

6.4: Answers to exercises

Correlation and linear models

Answer to the question of human traits. Inspect the data, load it and check the object:

Data is binary, so Kendall's correlation is most natural:

We will visualize correlation with `Pleid()`, one of advantages of it is to show which correlations are connected, grouped—so-called “correlation pleiads”:

(Look on the title page to see correlations. One pleiad, `CHIN`, `TONGUE` and `THUMB` is the most apparent.)

Answer to the question of the linear dependence between height and weight for the artificial data. Correlation is present but the dependence is weak (Figure 6.4.1):

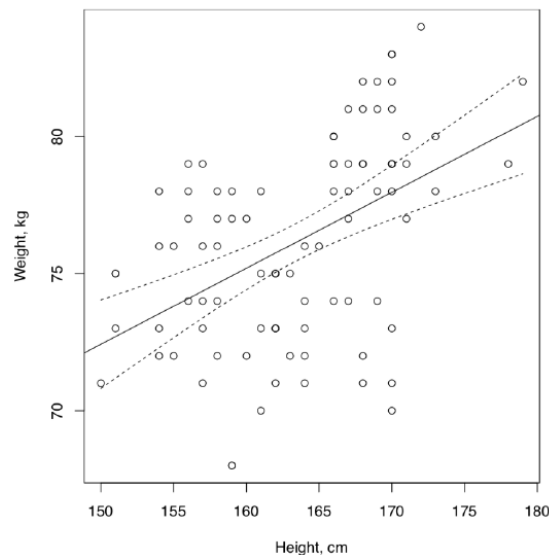


Figure 6.4.1 The dependence of weight from height (artificial data)

The conclusion about weak dependence was made because of low R-squared which means that predictor variable, height, does not explain much of the dependent variable, weight. In addition, many residuals are located outside of IQR. This is also easy to see on the plot where many data points are distant from the regression line and even from 95% confidence bands.

Answer to spring draba question. Check file, load and check the object:

Now, check normality and correlations with the appropriate method:

Therefore, `FRUIT.L` and `FRUIT.MAXW` are best candidates for linear model analysis. We will plot it first (Figure 6.4.2):

(`Points()` is a “single” variant of `PPoints()` from the above, and was used because there are multiple overlaid data points.)

Finally, check the linear model and assumptions:

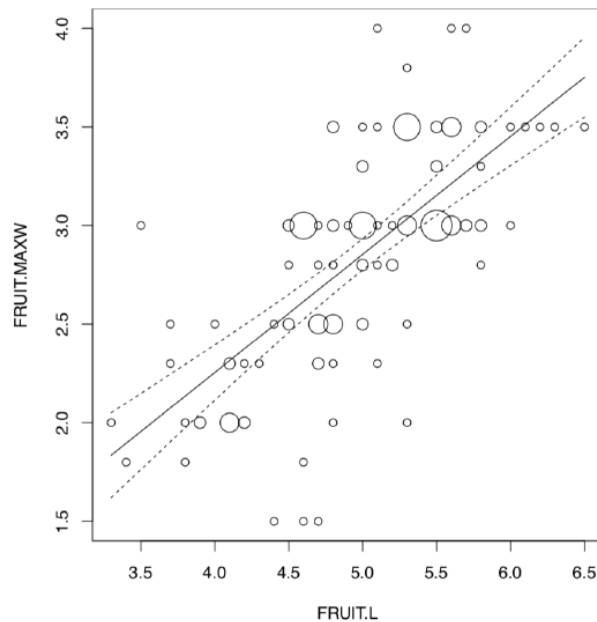


Figure 6.4.2 Linear relationship between fruit characteristics of spring draba.

There is a reliable model ($p\text{-value} < 2.2e-16$) which has a high R-squared value ($\sqrt{0.4651} = 0.6819824$). Slope coefficient is significant whereas intercept is not. Homogeneity of residuals is apparent, their normality is also out of question:

Answer to the heterostyly question. First, inspect the file, load the data and check it:

This is how to visualize the phenomenon of heterostyly for all data:

(Please review this plot yourself.)

