Comments on Two-ways Gaussian Models

R. Labouriau

November 2020

Here I present a range of general comments two-ways Gaussian models.

(1) On the description of Gaussian two-ways classification models - Here I explain how we can specify two-way classification models. These models arise naturally when there are two classification factors in plain, as I explain below using a fictive scenario.

Consider a situation where we compare the responses for three treatments, say A, B and C. Suppose that those treatments are applied for two types of individuals, say type x and type y (*e.g.*, females and males, or young and old). We then have six possible combinations the two classification factors as expressed in the representation below. Suppose, moreover, that we have two repetitions of each of the six combinations (*i.e.*, we apply the treatment A to four individuals, two of type x and two of type y, four individuals receive treatment B, and so on). The way to describe this experiment mathematically is to create a symbol for each of the 12 results of the experiment. We will use a capital letter, say Y, with three subindices indicating the treatment, type and the replication. The first subindex will indicate the treatment and is indicated by the letter t (so t can take the values A, B and C), the second index points to the type of individual and is represented by the letter i (so i is equal to x or y), and the last index, represented by r indicates the replicate (r is equal to 1 or 2). Using this convention we represent the response of the first and the second replicate of the individuals of type x that received the treatment A by Y_{Ax1} and Y_{Ax2}, respectively. See the table below representing all the possibilities.

	A rep. 1	A rep. 2	B rep. 1	B rep. 2	C rep. 1	C rep. 2
Type x	Y_{Ax1}	Y_{Ax2}	Y_{Bx1}	Y_{Bx2}	Y_{Cx1}	Y_{Cx2}
Type y	Y_{Ay1}	Y_{Ay2}	Y_{By1}	Y_{By2}	Y_{Cy1}	Y_{Cy2}

We write then (using the mathematical slang):
"The random variable Y_{tip} represents the response of the r^{th} replicate ($r = 1, 2$) of the individuals of type i ($i = x$ or y) that received the t^{th} treatment ($t = A, B$, or C)."
In this way, we established a basic notation that allow us to describe a statistical model. For example, we might want to say that the random variable that represents the response of the second replicate ($r = 2$) of the individuals of type x ($i = x$) that received the treatment A ($t = A$) is normally distributed with expectation μ and variance σ^2. You can see from the text above that this way of describing (part of a model) is too wordy and is definitely not practical. Now, using the mathematical notation discussed in the course the same idea can be expressed in a tighter way, we can just write "$Y_{AI2} \sim N(\mu, \sigma^2)$".

We can also express what happens in the entire experiment is a very neat way if we write:
"According to the model, for $t = A, B, C$, $i = x, y$ and $r = 1, 2$,

$$Y_{tir} \sim N(\mu_{ti}, \sigma^2) .$$

Moreover, we assume that the random variables $Y_{Ax1}, Y_{Ax2}, \ldots Y_{Cy2}$ are independent". Note that the text in red above completely describes a Gaussian two-ways classification model with interaction (or effect modification).

Here are two immediate consequences of the definition given above: the expectation and the variances of the responses are given by $E(Y_{tir}) = \mu_{ti}$ and $Var(Y_{tir}) = \sigma^2$. That is, according to this model the expectations are represented as in the table below.

	A rep. 1	A rep. 2	B rep. 1	B rep. 2	Crep. 1	C rep. 2
Type x	μ_{Ax}	μ_{Ax}	μ_{Bx}	μ_{Bx}	μ_{Cx}	μ_{Cx}
Type y	μ_{Ay}	μ_{Ay}	μ_{By}	μ_{By}	μ_{Cy}	μ_{Cy}

In the next section we will introduce a model (the additive model) that defines another pattern in the scheme above.

(2) Testing the effect of a type of individual - Here I introduce a model that represents the situation where there are no differences between the type of individuals. We will assume that the observations are independent, normally distributed, have the same variance and have the expectations depending on the treatment but not on the type. According to this model, the following pattern is formed in the representation of the expected values:

	A rep. 1	A rep. 2	B rep. 1	B rep. 2	Crep. 1	C rep. 2
Type x	μ_A	μ_A	μ_B	μ_B	μ_C	μ_C
Type y	μ_A	μ_A	μ_B	μ_B	μ_C	μ_C

Note that the two rows in the table above are equal. We informally say then that ”there is *no* effect of the type of individual”.

