transformed data provides quite different results (Table 10.1.).

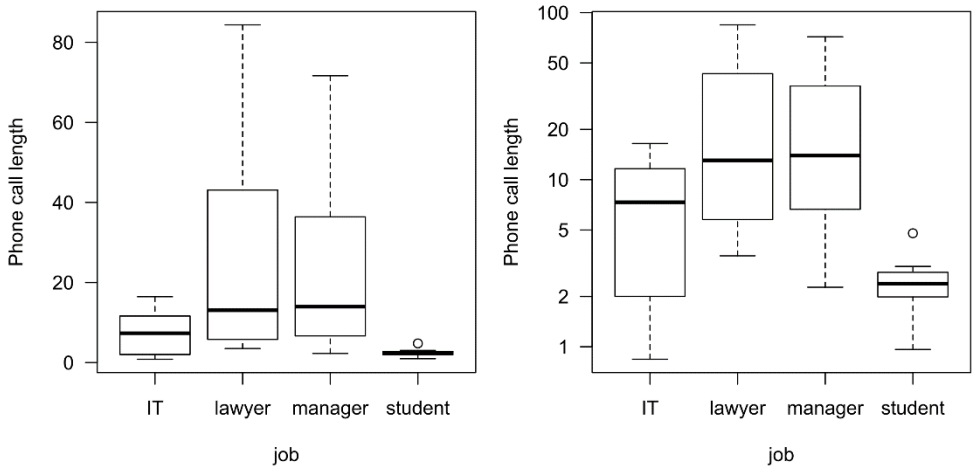

Fig. 10.1. Example of a log-normal variable: length of phone calls in dependence of job of the person calling. Left panel shows the boxplot on the ordinary linear scale, while the right panel shows the same values on the log-scaled y-axis.

Table 10.1. Summaries of ANOVA applied on non-tansformed and transformed data displayed on Fig. 10.1.

Analysis	R^2	F	DF	p
non-transformed	0.26	4.13	3,36	0.013
log-transformed	0.42	8.72	3,36	0.0002