Now we need to visualize linear relationships of question. There are many overlaid data points so the best way is to employ the `PPoints()` function (Figure 6.4.3):

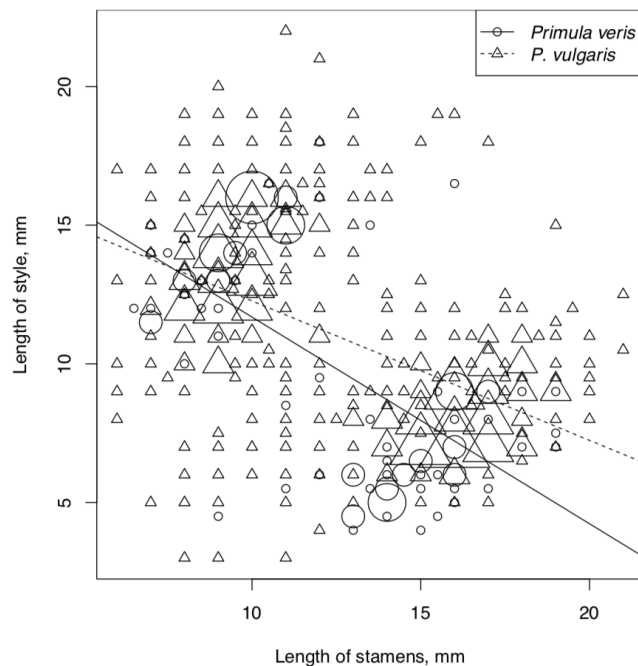


Figure 6.4.3 Linear relationships within flowers of two primrose species. Heterostyly is visible as two dense “clouds” of data points.

Now to the models. We will assume that length of stamens is the independent variable. Explore, check assumptions and AIC for the full model:

Reduced (additive) model:

Full model is better, most likely because of strong interactions. To check interactions graphically is possible also with the *interaction plot* which will treat independent variable as factor:

This technical plot (check it yourself) shows the reliable differences between lines of different species. This differences are bigger when stamens are longer. This plot is more suitable for the complex ANOVA but as you see, works also for linear models.

Answer to the question about sundew (*Drosera*) populations. First, inspect the file, then load it and check the structure of object:

Since we are required to calculate correlation, check the normality first:

Well, to this data we can apply only nonparametric methods:

(Note that "pairwise" was employed, there are many NAs.)

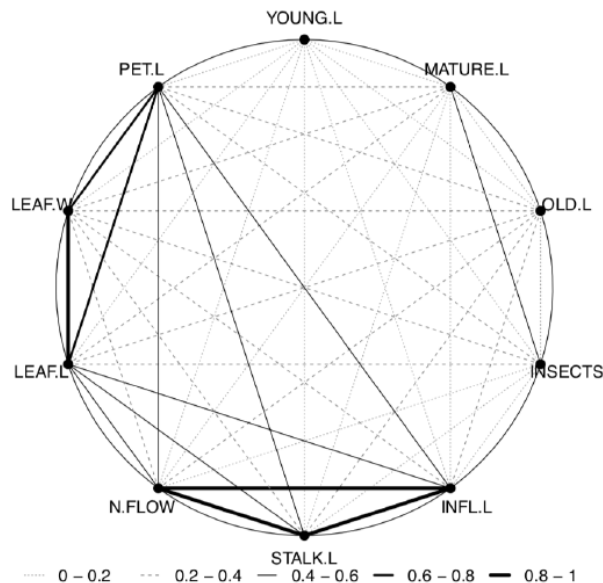


Figure 6.4.4 Correlations in sundew data.

The last plot (Figure 6.4.4) shows two most important correlation pleiads: one related with leaf size, and another—with inflorescence.

Since we know now which characters are most correlated, proceed to linear model. Since in the development of sundews stalk formed first, let us accept `STALK.L` as independent variable (influence), and `INFL.L` as dependent variable (response):

Reliable model with high R-squared. However, normality of residuals is not perfect (please check model plots yourself).

Now to the analysis of leaf length. Determine which three populations are largest and subset the data:

Now we need to plot them and check if there are visual differences:

Yes, they probably exist (please check the plot yourself.)

It is worth to look on similarity of ranges:

The robust range statistic, MAD (median absolute deviation) shows that variations are similar. We also ran the nonparametric analog of Bartlett test to see the statistical significance of this similarity. Yes, variances are statistically similar.

