8. Linear models

"Linear model" is an umbrella term for models that relate two sets of random variables with linear relations. One set of variables is called *response variables* and they are to be predicted by the model from the second set of variables, called *predictor variables* or sometimes *covariates*. The standard approach is to find linear combinations of predictors that result in a response. With very many variables in either group, this becomes a typical task for various machine learning algorithms, which is beyond the scope of this course.

Both the predictors and the covariates can be compositional, in which case compositional analysis must be applied rather than standard methods. A special case is of particular relevance for metagenomic data analysis, namely when the response is a composition and the covariate is a categorical variable. This situation arises, for instance, if we take a number of blood samples from a group of people and, after sequencing, extract the metagenomes. These are the compositional responses. The categorical variable used as a covariate could be the gender of the persons. The linear model would try to predict which genetic elements could be predicted by the covariate. In this case, the answer would be the abundance of the Y chromosome. Inthis situation, ANOVA is a popular choice for a linear model. In this chapter, we will deal with the simple case of a single predictor yielding a response. In reality, samples are described by a large number of covariates, and multivariate methods need to be applied.

8.1 Linear regression with compositional response

A common type of linear model is regression. However, as we have seen many times so far, we cannot do linear regression directly on compositional data because the method of least squares, which is the typical way to fit the regressions, is based implicitly on the Euclidean metric. The assumption of linear regression is that the data deviates randomly from a mean model, and the method of least squares seeks to determine the mean model by minimizing the residuals between the model and the data. These residuals, the sum of squares, are the

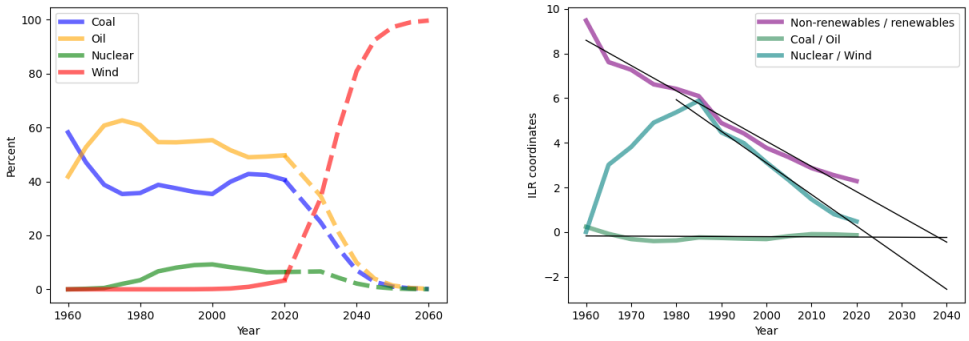

Figure 8.1: The left panel shows a 4-part sub-composition of the world's energy production between 1960 and 2020. The dashed lines extending beyond 2020 are linear extrapolations. The right panel shows the ILR coordinates on a basis where oil and gas (non-renewables) are split from nuclear and wind (renewables). The black lines are linear regression models of the ILR coordinates.

(squared) distances between the model and the data.

As usual, we can choose to implement a least squares algorithm using the Aitchison distance, or we can choose to transform our data into $\mathbb{R}^{D-1}$ and perform the model fitting using the normal least squares method.

For a practical application of compositional linear regression, we consider the world's energy production from four sources, coal, oil, nuclear, and wind, between 1960 and 2020. We close the 4-part sub-composition to 100, and the result is shown in the left panel of Fig. 8.1 as solid colored lines. It is difficult to tell by looking at the proportions whether the parts evolve linearly in time. It is also clear that if we tried to do linear regression directly on the proportions, this would lead to a nonsensical (and non-closed) result.

Instead we build a binary partition and use it to form an orthonormal basis to be used for ILR transforming the data. The partition we chose is the following,

	Coal	Oil	Nuclear	Wind
v_1	1	1	-1	-1
v_2	1	-1	0	0
v_3	0	0	1	-1

(8.1)

where we have chosen to split coal and oil from nuclear and wind (non-renewables from renewables; it is obviously debatable whether nuclear can be considered a renewable), and then split coal and oil and finally nuclear and wind. With this partition, we ILR transform the data, and the resulting coordinates are plotted in the right panel in Fig. 8.1. It can clearly be seen that the first two ILR coordinates are well approximated by linear functions over the entire period. The third coordinate (nuclear over wind), however, is clearly not linear over the entire period. We therefore only use the data from 1980 to 2020 to do the regression on that coordinate.