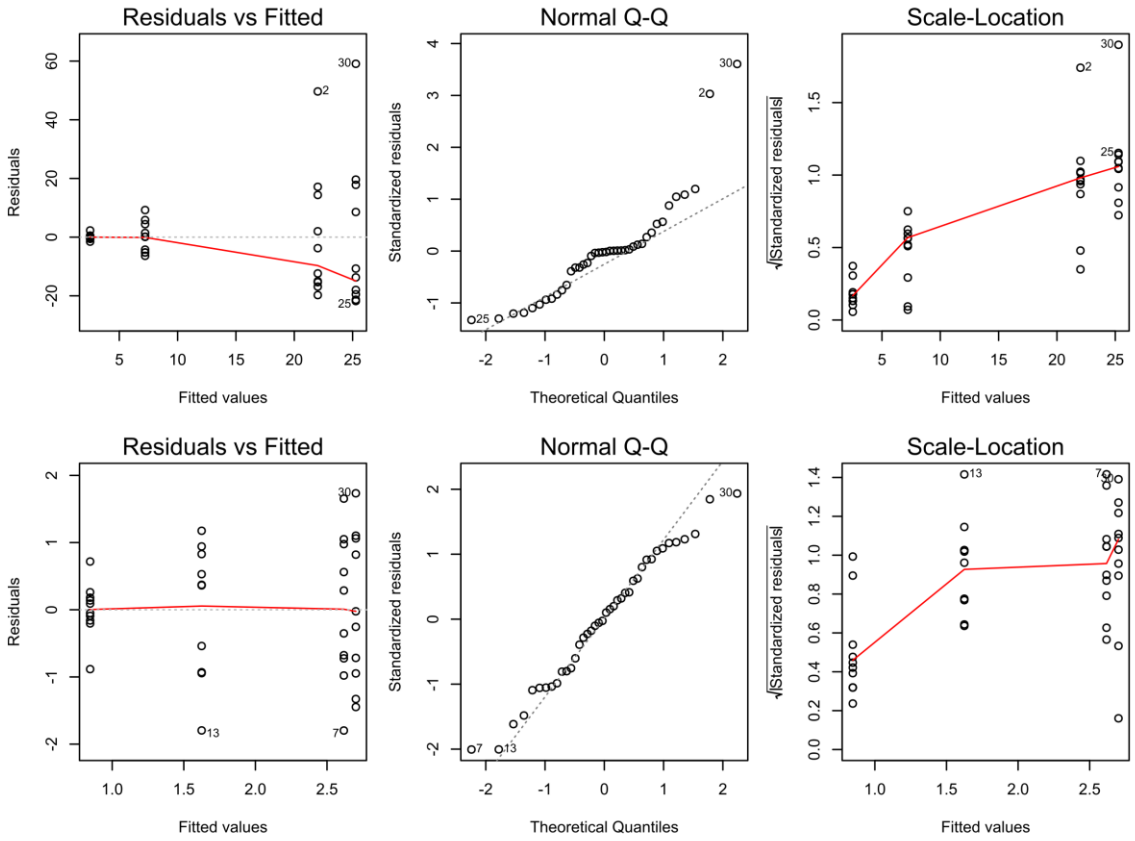

Fig. 10.1. Diagnostic plots of ANOVA models applied on non-transformed (upper row of plots) and log-transformed data (lower row of plots). Note improved normal fit on the QQplot and homogeneity of variances after transformation (Residuals vs. Fitted and Scale-Location plots).

Note, that log-transformation is not a simple utility procedure, it also affects the interpretation of the analysis. Log-transformation changes the scale from additive to multiplicative, i.e. we test the null hypothesis stating that the ratio between population means is 1 (instead of difference being 0). We also consider different means – analysis on log-scale implies testing geometric means on the original scale. The same applies for regression coefficients, which become relative rather than absolute numbers e.g. the slope indicates how many times the response variable will change with a change in predictor. An example with log-transformation in linear regression is displayed on Fig. 10.3., 10.4. and Table 10.2.

Log-transformation is sometimes used also for data, which are not log-normally distributed, but are just positively skewed. Such data may contain zeros and thus are not log-transformable. Instead log (x + constant) transformation must be used. Alternatively, square-root transformation may be considered for such data.

Note, that the analysis results do not depend on logarithm used – natural and decadic logarithms are used most frquently. Just beware to be consistent in using the same logarithm throughout the analysis.

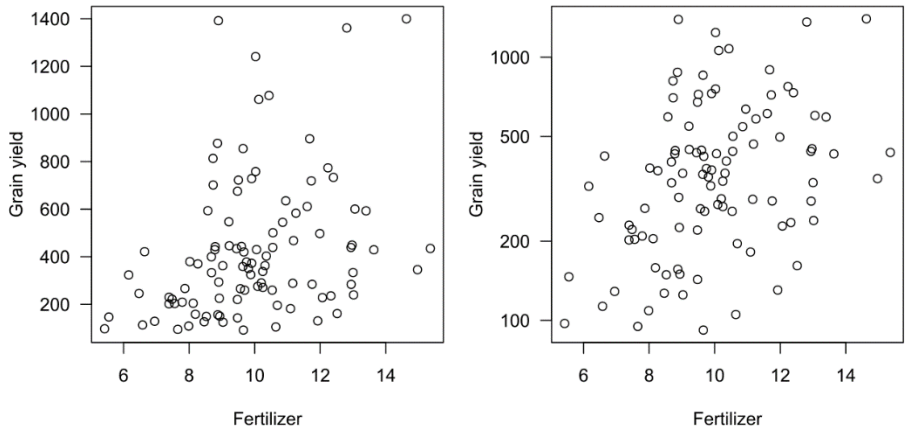

Fig. 10.3. Example of a regression with log-normal variable: how grain yield of maize depends on amount of fertilizer applied. Left panel shows the scatterplot on the ordinary linear scale, while the right panel shows the same values on the log-scaled y-axis.

Table 10.2. ANOVA tables of linear models fitted on non-tansformed and transformed data displayed on Fig. 10.3.

