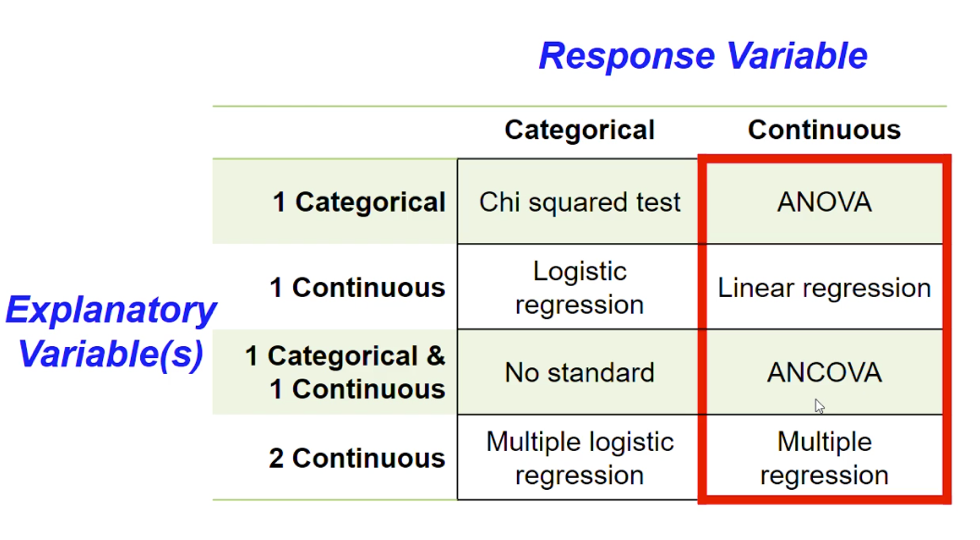
**PS: Linear models, generalised linear models and assessing model fit**



* P value can facilitate the conversation
* R2: proportion of variation explained by variable

Text, letter

Description automatically generated

ANOVA:

* Assumption: equal amount of variance
* Assumption: all observations are independent
* The variances of the residuals in different groups are the same
* The residuals are normally distributed

🡪 want to use the adjusted R-squared

ANCOVA: general linear model using one continuous and one categorical

🡪 interpreting X-Y Scatterplots: two groups within the data: adding categorical variable by grouping

Chart, scatter chart

Description automatically generated

🡪 avoid using multiple explanatory variables that are strongly correlated with each other: multicollinearity.

🡪 use priori inspection of correlations or post-hoc metrics: variance inflation factors

🡪 use prior knowledge of the system: what is the ecological basis

Assessing distribution of data:

* Normally distributed data
* Poisson distribution: skewed distribution because you don’t have negative values: count data
* Bionomical Data: presence/ absence data

Normal distribution:

Text, letter

Description automatically generated

Poisson distribution:

Text, letter

Description automatically generated

Bernoulli distribution: binomial data

Text, letter

Description automatically generated

**Assessing model fit:**

Akaike Information Criterion (AIC):

Logo

Description automatically generated with medium confidence

Text

Description automatically generated

🡪 AIC values are relative to the data being considered. They cannot be considered in an absolute sense like R-squared values. One compares different models with identical response data to assess which model is best.

🡪 some conventional interpretations: difference in AIC between models:

A close-up of a document

Description automatically generated with low confidence

🡪 also, AICc, BIC, SIC, DIC: other information criterion

**In person lecture**

Poisson: non-continuous, often zero-inflated, whole number (represented by integers)

Tutorial 1:

The anova() function gives you an ANOVA table, with all its constituent components, including the results of an F test. The summary function gives you all of the above and more. It first gives you the formula, followed by some summaries of the distribution of residuals. Then you get a table for the coefficients, which includes the estimated value of the coefficients, the standard errors around those estimates, a t-value for how far each coefficient departs from zero, and a test for significance. Don’t worry too much about those significance tests right now. Lastly, at the bottom, one gets information about the residual standard error (which essentially measures how much of the variation in your data is not explained by your statistical model), degrees of freedom (here, total number of data points – number of categories), R2 values, and the main results of an ANOVA, which tests whether fertiliser type significantly affects seedling height.

Next you want to determine if the data satisfy the assumptions necessary to conduct an ANOVA. You will need to use the function resid() to obtain the residuals, while the function shapiro.test() can be used to test the normality of the residuals, and the function bartlett.test() can be used to check equality of variances.

The first plot you see is the residual versus fitted plot. This lets you readily assess if you have constant variances. If your dependent variable responds to your independent variable in a non-linear manner, that will also be evident here (particularly relevant for continuous predictor variables).

We may be interested in statistically assessing which fertiliser(s) are driving our significant ANOVA result. To do this, we can perform a Tukey’s test using the TukeyHSD() function in R.

**Exercise 2: More on Analyses with Single Categorical Explanatory Variable and Continuous Response**

How could we solve the problem that the assumptions of the ANOVA may have been violated? (**HINT:** Log-transforming data often helps when the assumptions are violated). Re-conduct the analyses once you have transformed the data and assess if the data now follow the assumptions of an ANOVA.

**Exercise 3: Relationships between Continuous Variables**

This exercise focuses on whether base saturation and clay show relationships with soil pH.

**Exercise 4: Non-parametric Alternatives to Simple Linear Models**

We now have a significant result! But, we still have some issues of non-normality, heteroscedasticity, and an outlier. These are difficult data! In such times, we might resort to a non-parametric statistical test. Let’s try using a Kruskal-Wallis test.

> kruskal.test(Sodium~ Habitat,data=soils)

What do you infer based on whether the Spearman’s rho is positive or negative?