A regular least squares fit can be done on the ILR coordinates, and the resulting models are shown as thin black lines on top of the data. Notice how the fit to the third coordinate only covers the data from 1980 and onward. The regression models can be

extrapolated into the future, and we can do an inverse ILR transformation of these future values back to the simplex to obtain predicted proportions in the future. This is shown as dashed line segments in the left panel of Fig. 8.1. There are a few caveats, though. This extrapolation predicts that wind will reach 100 percent by 2060, but remember that this is in the sub-composition of only these four energy sources and the fact that nuclear has actually been declining since the 1990s while wind has grown tremendously over the same period, something which is unlikely to continue at the same rate over the next 40 years.

It should also be noted that this extrapolation depends on our choice of basis. If we had split our energy sources differently, our predicted energy budget would be different.

8.2 Analysis of Variance (ANOVA)

Statistical models that are used to analyze the difference between variation within a group and variation between groups are collectively known as ANOVA. ANOVA originated in evolutionary biology, where it was used to test whether variation between two groups of animals were larger than the variation within each group and use the result to identify new species. Testing two groups of responses that are separated by a categorical covariate is one obvious application of ANOVA, making it highly applicable for discovering effects within a set of metagenomic samples.

8.2.1 Hypothesis testing

The core of ANOVA is the concept of hypothesis testing. The idea behind hypothesis testing is that an expectation is made, typically expecting no effect, in which case it is called a null hypothesis, and a statistical test is conducted to see if a sample deviates significantly from the null hypothesis. If it does, the null hypothesis is rejected, and if it is not, the premise of the null hypothesis is true (from a statistical point of view).

In applications of ANOVA that are typically relevant for genomics, two sets of samples are tested against each other, and the null hypothesis is that the two sets of samples are identical. For example, consider a number of patients who are infected with a bacterial pathogen. The patients are divided into two groups, and one group is given an antibiotic agent. Afterwards, samples are taken from all patients, and they are analyzed for the presence of the pathogen. Our null hypothesis is that the drug has no effect, and that the pathogen will be present at equal levels in both groups. However, after a statistical test, we may find that the patient group that received the drug has a significantly reduced level of infection, in which case we reject the null hypothesis and conclude that the drug has an effect.

In order to make the decision to reject or accept the null hypothesis, we need to establish a significance level α. The significance level determines how certain we are of our decision. The statistical test returns a so-called p-value, and if $p < \alpha$ we reject the null hypothesis. The default value for α is 5%, which means that out of 100 tests, only 95 will result in a correct decision, or in other words, we can be 95% certain that our decision is correct.

Two types of errors occur in hypothesis testing: (I) rejecting a true hypothesis (true negative) and (II) accepting a false hypothesis (false positive). The rate at which these errors occur depends on the value of α. If set too high, we get many type II errors, and if set too low, we get many type I errors. Finally, just because a hypothesis is rejected,

doesn't mean that the opposite is the only possible hypothesis or even the best, and that is why we reject hypotheses rather than accept them.

8.2.2 Student's t-test

The most common form of testing differences between two normal sample distributions is the Student's t-test. This test determines if there is a significant shift in the sample means, given the variance of the two sets of samples. If the variances of the two distributions differ, the test is referred to as Welch's t-test or unpaired t-tests.

In its simplest form, the test between two distributions, $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by

$$t = \sqrt{n}\frac{\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2}{\sqrt{\mathsf{var}(\mathbf{x}_1) + \mathsf{var}(\mathbf{x}_2)}}, \tag{8.2}$$

where the bars indicate the distribution's (arithmetic) mean and n is the total number of samples in the two distributions. The t value can be converted into a p value (probability of the null hypothesis being wrong) using the t-distribution,

$$f(t,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{8.3}$$

where ν are the degrees of freedom, $\nu = n_1 + n_2 - 2$. t tests are implemented in most programming languages, and they provide both the t and the p values.

8.2.3 F-test

Sometimes it may be useful to test whether the variances between two parts are significantly different (obviously before scaling with the total variance). The procedure is similar, but in this case we use the F-distribution and the corresponding F-test. The F-test has the form

$$F = \frac{\sum_{i=1}^{K} n_i(\bar{Y}_i - \bar{Y})^2/(K-1)}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2/(N-K)}, \tag{8.4}$$

where the nominator is the between-group variability and the denominator is the within-group variability. $\bar{Y}_i$ is the mean of the i'th group with n_i samples, and K is the number of groups. N is the total number of samples.

When there are only two groups ($K = 2$), $F = t^2$ where t is the t-test statistic.

