BIOL 422 & 680, Pedro Peres-Neto, Biology, Concordia University

ANOVA, Regression and types of sum-of-squares (version 2022)

The General Linear Model (not to be confused with generalized linear models) includes a multitude of statistical models (all linear) such as t-test, ANOVA (analysis of variance), regression (simple and multiple), ANCOVA (analysis of covariance) and any other (linear) approach based on normality, homoscedastic and sample independence assumptions. The goal of all these analyses is to model (linearly) the variation of a response variable (or multiple response variables) against a predictor variable (or a set of multiple predictors). The response (dependent) variable is always continuous and the predictors can be categorical (e.g., t-test, ANOVA for comparing group means), continuous (e.g., regression) and a mix of both (e.g., multiple regression, ANCOVA). The goal of general linear models is to model the variation in the response variable as a function of the predictor variables. Note that transformations (e.g., rank) and other approaches to deal with non-normality, homoscedasticity (e.g., weighted least squares; mixed-models) and sampling independency (e.g., generalized least squares) can be performed in all analyses under the family of general linear models.

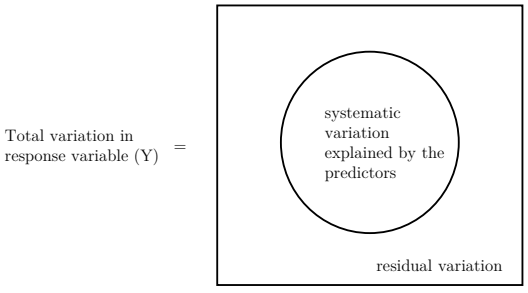
The General Linear Model analyzes (models) the total variation in the response variable by splitting it into two general components (or sources of variation): (1) systematic variation that is explained by the predictor(s); and (2) residual variation that is not explained by the predictors. The F-statistic is calculated as the ratio between the systematic variation and the residual variation. Depending on the type of analysis (e.g., ANOVA for contrasting group means and ANOVA for regression), the systematic and residual components will be calculated in different ways. Note, however, that they both represent the same types of variation (i.e., systematic = explained by predictors & residual = non-explained by residuals). Note each type of variation is corrected by their own degrees of freedom. Remember that the F-statistic has two types of degrees of freedom. If the F statistic is large (for a given set of degrees of freedom), then the systematic (explained) variation is large in contrast to the residuals (unexplained) variation. Two relevant questions arise here. The first one is whether the variation explained by the systematic variation is significantly greater than the residual variation; the second question is what is the percentage of the total

variation in the response variable explained by the systematic variation. The second question is often asked in regression analysis but not in ANOVAs and ANCOVAs (a mistake in my mind).

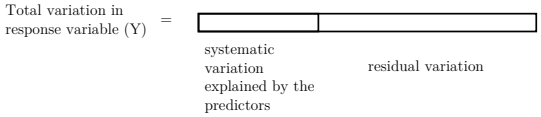
ANOVA. In the case of ANOVA for comparing means among groups, the systematic variation is commonly referred "variation between groups" and the residual variation as "variation within groups" (groups here refer to groups of individuals). The term "variation between groups" in ANOVA is used because the categorical predictors serve to separate variation among groups in terms of their mean values (i.e., how much group means vary among each other). Sometimes, in ANOVA, the residual variation is also referred as to random variation. This reference is due to the fact that ANOVA is often used in experiments where it is assumed that the variation within groups (i.e., not the result of the experimental factors imposed to the groups) is due to random chance.

Regression. In the case of regression, the goal is to model the variation in the response variable as a function of the variation of the predictors. The systematic variation in regression is commonly referred as "regression variation" and the residual variation as "residual error".

Summarizing, the total variation in ANOVA is divided into between and within components; and in regression the total variation is divided into regression and residual. This variation is often graphically expressed in one of the two forms:



the size of the circle represents the percentage of the total variation due to the systematic variation



the size of the rectangle represents the percentage of the total variation due to the systematic variation

The total variation in the response variable as well as the systematic and residual sources of variation are based on sum of squares (by the way, we often hyphenate it as sum-of-squares) that vary according to the analysis (e.g., ANOVA comparing means versus regression). Moreover, as we will see below, there are different types of sum-of-squares even for the same type of analysis (e.g., ANOVAs). This issue is not well covered in standard Introduction-level books on Biostatistics but

they need to be covered in advanced levels. Let's start by the total variation in the response variable Y in both ANOVA and regression is simply:

