Family of linear regression mixture models stratified along the outcome[*]

R. Priam [†]

August 30, 2024

Abstract

Linear regression is one of the most studied model, it assumes a clear hypothesis of linearity. Underlying issues coming from Yule-Simpson's paradox or more generally hidden nonlinearities lead to spurious correlations difficult to detect in practice and prone to induce a mistaken linear model. The concern is when the model for explaining/predicting the outcome cannot be kept the same for the whole sample, it changes accordingly to the dependent variable. Hence, it is proposed a stratification of the outcome which leads to a new family of mixture models of regressions. A break or more along the outcome changes the linear regression into several components instead of one. A difference with the existing mixture models of regressions is that the partioning now depends mainly on the outcome. A double check of the change is obtained via an additional ordinal model and a discretization of the outcome. For the validation of the mixture, it is required a decrease of the bic, the aic and a mse or mae for both the continuous and discretized outcomes. Graphically, it is also shown these indicators plus the determination coefficient for moving thresholds in order to visualize the change between intervals of outcomes. With a threshold equal to the median, the approach is illustrated for several real datasets in the presented experiments. It is applied with a medical dataset from the Covid-19 lockdown in spring 2020.

1 Introduction

Regressions models [1] are often applied in many domains such as medecine, biostatistics or social sciences. The usual approach is to consider one unique regression for a whole data sample. The results are justified in the literature by the gaussianity of the residuals and often by some robutness of the regression against a misspecified model. A source of bias comes from noisy variables or outliers [2]. Another source comes from over-or-under-represented classes among the modalities of categorical variables. When the observations of tabular data are noisy or some modalities are missbalanced, the population values are not retrieved accurately by most of the current usual methods in practice. A solution is sometimes to add weights in order to robustify the fitting. Regression per groups such as gender or work occupation would allow to avoid this issue, but it would ask for a relevant method of selection of the similar variables across the groups in order to help the comparisons. Unfortunately, such methods may not exist currently and also the subsample sizes may be too small: this may explain why an unique regression is usually preferred in the literature despite that it is less informative. Another source of bias comes from nonlinearities which may be tackled either by transforming the variables either by partitioning the sample.

Herein, the bias of interest comes from the correlations between the outcome and the main explaining variables. They may be spurious for the regression such that the hypothesis of linearity is mistaken. In particular a break happens according to the outcome (response, target, dependent) variable instead of the independent (explaining, predictive) variables. Thus this problem is tackled in the next sections via a mixture approach because different regressions are fitted in each interval or group defined from the break(s). The case of continuous and ordered discretized outcomes are studied together and are seen complementary in order to check for a break. In the numerical experiments, the corresponding threshold is chosen equal to the median of the distribution for the outcome variable. To our knowledge, such approach was not adressed before in the literature despite its utility in practice. The mixture is for a target variable, either continuous either ordinal, where in both cases such model does not exist yet because usually the clusters are not according to only the outcome, on the contrary to mostly herein. The

[*] A version of this document is under review process but will be perhaps rejected if the contents is not relevant or not suitable. Reviewers have suggested improvement of the text at a first stage but this document is not yet accepted. This version may be the last one available.

[†] rpriam@gmail.com.

approach is different from the usual mixture models of regressions [3, 4] which do not propose a stratification of the observed outcomes y_i but a clustering of the observed pairs (x_i, y_i). Similarly the segmented (also called piecewise or broken-stick) regressions [5, 6, 7] are for breaks at the level of the independent variables x_i instead of the dependent variable y_i. In the literature it is not rare to find regression per groups, but not when the groups come from the outcome. Some exceptions are non generalized ways such as in the two-parts models [8] where only a part of the outcome is modeled on the contrary to herein.

Nextafter it is supposed that one, two (or more) component(s) are either or not relevant for a multiple regression according to the observed available sample. In a real example given at the next section, it is computed a weak correlation for a part of the distribution and nearly moderate one for another part of the distribution such that the overall correlation is increased in absolute value without inversion of the sign. Note that the idea of Berkson's bias or Yule-Simpson's paradox [9, 10, 11, 12, 13, 14] for regression may be related to the proposed mixture model along the outcome but this is not discussed herein. The proposed mixture is able to handle several cases of underlying groups by looking automatically for a change of the model which occurs along the outcome variable. Hence this is different from the usual change-point, random slope, mixture of regressions and mixture-of-experts or two-parts models in the current literature.