We can describe this model using the mathematical notation (*i.e.*, mathematical slang) by writing ”According to the model, for $t = A, B, C$, $i = x, y$ and $r = 1, 2$,

$$Y_{tir} \sim N(\mu_t, \sigma^2)\,.$$

Moreover, we assume that the random variables $Y_{Ax1}, Y_{Ax2}, \ldots Y_{Cy2}$ are independent”.

Please compare the definition above with the definition of the model described in the last section. You will see that the only change is in the subindex of the expectation (μ_{ti} becomes μ_t), a tiny change in the notation, but huge change in the model.

Note that the scenario described by this model (absence of effect of type) can be represented by the last model (by making $\mu_{Ax} = \mu_{Ay}$, $\mu_{Bx} = \mu_{By}$ and, $\mu_{Cx} = \mu_{Cy}$). In order to test whether there are differences between the individual types we use an F-test to test whether one can reduce the model in the last paragraph to the model described here.

(3) Testing the effect of treatment - Here I introduce a model that represents the situation where there are no differences between the treatments. We will assume that the observations are independent, normally distributed, have the same variance and have the expectations depending on the treatment but not on the type. According to this model, the following pattern is formed in the representation of the expected values:

	A rep. 1	A rep. 2	B rep. 1	B rep. 2	Crep. 1	C rep. 2
Type x	μ_x	μ_x	μ_x	μ_x	μ_x	μ_x
Type y	μ_y	μ_y	μ_y	μ_y	μ_y	μ_y

Note that the two columns in the table above are equal. We informally say then that ”there is *no* effect of the type of treatment”.

We can describe this model using the mathematical notation (*i.e.*, mathematical slang) by writing:

”According to the model, for $t = A, B, C$, $i = x, y$ and $r = 1, 2$,

$$Y_{tir} \sim N(\mu_I, \sigma^2)\,.$$

Moreover, we assume that the random variables $Y_{Ax1}, Y_{Ax2}, \ldots Y_{Cy2}$ are independent”.

In order to test whether there are differences between the treatments we use an F-test to test whether one can reduce the model in the first paragraph to the model described here.

(4) The Notions of Effect Modification (interaction) and Additivity
- The idea of additivity is central in this course as I explain below. In the previous sections I described (through a fictive example) three models: the first stating that the expectation of the response depend on the combination of the treatment and the type (there were then six expectations in play); the second and the third models described the situation were there were no effects of type (in section (2), with three different expectations) or no effect of treatment(in section (3), with only two expectations). There is, however, a situation of practical importance that we have not considered yet. Suppose that we have effect of both type and treatment, but the differences between the expected values of two different treatments are the same among the individual of type x and among individuals of type y. More precisely, suppose that the representation of the expected values for each observation takes the form below.

	A rep. 1	A rep. 2	B rep. 1	B rep. 2	Crep. 1	C rep. 2
Type x	$\tau_A + \beta_x$	$\tau_A + \beta_x$	$\tau_B + \beta_x$	$\tau_B + \beta_x$	$\tau_C + \beta_x$	$\tau_A C + \beta_x$
Type y	$\tau_A + \beta_y$	$\tau_A + \beta_y$	$\tau_B + \beta_y$	$\tau_B + \beta_y$	$\tau_C + \beta_y$	$\tau_A C + \beta_y$

That is, we assume that the expectation of a particular observation is a sum of a quantity depending only on the treatment (here τ_A, or τ_B or τ_C) and a quantity depending only on the type (here β_x or β_y). We write then,
"for $t = A, B, C$, $i = x, y$ and $r = 1, 2$, $E(Y_{tir}) = \tau_t + \beta_i$."

This model is called the additive model because the expectations are written sums. Note that the differences between the expectations of two treatments among the individuals of type x are the same as the differences of two two treatments among the individuals of type y.

To test whether there is additivity we compare the model described in section (1) with the additive model described here using an F-test. If we have additivity, then we might make a test for the effect of type (given that we have additivity) by comparing the additive model to the model described in section (2). Moreover, we might test the effect of treatments by comparing the additive model to the model described in section (3). These two last tests are preferable than the test comparing the model described in (1) to the model described in (2) or the model described in (3). This is the sequence of analysis that I used in my codes.