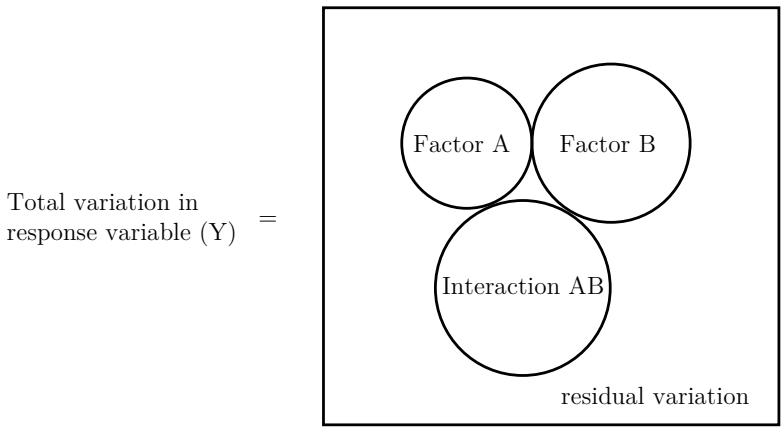
$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

where[1] TTS is the total sum of squares of a variable Y, n is the number of observations in Y and $\overline{Y}$ is the average of Y [[1] when explaining terms in an equation, the proper "mathematical etiquette" is that we start without a paragraph space with "where" in lower caps]. TTS is then divided (corrected) by its appropriate degrees of freedom (n – 1) given that we are subtracting each observation by the mean. Without correcting for the appropriate degrees of freedom, the sample-based TTS is a biased estimate of the true population TTS (this issue was covered in our lecture about degrees of freedom and variance estimation; remember that TTS is the numerator of the variance estimator). TTS/(n-1) becomes the total mean sum-of-squares (acronym is MST). So, MST is simply the variance of Y. Although the sum-of-squares will change for the systematic component(s) across different types of analyses (e.g., ANOVA contrasting means versus regression), they are the same for the total sum-of-squares (presented above) and residuals (not presented here). However, for each type of analysis, the systematic component(s) is based on appropriate sum-of-squares and their divisions (correction) by the appropriate degrees of freedom. Note that the systematic sum-of-squares plus the residual sum-of-squares equals the total sum-of-squares for almost all types of analyses (exceptions are very intricate analytical tools not covered in this course).

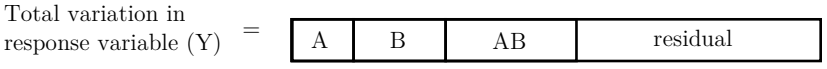
The systematic source of variation can be further divided into the relative importance of each predictor (different factors in ANOVA or different predictors in regression). Let's consider first the case of ANOVA. In the case of one-way ANOVA, there is only one factor but multiple predictors when we have more than two groups (see lecture on contrasts). In this case, the systematic variation of interest is the total variation across all contrasts (predictors) of the single factor (i.e., one-way ANOVA; represented by the first figure above).

Let's move on to a two-way factorial ANOVA with two factors A and B, and their interaction AB. In this case we have four sources of variation. Three sources are systematic (A, B and AB) and the fourth is the residual variation. In the standard way to compute the appropriate

sources of variation in an ANOVA, the design is assumed balanced (i.e., there is an equal number of observations for all possible combinations of levels). Consider the Sandberg et al. (2000) study on gene expression. Their study represents a balanced design because for each factor and associated levels (factor mice strain: 2 strains; factor brain region: 6 regions) they have exactly two observations. If one single out of any two observations for any combination of levels (e.g., the hippocampus of strain 129SvEv) would be missing, the design would be unbalanced. Note that the issues of balanced *versus* unbalanced design obviously only affects multi-factorial ANOVA designs and not one-way or single-factorial ANOVA designs). Let's consider first the case of a balanced design. In this case, all the contrasts (predictors) required to model the systematic variation would be fully orthogonal, i.e., the correlation between them would be zero. In this case, we can represent the variation as below. The circle representation of the variation due to the different components of is done via a Venn diagram. Either representation is often used in regression analysis.



the size of the circle represents the percentage of the total variation due to the systematic variation. Here, the variation of factor A is smaller than B which in turn is smaller than the interaction AB.



the size of rectangles represents the percentage of the total variation due to the systematic variation. Here, the variation of factor A is smaller than B which in turn is smaller than the interaction AB.

We can note a few things: 1) The total systematic variation is the variation of (A) + (B) + (AB); 2) That the overlap between each factor and their interaction is independent (i.e.,

orthogonal and as such, the circles and rectangles do not overlap). Because they are orthogonal, we say that they do not share any systematic variation regarding their contributions to the total variation in the response variable Y. A few additional (and somewhat relevant remarks) – even though a particular factor may not be statistically significant, we don't transform its contribution into zero (e.g., the contribution can be small and test lack statistical power to detect it). Finally, because these components are based on additive sum-of-squares (i.e., total sum-of-squares equals the systematic sum-of-squares + the residual sum-of-squares), each component can be translated (and often is) into a percentage of variation, which becomes a more comparable metric instead of the original values expressed as variance components. We will see this later in our module on multiple regressions.