How do we compare different statistical models, which incorporate different explanatory variables, to decide which model is ‘best’? One sensible approach is to build a model that incorporates all explanatory variables that we think are important to explaining the response variable. If some are insignificant, they will not have much ‘impact’ on the model.

Another common approach is to simply include whichever variables seems to be significant in univariate models using an ANOVA framework or t-tests on the associated parameters (to determine if they are significantly different from zero). While this is not really the approach I recommend, building univariate models and plotting out relationships between pairs of variables is one of the most straightforward ways to come to understand your data. The more time you spend with a given dataset, the better will be your analysis.

Yet another approach is to build the most complex model possible and then to exclude non-significant variables. However, in large datasets sometimes a variable (or parameter) will be ‘significant’, but it actually does not increase the explanatory value of the overall model too much. With linear models or general linear models, the r squared value (or better, the adjusted R squared value) can give you an idea of ‘model fit’, or how well the explanatory variables explain the response variable. One can remove variables from this complex model one at a time to determine which removals cause the r squared value to drop a lot. Those variables should stay in the model. If the r squared does not drop at all when a variable is removed, then there is no reason not to remove it.

🡪 So, it looks like the model with the interaction term is the best model. However, we cannot forget to evaluate the models using the plot function or otherwise. There are some MAJOR non-normality and heteroscedasticity issues here. Thus, we cannot really trust these results. How might you solve these issues? One would have to re-run these various models once a ‘fix’ has been tried (e.g. transforming the response variable). Also, once you come up with your final model, think about what the results actually mean. What does it mean if there is a significant interaction between river basin and habitat type? Also, in terms of our discussion above about over-fitting, we are approaching ‘danger territory’ here in trying to estimate so many parameters with only 19 data points (now that we have excluded some of the river basins).

**Assessing Model Fit using AIC and Constructing Generalised Linear Models**

**Exercise 1: Exploring Model Fit**

When one has many potential explanatory variables that could be included in an analysis, there are diverse ways of deciding which explanatory variables to include. One approach is to simply include whichever variables seem to be significant using an ANOVA framework or t-tests on the associated parameters (to determine if they are significantly different from zero).

However, in large datasets sometimes a variable (or parameter) will be ‘significant’, but it actually does not increase the explanatory value of the model too much. With linear models or general linear models, the R squared value (or better, the adjusted R squared value) can give you an idea of ‘model fit’, or how well the explanatory variables explain the response variable.

An alternative means for comparing statistical models is an information theoretic approach. The workhorse of such an approach is the Akaike Information Criterion (AIC) and its derivatives (e.g., AICc). The AIC is derived from the probability of the model, i.e. its likelihood. Let’s start off by comparing the various models we generated in the soil pH example with AIC, which can also be used for linear or general linear models.

An additional point: it is often useful to construct a null model when looking at AIC values, in order to assess if even your simple models are worth considering. This is a quick ‘sniff test’ of whether you have any ability to explain your response variable with the data you have available. To do so in R, you put the number ‘1’ in the explanatory variables part of the model (more on this below).

**Exercise 2: Generalised Linear Models (GLMs) with a Poisson Response**

Then there is the residual degrees of freedom which is the number of sample units minus the number of parameters we have estimated.

We have estimated two parameters in this case. Note that we have not estimated a variance (or error), because Poisson assumes the variance to be the same as the mean. But check:

* If the variance is greater than the mean in your data/model, then you have what is called overdispersion
* This can quickly be checked here by dividing the residual deviance (also given) by the residual degrees of freedom
* If this value is greater than one, then you have overdispersion.

🡪 In this case, we obtain a ratio greater than two, which indicates a fair bit of overdispersion and that we should consider other models (e.g. a negative binomial response).

You will see that R has actually performed some statistical tests on whether the estimated coefficients differ significantly from zero. However, these particular statistical tests assume that your Poisson model was appropriate (i.e., assume a dispersion ‘ratio’ of ~1). This is not the case, so the results of these tests should perhaps not be trusted. In any case, I urge you to use an information theoretic approach to test if this poisson model with habitat fits the data better than a null model. And this brings us onto null models in R. You can create a null model by only estimating the intercept, and this is done as follows:

🡪 Given that both habitat and soil pH seem to explain variation in abundance of Inga thibaudiana better than a null model, it is sensible to construct multivariate models with both of them and to check how that performs.

> mod3 <- glm(thibaudiana~Habitat + Soil\_pH,data=combo,family=poisson)

> mod4 <- glm(thibaudiana~Habitat \* Soil\_pH,data=combo,family=poisson)

> AIC(mod\_null,mod1,mod2,mod3,mod4)

What we see here is that including both variables results in a better model than including just one. However, including an interaction term actually worsens the model fit (gives a higher AIC value). What this means is that while including the interaction term may have increased the probability of our model slightly (which lowers the log-likelihood), the ‘penalty’ of adding the extra parameter (which results in adding two to the log-likelihood to calculate AIC) outweighed any benefit from including it

**Overdispersion:** Thus, we still have some overdispersion, which means that the data are more variable than predicted by our model.

* One approach would be to use a different distribution for the response variable (e.g., a negative binomial).
* Another approach could be to simplify the data to presence/absence data
* Another approach is to plot the predicted versus observed values to see how well they match.
* Plot out the raw data and see how our predicted values compare to it, in order to fully understand our analysis.

**Exercise 3: Generalised Linear Models (GLMs) with a Binomial Response**

**🡪** As mentioned above, one way to simplify noisy abundance data is to do presence/absence analyses.

In order to see if soil pH significantly affects the presence vs. absence of Inga auristellae, we should compare this model to a null model.

Now, let’s compare this to some more models, with habitat and other variables.

Also, in terms of further reading, if you want to try negative binomial models to deal with overdispersion issues, you can install the MASS library and check the glm.nb() function. We will cover how to install optional libraries next week