We can describe this model using the mathematical notation (*i.e.*, mathematical slang) by writing:

"According to the model, for $t = A, B, C$, $i = x, y$ and $r = 1, 2$,

$$Y_{tir} \sim N(\tau_t + \beta_i, \sigma^2)\,.$$

Moreover, we assume that the random variables $Y_{Ax1}, Y_{Ax2}, \ldots Y_{Cy2}$ are independent".

Additional remark: The most common notation to specify a model with interaction uses the symbol "*" connecting two classification factors. For example, in R the formula "Y cultivar * block" and in SAS the model statement "model Y = cultivar * block;" both specify a model with interaction between cultivar and block. This notation is NOT indicating the multiplication operation when using a two-ways classification model. The symbol "*" is just a syntactical symbol to specify the model. For instance, a completely equivalent way to specify the model above is to write "Y cultivar + block + cultivar :block", here we are not summing and dividing numbers. I have seen many examples of this misconception, therefore I write this remark.

(5) Using Models Based on Distributions Different than the Normal Distribution -

I discuss below the description of the exercise 4.4 (of Chapter 4). In this exercise a data on the abundance of worms in soil is studied. The total number of worms in samples of the same size of three different soils (termed $1, 2$ and 3) was determined using two different methods (labelled 1 and 2). The question is whether the abundance of worms differ for the three soil types and/or the determination method. The data-frame $Ex4.4$ contains the data of this exercise

As you might have realised, the counts of worms described in this assignment are **not** normally distributed. In fact, the counts are very well described by a Poisson distribution. Now, I will show that the technique for describing the Poisson models is very similar to the technique used to describe the models based on the normal distribution. To demonstrate that, I specify below a two-ways classification models similar to the model defined above, **but** using the Poisson distribution. You will probably realise that I used the old ”copy and paste” trick (with some obvious light edition).

”The random variable Y_{tip} represents the **counts** of the worms found in the r^{th} sample ($r = 1, \ldots, 20$) of the soils of type i ($i = 1, 2, 3$) using the t^{th} method ($t = 1$, or 2).” According to the model with *effect modification*, for $t = 1, 2$, $i = 1, \ldots, 20$ and $r = 1, \ldots, 20$,

$$Y_{tir} \sim Po(\mu_{ti}) \,.$$

We assume that the random variables $Y_{111}, Y_{112}, \ldots Y_{23\,20}$ are independent”.

The additive model is specified by stating that for $t = 1, 2$, $i = 1, \ldots, 20$ and $r = 1, \ldots, 20$,

$$Y_{tir} \sim Po(\tau_t + \beta_i) \,,$$

while the model without effect of method is specified by stating that for $t = 1, 2$, $i = 1, \ldots, 20$ and $r = 1, \ldots, 20$,