Since we have three populations to analyze, we will need something ANOVA-like, but nonparametric:

Yes, there is at least one population where leaf length is different from all others. To see which, we need a *post hoc*, pairwise test:

Population N1 is most divergent whereas Q1 is not really different from L.

Logistic regression

Answer to the question about demonstration of objects. We will go the same way as in the example about programmers. After loading data, we attach it for simplicity:

Check the model:

(Calling variables, we took into account the fact that R assign names like `V1`, `V2`, `V3` etc. to “anonymous” columns.)

As one can see, the model is significant. It means that some learning takes place within the experiment.

It is possible to represent the logistic model graphically (Figure 6.4.5):

We used `predict()` function to calculate probabilities of success for non-existent attempts, and also added small random noise with function `jitter()` to avoid the overlap.

Answer to the juniper questions. Check file, load it, check the object:

Analyze morphological and ecological characters graphically (Figure 6.4.6):

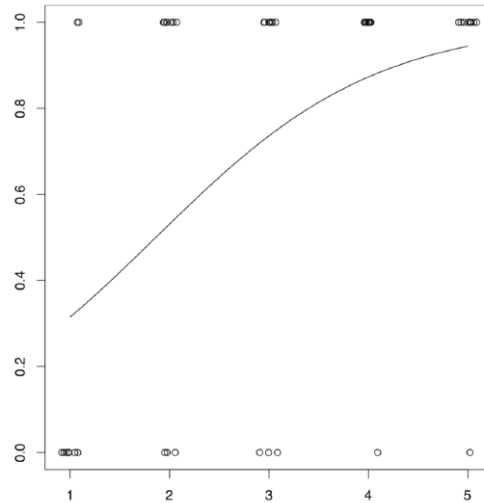


Figure 6.4.5 Graphical representation of the logistic model.

Now plot length of needles against location (Figure 6.4.7):

(As you see, spine plot works with measurement data.)

Since there is a measurement character and several locations, the most appropriate is ANOVA-like approach. We need to check assumptions first:

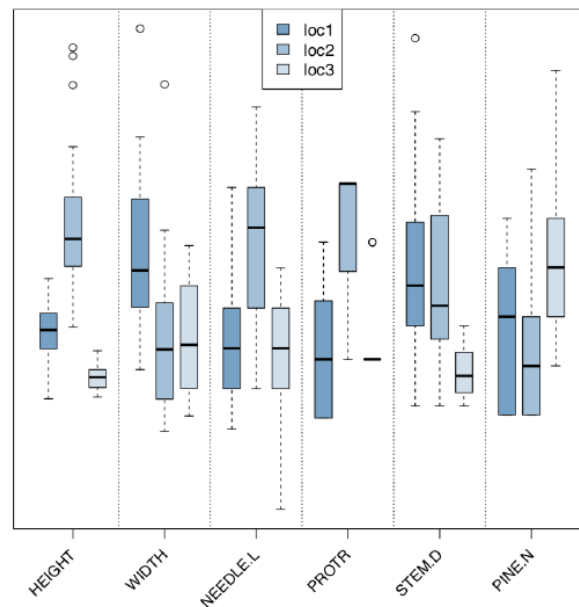


Figure 6.4.6 Boxplots show distribution of measurements among juniper populations. Since variation is not homogeneous, one-way test with post hoc **pairwise t-test is the best**:

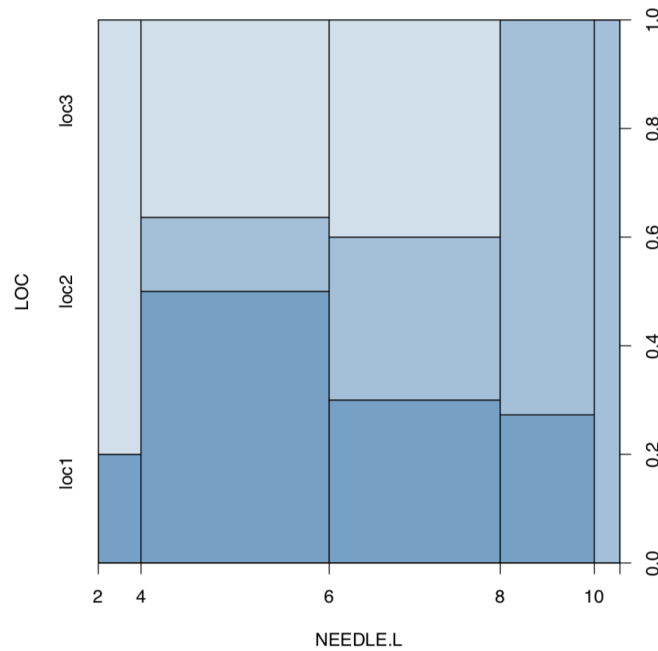


Figure 6.4.7 Spine plot: locality vs. needle length of junipers.

