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

Log-normally distributed data

Log-normal distribution is very common in many kinds of biological data. These are random variables logarithm of which follows the normal distribution. As a result, log-normal variables may range from the zero limit (excluding zero itself) to plus infinity – that is pretty realistic e.g. for dimensions, mass, time, etc. In contrast to the symmetric normal distribution, log-normal variables are positively skewed and display a positive correlation between mean and variance (Fig. 10.1). A straightforward suggestion for such data is to apply log-transformation to 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.).

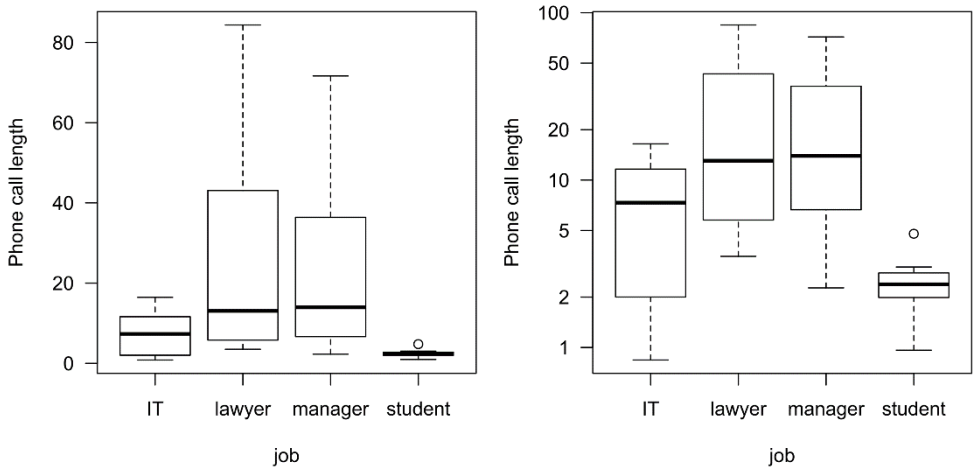

Fig. 10.1. Example of a log-normal variable: effect of job on the length of phone calls. The left panel shows the boxplot on the ordinary linear scale, while the right panel shows the same values on the log-scaled y-axis.

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

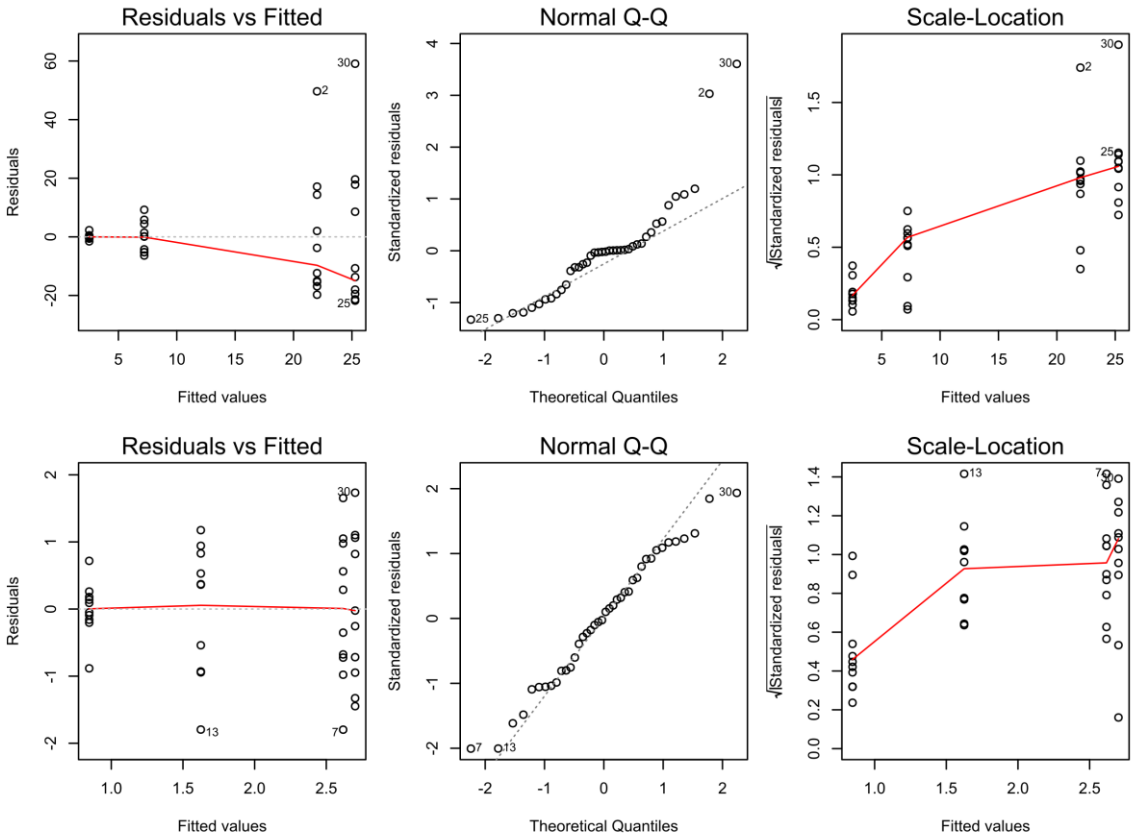

Fig. 10.2. Diagnostic plots of ANOVA models applied on non-transformed (upper row) and log-transformed data (lower row). Note the improved normal fit on the QQplot and homogeneity of variances after transformation (Residuals vs. Fitted and Scale-Location plots).