Analysis	R^2	F	DF	p
non-transformed	0.10	11.0	1,98	0.0013
log-transformed	0.14	16.05	1,98	0.0001

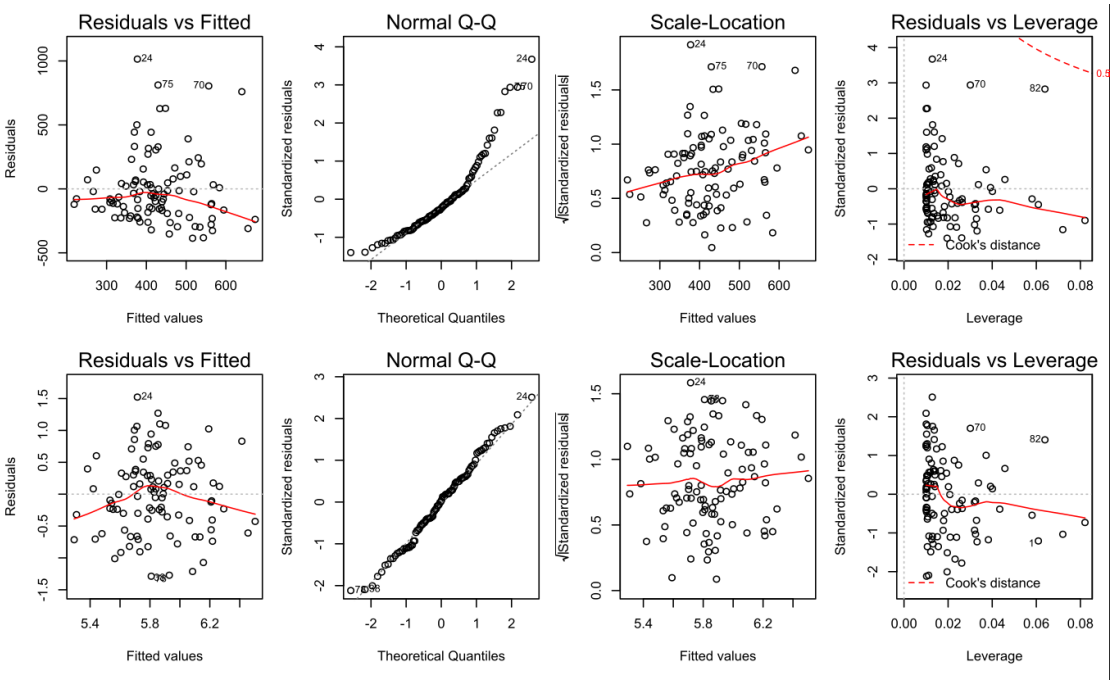

Fig. 10.4. Diagnostic plots of linear models fitted on non-transformed (upper row of plots) and log-transformed data (lower row of plots). Note improved normal fit on the QQplot and improved homogeneity of variances after transformation (Scale-Location plot).

Non-parametric tests

Some distributions cannot be approximated by normal distribution and simple transformations are not helpful. This applies e.g. on many data on ordinal scale, such as schoolgrades, subjective rankings etc. For such cases, non-parametric tests were developed (Table 10.3.). These tests replace original values by value order and use these data to test differences in central tendencies (which are not exactly means) between the samples based only on the assumption, that the samples come from the same distribution.

Table 10.3. List of parametric tests and treir non-parametric counterparts together with appropriare R functions.

Parametric test	Non-parametric test	R function
two-sample t-test	Mann-Whitney U test	wilcox.test
paired t-test	Wilcoxon test	wilcox.test with parameter *paired=T*
One way ANOVA	Kruskal-Wallis test*	kruskal.test
Pearson correlation	Spearman correlation	cor.test with parameter *method="spearman"*

* Dunn test may be used for post-hoc comparisons (function dunnTest in package FSA)

Permutation tests

Permutation tests represent useful alternatives to parametric tests. First, a statistic of difference from null hypothesis (between samples) is defined. That may be raw or relative difference or an F-ratio if multiple groups are analyzed. This statistic is computed for observed data (observed statistic). Subsequently, values of response variable are repeatedly permuted (reshuffled) and the same statistic is computed in each permutation. P-value is then determined by the formula:

$$p = \frac{x + 1}{n_{perm} + 1}$$

where *x* is the number of permutations in which test statistic was higher than observed test statistic and n_{perm} is the total number of permutations.

How to do in R

1. Log-scaling of graph axis: parameter log='axis to be log-scaled', i.e. mostly log='y'
2. Log-transformation: function **log** for natural logarithm, **log10** for decadic
3. Non-parametric tests: see Table 10.3.
4. Permutation tests are available in library coin:
    a. permutation-based ANOVA: function **oneway_test**
    b. permutation-based correlation: **spearman_test**
       Both methods require parameter distribution=approximate(B=number of permutations) to be set B is usually set to 999 or 9999.