8.3 ANOVA with compositional response

In the following, we will consider a model where the response is compositional and the covariates are categorical variables. The goal of the ANOVA is, in this case, to test if there is a significant difference in the compositional centers between different categories of the covariates.

A compositional ANOVA model can be written as

$$\hat{\mathbf{x}} = \beta_1 \oplus (I(z=2)) \odot \beta_2 \oplus \ldots \oplus (I(z=K) \odot \beta_K), \qquad \mathbf{x} \ominus \hat{\mathbf{x}} = \varepsilon, \tag{8.5}$$

for K different categories. The indicator function $I(z=k)$ equals 1 when the condition is true and 0 otherwise. ε is the compositional residual of the model.

ANOVA methods are typically implemented to work on real coordinates, so the compositional approach is to select a basis, ilr transform the compositions, and apply ANOVA on the resulting coordinates. Afterwards, coefficients can be inversely transformed to return to the compositional coefficients.

In chapter 6, we worked through an example based on protein consumption in the early 1980's in Europe. We will now revisit that data set and carry out an ANOVA to look for food sources that contribute significantly different amounts of protein consumption between eastern and western Europe. The first step is to ILR transform the data, and therefore to choose a basis. In this case, it is important how that basis is chosen because it may affect the outcome of the ANOVA. Therefore, if we have previous knowledge of associations between food sources and country location, for instance, from the PCA or from sub-compositional analysis, our basis should be informed by this. Prior knowledge, however, is not always available, so we will proceed as if no such information were available.

In this case, we start by picking an arbitrary hierarchical binary partition, which we normalize to form an orthonormal basis. From this basis, we can calculate ILR coordinates.

	RM	WM	E	M	F	C	S	N	V
v_1	-1	+1	0	0	0	0	0	0	0
v_2	-1	-1	+1	0	0	0	0	0	0
v_3	-1	-1	-1	+1	0	0	0	0	0
v_4	-1	-1	-1	-1	+1	0	0	0	0
v_5	-1	-1	-1	-1	-1	+1	0	0	0
v_6	-1	-1	-1	-1	-1	-1	+1	0	0
v_7	-1	-1	-1	-1	-1	-1	-1	+1	0
v_8	-1	-1	-1	-1	-1	-1	-1	-1	+1

It is now up to the analyst to pick an ANOVA model and fit the two coefficients that correspond to the two categorical variables, north-south and east-west. In this example, we chose a double-sided Student's t test with equal variance in the two groups. This is appropriate if we power the data by the inverse of the total variance, like we did in Chapter 6.

	v_1	v_1	v_3	v_4	v_5	v_6	v_7	v_8
β_1	-0.40	-0.13	-0.06	0.31	0.73	-0.25	1.28	0.51
β_2	-0.22	0.26	0.05	1.07	-0.81	-0.09	-0.59	-0.10

The coefficients come out in ILR coordinates, which may be slightly inconvenient, but, just like we can go from CLR to ILR by taking the dot product with the transposed basis (eq. 3.3.4), we can go from ILR to CLR by taking the dot product with the un-transposed basis. Thus, we can easily express our fitted coefficient in CLR values.

	RM	WM	E	M	F	C	S	N	V
clrβ_1	-0.04	-0.61	-0.48	-0.45	-0.05	0.47	-0.46	1.14	0.48
clrβ_2	0.05	-0.26	0.21	0.06	1.21	-0.63	0.01	-0.54	-0.09

These values are easily interpreted from their signs. The coefficient clrβ_2 corresponds to the east-west category, and food items with a positive sign are consumed more in the

west, while coefficients with a negative sign describe food that is consumed more in the east.

The problem with this result is that conclusions based on CLR values are not sub-compositionally coherent because the geometric mean moves, in general, when a sub-composition is considered, and therefore, this result is only valid for the full composition (which is a sub-composition in itself). Another problem is that we can't tell how significant this result is because we can't provide a p-value on individual items against each other. We must use this result as a guide to form another basis on which we can do statistical testing. We do this by grouping together components with similar CLR values. If we sort $\text{clr}\beta_2$ we get

	C	N	WM	V	S	RM	M	E	F
$\text{clr}\beta_2$	-0.63	-0.54	-0.26	-0.09	0.01	0.05	0.06	0.21	1.21

We group the items two by two, so that cereals and nuts, white meat and vegetables, red meat and starch, milk and eggs, and finally fish on its own. Our new tailored basis looks like this