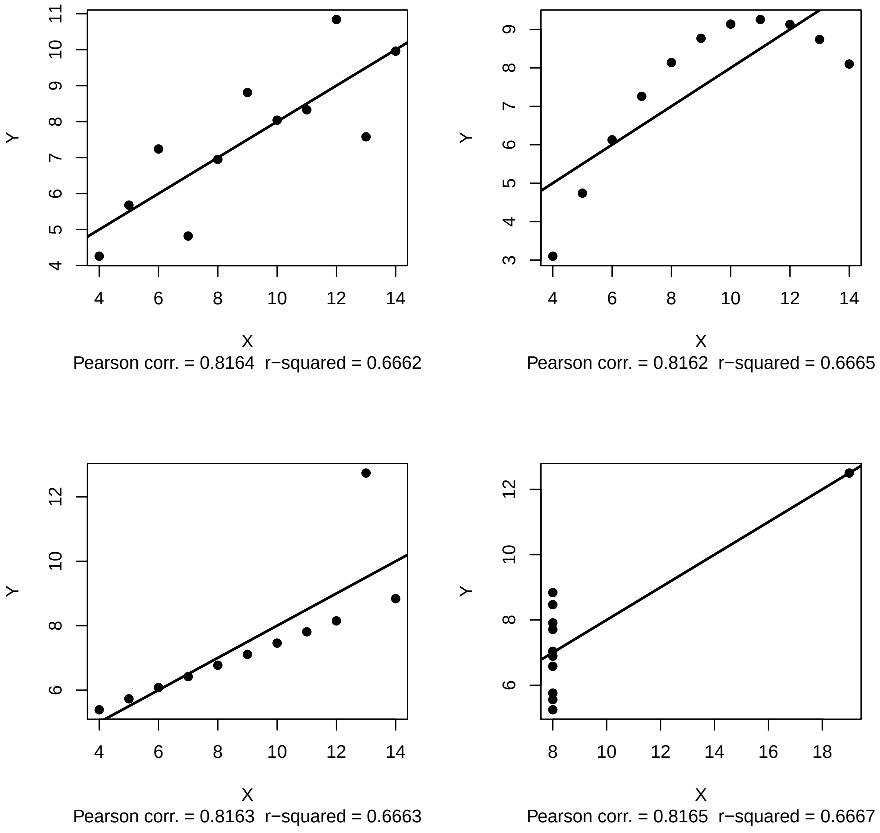
$$Y_{tir} \sim Po(\beta_i)$$

and so on ... .

Note that the patterns of expectations that arise with these are the same patterns described in the previous sections. Here we use the likelihood ratio test to compare (nested) models (instead of the F-test used in the normal models).

Additional remark: Note that it is not possible to specify a model based on the Poisson model using the classic formulation based on residuals. In the case of the normal distribution, writing ”$Y \sim N(\mu, \sigma^2)$” and writing ”$Y = \mu + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$” are completely equivalent methods for defining the models. This is not the case with the Poisson (and most of the distributions different than the normal distribution) because if we sum a constant to a Poisson distributed random variable, then what we get is a new random variable that is **not** Poisson distributed.

(6) On the R^2 coefficient - It is a common believe that models with high determination coefficients R^2 present a good fit. There are, however, many examples of models with reasonable R^2 that are clearly inadequate. One of these examples are the famous Anscombe Quartet with four examples of linear regressions that have the same relative high R^2 coefficients, one of the examples is reasonable, but the other three are inadequate in a caricatural way. See the tutorial on the anscombe quartet.



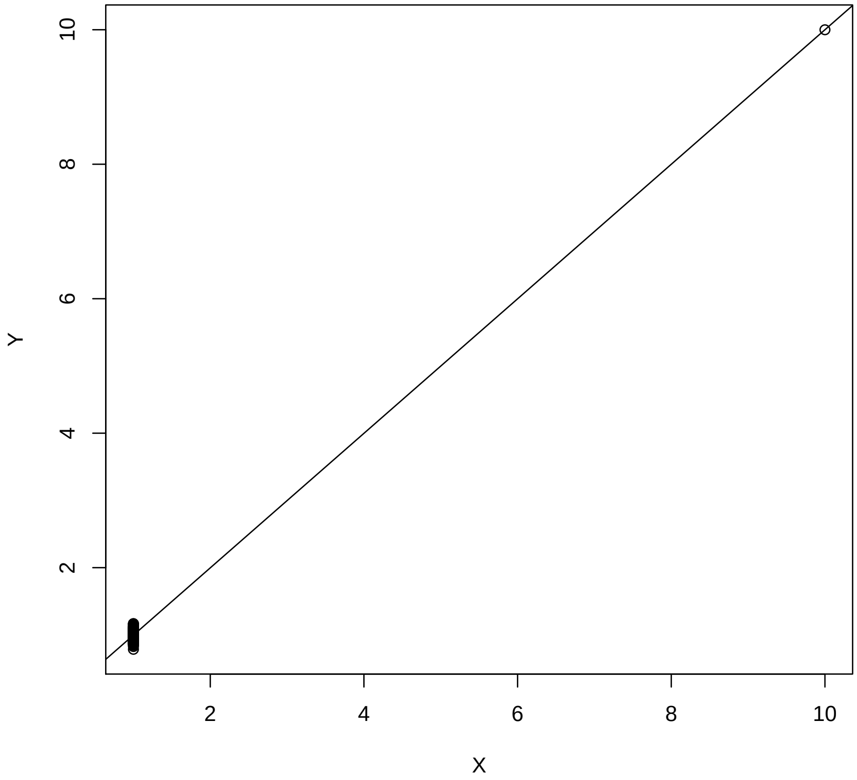
Pearson corr. = 0.8164 r−squared = 0.6662

Pearson corr. = 0.8162 r−squared = 0.6665

Pearson corr. = 0.8163 r−squared = 0.6663

Pearson corr. = 0.8165 r−squared = 0.6667

Observe also this regression that has an R^2 coefficient of 0.999.


10
8
6
4
2
Y
2
4
6
8
10
X