Analysis of Variance

Elise Arnaud elise.arnaud@univ-grenoble-alpes.fr

UGA Mosig

Table of contents

Different types of modeling

- Linear regression, which allows to explain a quantitative variable from quantitative explanatory variables (possibly also qualitative)
- The supervised classification, which allows to explain a qualitative variable from quantitative explanatory variables (possibly qualitative in addition). Attention, it must be distinguished from the unsupervised classification which is the clustering.
- **Analysis of variance, to analyze the influence of one or two qualitative explanatory variables on a quantitative variable.**

Analysis of variance

The aim here is to study the impact of a qualitative variable on a quantitative variable. we have wheat yields observed on 80 homogeneous and distant plots. Each of the 4 wheat varieties considered was planted on 10 plots with phytosanitary treatment and on 10 other plots without any treatment. The wheat yield on each of the plots was measured.

The dataset contains the following 3 variables:

- rdt: wheat yield (in quintals per hectare) ;
- ble : wheat variety (A, B, C, D) ;
- phyto: phytosanitary treatment (1 if positive, 0 otherwise).

Our objective here is to evaluate the sensitivity of the yield according to the wheat variety, and possibly according to the couple wheat variety-plant protection treatment.

We ask is the following: Does the wheat variety have an impact on the yield?

Analysis of variance

The model considered here is written in the following form:

$$y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$$

for $i \in \{1, \ldots, 4\}$ et $j \in \{1, \ldots, n_i\}$

In this model :

- y_{ij} is the yield of the plot j with the variety i;
- μ is the average yield;
- α_i depends only on the wheat variety.

It is assumed here that the plots are homogeneous and independent, and that the yield of variety i follows a normal distribution with mean $\mu + \alpha_i$ and variance σ^2 (identical variance for all the wheat varieties).

In the case under study, $\forall i \in \{1, \ldots, 4\} : n_i = 20$.

Analysis of variance

1. **one-factor analysis of variance**
   To find out if there is a variety effect, we will construct a statistical test whose null hypothesis is:

   $$H_0 : \alpha_1 = \ldots = \alpha_4 = 0$$

   This null hypothesis amounts to considering that the 4 varieties lead to an average yield equal to μ.

   - If this is the case, it means that the wheat variety has no impact on the yield.
   - If, on the contrary, one of them is non-zero, it means that the wheat variety has an effect on the yield.

2. **two-factor analysis of variance**
   We can also carry out this analysis by considering the influence of the wheat variety and the phytosanitary treatment

one factor ANOVA

We place ourselves in the more general case where the qualitative variable has I levels ($I = 4$ for the variety of wheat, in the case of study).

We consider that we have n_i observations for the i modality of the variable. This is referred to as an experimental design:

- complete if $\forall i \in \{1, \ldots, I\} : n_i > 0$;
- balanced if $n_1 = \ldots = n_I = r$.

In the case under study, the plan is balanced (therefore necessarily complete).

one factor ANOVA

The model is written:

$$y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$$

for $i \in \{1, \ldots, I\}$ et $j \in \{1, \ldots, n_i\}$ $\mu, \alpha_1, \ldots, \alpha_I$ are unknown parameters, and the $\varepsilon_{i,j}$ are independent r.v of law $\mathcal{N}\left(0, \sigma^2\right)$, where $\sigma^2 > 0$ is unknown
The following notations are considered hereafter:

$$y_{i,\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{i,j} \quad \text{(average on the modality } i \text{)}$$

$$y_{\cdot,\cdot} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{i,j} \quad \text{(overall average)}$$

one factor ANOVA

In order to solve the problem, an additional constraint must be imposed, for example :

- $\mu = 0$
- $\alpha_1 = 0$ (we can choose another cell than the first)
- $\sum_{i=1}^{I} n_i \alpha_i = 0$
- $\sum_{i=1}^{I} \alpha_i = 0$

The estimators of $(\mu, \alpha_1, \ldots, \alpha_I)$ are then :