Now, when designs are not balanced, we have an issue with the traditional way used to calculate systematic variation (i.e., the one assuming orthogonal designs). In this case, we say that the factors are "not fully orthogonal", i.e., they correlate and as such it would not be mathematically possible to estimate the independent variation of any factor and their interactions by using the calculation of sum-of-squares based on a balanced design. In this case, the circles (or rectangles) in the representation above would share some variation.

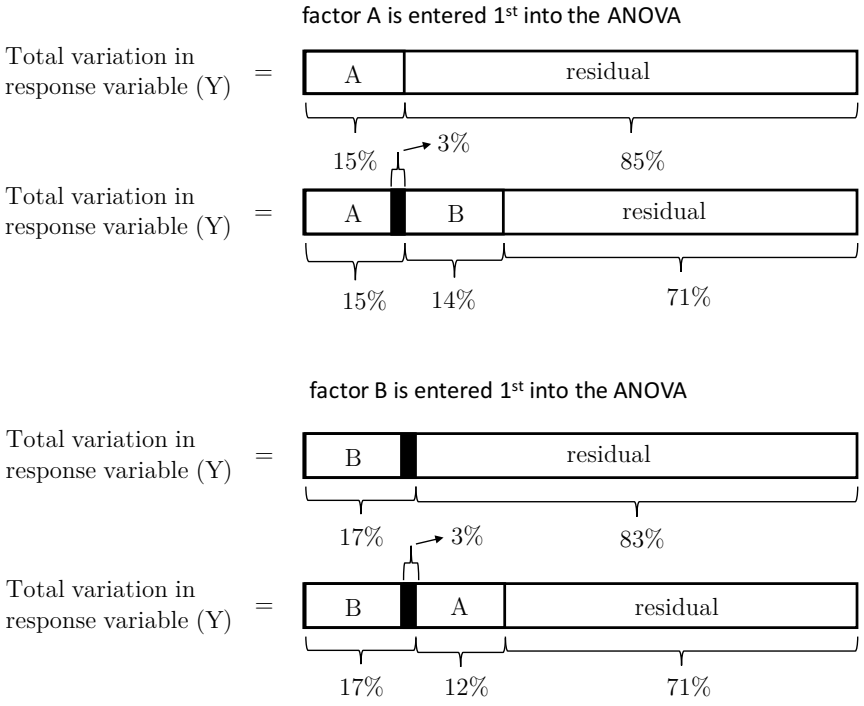
This is a very important notion that becomes even more relevant in regression analysis, particularly model selection, because unlike categorical predictors, continuous predictors are almost never uncorrelated (unless by design or mathematically built that way, e.g., principal component regression). One important note here. This discussion of sum-of-squares is often not done in regression because the ANOVA model applied to regression is simply based on the sum-of-squares of all combined systematic component (i.e., all regression predictors together) and not each systematic component separately, unless in some important steps such as model selection. We will go through this also in our multiple regression module.

Now we need to understand the consequences of calculating sum-of-squares for the relative contributions of each factor and interaction assuming a balanced ANOVA design for data that are not balanced, i.e., when the contribution of different factors and their interactions are not orthogonal. The name of the sum-of-squares for balanced ANOVA design is "Type I sum-of-squares"

(or simply Type I SS). Worth noting (again) is that neither the residual sum-of-squares or the total sum-of-squares are affected by the order or the total contribution of systematic errors in balanced and non-balanced standard ANOVA designs. Only the relative contributions of the systematic components (factors and interactions) are affected by unbalanced designs. The issue with using Type I SS for unbalanced designs (i.e., when factors or predictors are not orthogonal) is that the order of entrance of factors (or predictors) influences the calculation. For this reason, the Type I SS is known also as "sequential" sum-of-squares. Let's consider a small example. For simplicity, we will only consider the effects of two factors A and B, but not their interactions. The reason is that it would be tricky to do that graphically without explaining a lot of little additional details. But I'll explain this decision once you get the simple example; hopefully you'll easily understand the issue of complexity graphically. Also, instead of using absolute values for the sum-of-squares, we will use relative contributions (%) as discussed earlier. In this case, the total sum-of-squares becomes 100%.