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 the difference being 0). We also consider different means – analysis on log-scale implies testing the geometric means on the original scale. The same applies to 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 the predictor. An example with log-transformation in linear regression is displayed in Fig. 10.3., 10.4. and Table 10.2.

Log-transformation is sometimes used 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 the logarithm used – natural and decadic logarithms are used most frequently. Just beware of being consistent in using the same logarithm throughout the analysis.

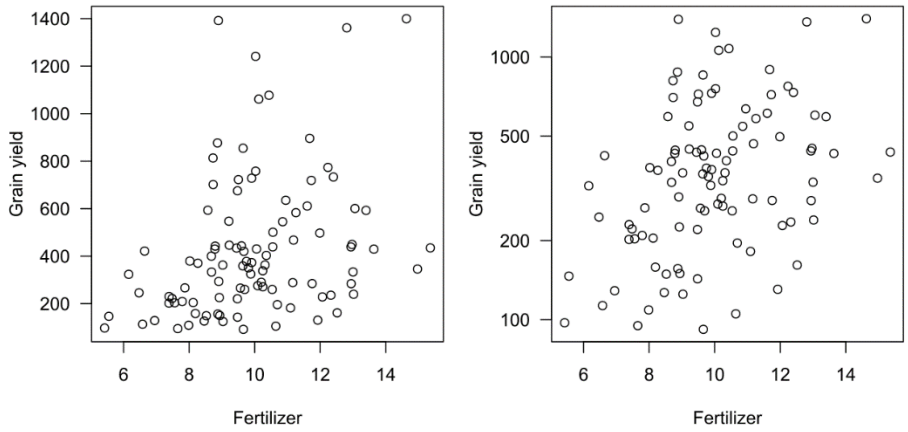

Fig. 10.3. Example of a regression with log-normal variable: how grain yield of maize depends on the amount of fertilizer applied. The left panel shows the scatterplot on the ordinary linear scale, while the right panel shows the same values on the log-scaled y-axis.

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

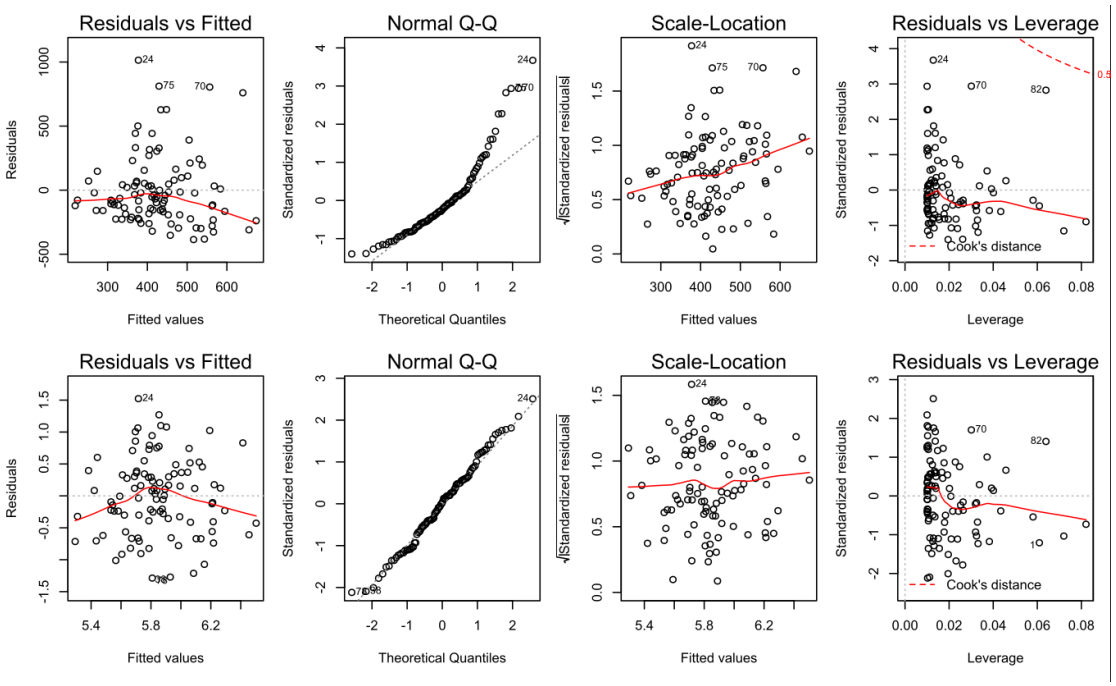

Fig. 10.4. Diagnostic plots of linear models fitted on non-transformed (upper row of plots) and log-transformed data (lower row of plots). Note improved normal fit on the QQplot and improved homogeneity of variances after transformation (Scale-Location plot).

Non-parametric tests

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

Table 10.3. List of parametric tests and their non-parametric counterparts together with appropriate 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 valuable alternatives to parametric or non-parametric tests. First, a statistic of difference from the null hypothesis (between samples) is defined. That may be the 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 the response variable are repeatedly permuted (reshuffled), and the same statistic is computed in each permutation. The 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 functions require this parameter
       distribution=approximate(B=number of permutations)
       to be set; B is usually set to 999 or 9999.