- $\widehat{\mu} = 0; \forall i \in \{1, \ldots, I\} : \widehat{\alpha}_i = y_{i,}$
- $\widehat{\mu} = y_{1,.}; \widehat{\alpha}_1 = 0; \forall i \in \{2, \ldots, I\} : \widehat{\alpha}_i = y_{i,.} - y_1$
- $\widehat{\mu} = y_{.,.}; \forall i \in \{1, \ldots, I-1\} : \hat{\alpha}_i = y_{i,.} - y_{.,.}; \hat{\alpha}_I = \sum_{i=1}^{I-1} \frac{n_i \hat{\alpha}_i}{n_I}$
- $\hat{\mu} = \frac{1}{I} \sum_{i=1}^{I} y_{i,.}; \forall i \in \{1, \ldots, I-1\} : \hat{\alpha}_i = y_{i,.} - \frac{1}{I} \sum_{i=1}^{I} y_{i,.}; \hat{\alpha}_I = \sum_{i=1}^{I-1} \widehat{\alpha}_i$

The estimator of σ^2 is in all cases:

$$s^2 = \frac{1}{n-I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{i,j} - y_{i,.})^2$$

one factor ANOVA

- We are not so much interested in the estimation of the parameters μ, α_i, σ^2
- We are interesting in the ability to test a hypothesis H_0 such as the wheat variety has no effect, which translates statistically as:

$$H_0 : \alpha_1 = \ldots = \alpha_I = 0$$

- This hypothesis is all the more easily rejected as the means are different from one another. The test statistic used for this purpose is :

$$F = \frac{\mathrm{MSM}}{\mathrm{MSE}}$$

one factor ANOVA

$$F = \frac{\mathrm{MSM}}{\mathrm{MSE}}$$

where:

- Interclass variation: SSM (Sum of Squares of the Model)

$$\mathrm{SSM} = \sum_{i=1}^{I} n_i \left(y_{i,\cdot} - y_{\cdot,\cdot}\right)^2$$

- Intra-class variation : SSE (Sum of Squares of the Error)

$$\mathrm{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{i,j} - y_{i,\cdot}\right)^2$$

-

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{n - I} \qquad \mathrm{MSE} = \frac{\mathsf{SSE}}{I - 1}$$

It can also be shown that

$$\mathrm{SST} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{i,j} - y_{\cdot,\cdot}\right)^2 = SSR + SSE$$

Test the overall significance of the model

It can be shown that under H_0 :

$$F \sim \mathcal{F}(p-1, n-p)$$

- We decide to reject H_0 at the α test level if $f > f_{(p-1,n-p),1-\alpha}$.
- Here is the analysis of variance table:

Source	df	SC	MS	F	p-valeur
Model	$I-1$	SSM	MSM	$\frac{MSM}{MSE}$	$\mathbb{P}(\mathcal{F}(p-1, n-p) > f)$
Residuals	$n-I$	SSE	MSE		
Total	$n-1$	SST			

- we reject H_0 at the α test level if p-value $< \alpha$.
- In practice, rejecting H_0 is equivalent to declaring that the qualitative variable has a significant effect on our phenomenon (Y).

ANOVA and F-statistic

The ANOVA produces an F-statistic, the ratio of the variance calculated among the means to the variance within the samples.

- If the group means are drawn from populations with the same mean values, variance between the group means should be lower than the variance of the samples
- A higher ratio therefore implies that the samples were drawn from populations with different mean values

two-factor ANOVA

We now wish to study the influence of two qualitative factors A and B, with I and J modalities respectively, on a quantitative variable. We assume here that we have a balanced design (with r observations for each crossing of the factors). The model considered is the following:

$$y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \varepsilon_{i,j,k}$$

for $i \in \{1, \ldots, I\}, j \in \{1, \ldots, J\}$ et $k \in \{1, \ldots, r\}$

The μ, the α_i, the β_j and the $\gamma_{i,j}$ are unknown parameters, and the $\varepsilon_{i,j,k}$ are independent r.v. of distribution $\mathcal{N}\left(0, \sigma^2\right)$, where $\sigma^2 > 0$ is unknown.