(Note how we calculated eta-squared, the effect size of ANOVA. As you see, this could be done through linear model.)

There is significant difference between the second and two other locations.

And to the second problem. First, we make new variable based on *logical expression* of character differences:

There are both “species” in the data. Now, we plot conditional density and analyze logistic regression:

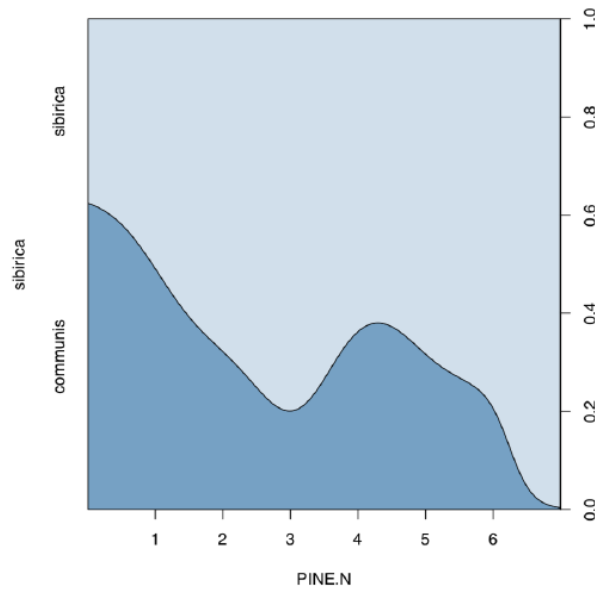


Figure 6.4.8 Conditional density of being *Juniperus sibirica* with the presence of some pine trees.

Conditional density plot (Figure 6.4.8) shows an apparent tendency, and model summary outputs significance for slope coefficient. On the next page, there is a table (Table 6.4.1) with a key which could help to choose the right inferential method if you know number of samples and type of the data.

Type of data	One variable	Two variables	Many variables

Type of data	One variable	Two variables	Many variables
Measurement, normally distributed	t-test	<i>Difference</i> : t-test (paired and non-paired), F-test (scale) <i>Effect size</i> : Cohen's d, Lyubishchev's K <i>Relation</i> : correlation, linear models	Linear models, ANOVA, one-way test, Bartlett test (scale) <i>Post hoc</i> : pairwise-test, Tukey HSD <i>Effect size</i> : R-squared
Measurement and ranked	Wilcoxon test, Shapiro-Wilk test	<i>Difference</i> : Wilcoxon test (paired and non-paired), sign test, robust rank order test, Ansari-Bradley test (scale) <i>Effect size</i> : Cliff's delta, Lyubishchev's K <i>Relation</i> : nonparametric correlation	Linear models, LOESS, Kruskal-Wallis test, Friedman test, Fligner-Killeen test (scale) <i>Post hoc</i> : pairwise Wilcoxon test, pairwise robust rank order test <i>Effect size</i> : R-squared
Categorical	One sample test of proportions, goodness-of-fit test	<i>Association</i> : Chi-squared test, Fisher's exact test, test of proportions, G-test, McNemar's test (paired) <i>Effect size</i> : Cramer's V, Tschuprow's T, odds ratio	<i>Association</i> tests (see on the left); generalized linear models of binomial family (= logistic regression) <i>Post hoc</i> : pairwise table test

Table 6.4.1 Key to the most important inferential statistical methods (except multivariate). After you narrow the search with couple of methods, proceed to the main text.

This page titled 6.4: Answers to exercises is shared under a [Public Domain](#) license and was authored, remixed, and/or curated by Alexey Shipunov via source content that was edited to the style and standards of the LibreTexts platform.