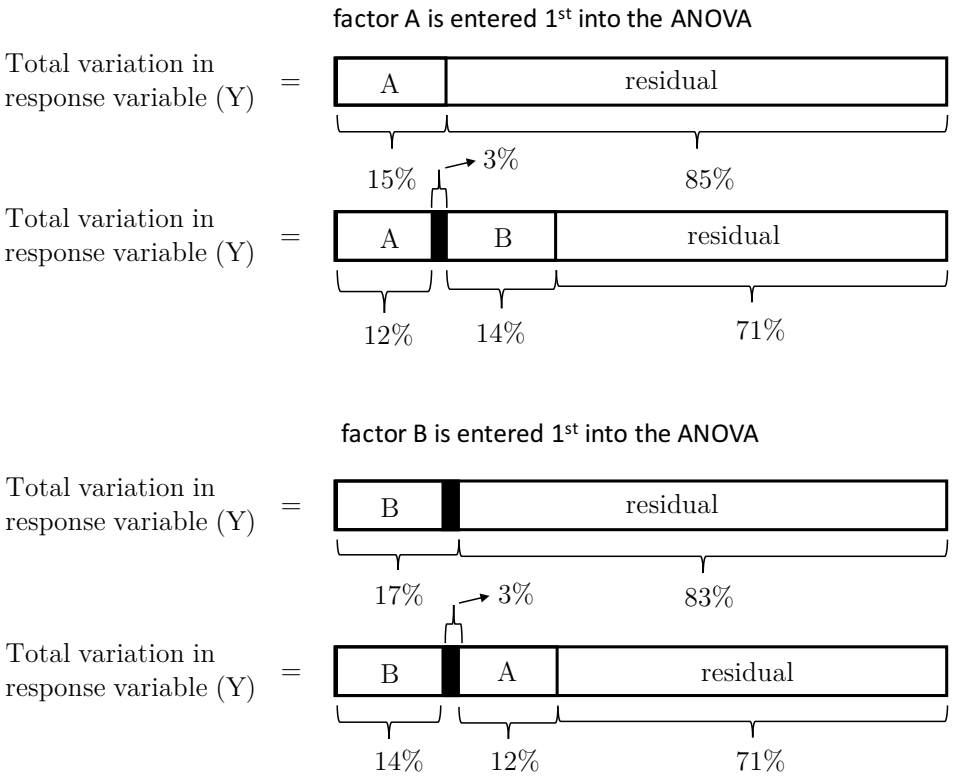
The next figure shows a contrast between the situation in which factor A is entered first in the ANOVA and then factor B, and the reverse situation (factor B entered first and then A). Let's say that when entering factor A first, it explains 15% of the total variation in Y (out of 100%). Factor B explains alone 17%, but when entered in the model, parts of this 17% overlaps with A (say 3%) because factors A and B are not orthogonal (i.e., they share variation in common that explains part of the variation in Y). In type I SS, the overlap fraction is kept as contribution of factor A. So, factor A remains at 15% and factor B contributes with 14% (17% - 3%). Now, let's consider the case where factor B enters first. In this case, factor B's contribution is kept at 17%. Then, factor A is entered in the model, but this time, the shared variation due to lack of orthogonality is kept with factor B. As such, factor A's contribution is 12% (i.e., 15% - 3%). Now you understand why the order of entrance changes the calculation of sum-of-squares of each factor (remember that the percentage of contribution is directly proportional to the sum-of-squares of each factor). It is also easier to understand why we call Type SS as "sequential". Note that in both cases, the residual variation left unaccounted for (unexplained) is 71% regardless of which factor (A or B) entered first in the model. Remember that we did not consider the interaction because the interaction in non-balanced designs will have overlapping fractions of their sum-of-squares with both A and B, making

it difficult to represent into rectangles. Venn diagrams facilitate expressing shared fractions among 3 factors, but not easily in terms of ordering entrance. As such, we kept it simple with just two factors. Using the Type I SS (below), the contributions of factors A and B are different depending on the order in which they are entered into the analysis (model).



Now let's consider a sum-of-squares for each factor is kept the same regardless of their order of entrance in the model. This type of sum-of-squares is called Type III SS (or marginal or orthogonal). We won't cover types II and IV sum-of-squares for the moment. The figure below shows what happens when we use Type III SS. In this way of calculating sum-of-squares, the shared fraction is not included in any of the factor's contribution. This lowers the sum-of-squares of all factors, but it keeps them consistent regardless of the entrance in the model. As such, the contributions are marginal (or orthogonal) to one another.

Using the Type III SS (below), the contributions of factors A and B are equal regardless of the order in which they are entered into the analysis (model).



As such, we should use the type III SS for unbalanced designs.