two-factor ANOVA

We consider the following quantities (averages: global, by modality on A and on B):

$$y_{i,j,} = \frac{1}{r} \sum_{k=1}^{r} y_{i,j,k}$$

$$y_{i,\cdot,\cdot} = \frac{1}{Jr} \sum_{j=1}^{J} \sum_{k=1}^{r} y_{i,j,k}$$

$$y_{\cdot j,\cdot} = \frac{1}{Ir} \sum_{i=1}^{I} \sum_{k=1}^{r} y_{i,j,k}$$

$$y_{\cdot,\cdot,} = \frac{1}{IJr} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{r} y_{i,j,k}$$

two-factor ANOVA

The tests carried out are the following:

$$H_0^A : \alpha_i = 0, \forall i \in \{1, \dots, I\}$$
$$H_0^B : \beta_j = 0, \forall j \in \{1, \dots, J\}$$
$$H_0^{AB} : \gamma_{ij} = 0, \forall i \in \{1, \dots, I\}, \forall j \in \{1, \dots, J\}$$

They allow respectively to test the influence of the factor A, the factor B and the interaction of the factors A and B.

two-factor ANOVA

We consider the following quantities:

$$\text{SSM}_A = Jr \sum_{i=1}^{I} (y_{i,\cdot,\cdot} - y_{\cdot,\cdot,})^2$$

$$\text{SSM}_B = \text{Ir} \sum_{j=1}^{J} (y_{\cdot j,\cdot} - y_{\cdot,\cdot,\cdot})^2$$

$$\text{SSM}_{AB} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{r} (y_{i,j,\cdot} - y_{i,\cdot,\cdot} - y_{\cdot j,\cdot} + y_{\cdot,\cdot,\cdot})^2$$

$$\text{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{r} (y_{i,j,k} - y_{i,j,})^2$$

$$\text{SST} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{r} (y_{i,j,k} - y_{\cdot,\cdot,\cdot})^2$$

two-factor ANOVA

The results are classically presented in the form of a table:

Source	df	SS	MS	F	p-value
$\text{Model}_{\mathrm{A}}$	$I-1$	$\mathrm{SSM}_{\mathrm{A}}$	MSM_A	$\frac{\mathrm{MSE}_A}{\mathrm{CMR}}$	influence of A
$\text{Model}_{\mathrm{B}}$	$J-1$	$\mathrm{SSM}_{\mathrm{B}}$	MSM_B	$\frac{\mathrm{MSE}_B}{\mathrm{CMR}}$	influence of B
$\text{Model}_{\mathrm{AB}}$	$(I-1)(J-1)$	$\mathrm{SSM}_{\mathrm{AB}}$	MSM_{AB}	$\frac{\mathrm{MSM}_{AB}}{\mathrm{MSE}}$	influence of AB
residuals	$n-IJ$	SSE	MSE		
T	$n-1$	SST			

This is how the analysis of variance works in principle.

Exercice

Now, open up your favorite code program and we'll perform an analysis of variance to understand what influences wheat yields.

Conclusion

- linear regression : effect of quantitative explanatory variables on a quantitative variable
- anova : effect of qualitative explanatory variables on a quantitative variable

- effect on a qualitative variable : classification or logistic regression

Conclusion

- Possibility to combine quantitative and qualitative explanatory variables (exemple)
- lm(Y∼ sex+age)

$$Y_i = \alpha_{sex_i} + \beta X_{age_i}$$

- lm(Y ∼ sex*age)

$$Y_i = \alpha_{sex_i} + \beta_{sex_i} X_{age_i}$$

Conclusion

1. You need a model to perform your regression
2. You need to check whether the underlying hypothesis of this model are reasonable or not

This model will allow you to:

1. Assess and quantify the effect of parameters on the response
   - Parameters are estimated as a whole, using **all** the measurements
2. Extrapolate within the range of parameters you tried
3. Detect outstanding points (those with a high residual and/or with a high lever)

This model will guide on how to design your experiments:

- e.g., the linear model assumes some uniformity of interest over the parameter space range
- if your system is heteroscedastic, you should perform more measurements for parameters that lead to higher variance