	RM	WM	E	M	F	C	S	N	V
v_1	0	-1	0	0	0	0	0	0	+1
v_2	-1	0	0	0	0	0	+1	0	0
v_3	0	0	-1	+1	0	0	0	0	0
v_4	0	0	0	0	0	-1	0	+1	0
v_5	-1	0	+1	+1	0	0	-1	0	0
v_6	+1	-1	+1	+1	0	0	+1	0	-1
v_7	+1	+1	+1	+1	0	-1	+1	-1	+1
v_8	+1	+1	+1	+1	-1	+1	+1	+1	+1

We proceed with a t test based ANOVA to get the statistics of the coordinates in the tailored basis.

	v_1	v_1	v_3	v_4	v_5	v_6	v_7	v_8
β	0.02	0.01	-0.11	<0.01	0.14	0.35	0.88	-1.29
t value	0.10	0.06	-0.92	0.03	1.05	1.90	2.66	-4.36
$P(>\|t\|)$	0.92	0.96	0.37	0.98	0.31	0.07	0.01	<0.001

If we decide on a confidence interval of 95%, we can remove all coefficients with p-values greater than 0.05. This leaves us with only two balances: fish versus everything else and cereal-nuts versus everything else but fish. The sign of the t-values determines the direction of the overconsumption, which means that fish is consumed more in the west than in the east, while cereals and nuts are consumed more in the east. There is no significant difference in the consumption of meat, dairy, starch, and vegetables between the east and west.

8.3.1 Effect plot

Effect plots are a way to visualize the result of an ANOVA. In essence, an effect plot is the variance between groups plotted against the variance within the groups. Parts (or groups) are plotted as points and can be colored by p-value to show which ones are significant. Figure 8.2 shows the protein data in an effect plot, where we have simply taken the variances on the CLR values.

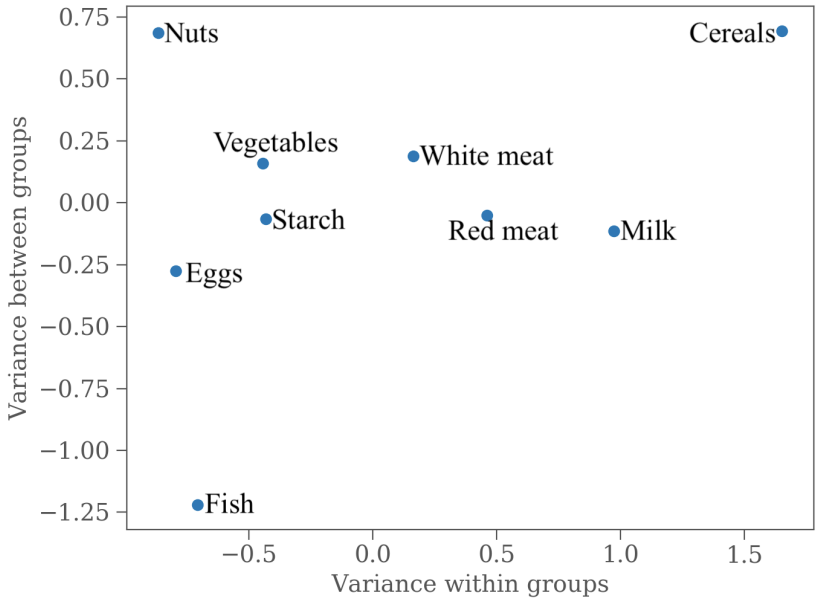

Figure 8.2: Effect plot of the protein data.

8.4 Exercises

Exercise 8.1 Repeat the example in Sect. 8.1 of linear extrapolation of energy production using a different sequential binary partition, specifically the one where nuclear is considered non-renewable:

	Coal	Oil	Nuclear	Wind
v_1	+1	+1	+1	-1
v_2	+1	+1	-1	0
v_3	+1	-1	0	0

The data can be found in the file `world_energy.csv`. ■

Exercise 8.2 Calculate the sample centers for the nutrition data separately for eastern and western countries. Calculate the perturbation difference between the two sample center compositions. Do an inverse CLR transform of the first set of $clr\beta_2$ coefficients in Sect. 8.3. Compare the resulting β_2s to the perturbation difference. The data can be found in the file `protein.csv`. ■

Exercise 8.3 Repeat the ANOVA example for the nutrition data, but check for differences between north/south instead of east/west.

Hint: start from the clrβ table. Sort the parts by clrβ_1 and build a basis on similar parts. ■

Exercise 8.4 Redefine "northern" and "southern" countries based on significant differences only within animal products (meats, eggs, milk, and fish). This exercise has no unique solution. ■