The plan of the paper is as follows. In a second section, after the introduction the main purpose is further presented and explained. In a third section, the proposed family of regression models is described. In a fourth section, the parameters inference is discussed after a summarizing table of the proposed family of models with four members having or not constraints. In a fifth section, the experiments confirm the interest of the models with several real datasets, while in a last section the conclusion is with perspectives.

2 Stratification for non linearities

In the case when the relation between the leading independent variables and the outcome variable is not linear, a stratification is proposed in order to model a break in the regression model. In this section, it is studied and illustrated a correlation from bivariate data when defined after partitioning the sample.

2.1 Segmented correlations from subsamples

With z the outcome, if the correlation between the variable z and a main variable x in the regression is spurious, typically a transformation such as x^2 or $log(x)$ is likely to be tried according to the shape of the scatterplot. This may be not always taken care of because for instance, a bias sample induced biased correlations or a linear relation remained a good approximation. An example of such situation is typically when for M_z a given threshold value, the (population or sample, as supposed identically denoted) correlations are as follows:

$$cor(x,z) = \begin{cases} \rho_{xz} & \text{for All } z \\ \rho_{xz}^{-} & \text{if } z \leq M_z \\ \rho_{xz}^{+} & \text{if } z > M_z \end{cases} .$$

Here, an example of hypothesis is that the correlation is weak for a part and weak or almost moderate [15] for the second part, such that for instance with same signs for the correlations,

$$|\rho_{xz}| \gg |\rho_{xz}^{+}| \gg |\rho_{xz}^{-}| .$$

Note that the relation is eventually polynomial or more generally nonlinear for one or both of the subsamples but with a different function. With this configuration, the resulting overall correlation may look like also moderated too. An example of such altered correlation is presented in the experiment part in table 6 at the experiments section and below in Table 1 in a dedicated subsection . For instance the main correlation is equal to $\rho_{xz} = -0.47$ for the full sample while equal to only $\rho_{xz}^{+} = -0.33$ and $\rho_{xz}^{-} = -0.20$ for the subsamples. With sample sizes enough large, such difference suggests that one unique model is expected to not fit the data as well as two distinct models with their own regression coefficients.

2.2 Correlations and conditional expectations

With x and z denoting the random variables generating the sample above, the correlation is usually writen like:

$$\rho_{xz} = \frac{\text{cov}(x,z)}{\sqrt{\text{cov}(x,x)}\sqrt{\text{cov}(z,z)}} .$$

Here, $\text{cov}(x, z)$ denotes the covariance c_{xz} with the mean expectations m_x and m_z, such that:

$$Cov \begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} c_{xx} c_{xz} \\ c_{zx} c_{zz} \end{bmatrix} = C_{xz} \text{ and } E \begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} m_x \\ m_z \end{bmatrix} = m_{xz} .$$

In the case when the sample is in two parts, one may write a mixture model in order to respect the true generating process, such that one gets two (or more) random variables. When $\square$ is for $+$ or $-$ this leads to conditional statistics,

$$Cov\begin{bmatrix}x^\square\\ z^\square\end{bmatrix} = C^\square_{xz} \text{ and } E\begin{bmatrix}x^\square\\ z^\square\end{bmatrix} = m^\square_{xz}\,.$$

Thus when B is the random variable for choosing between the two component, with expectation α_+, and with $\bar{m} = \alpha_+ m^+_{xz} + (1-\alpha_+)m^-_{xz}$ one may get:

$$\begin{aligned} Cov\begin{bmatrix}x\\ z\end{bmatrix} &= E_B\left(\mathrm{Cov}\left[\begin{matrix}x^\square\\ z^\square\end{matrix}\middle| B\right]\right) + \mathrm{Cov}_B\left(E\left[\begin{matrix}x^\square\\ z^\square\end{matrix}\middle| B\right]\right)\\ &= \alpha_+ C^+_{xz} + (1-\alpha_+)C^-_{xz}\\ &+ \alpha_+(m^+_{xz} - \bar{m}_{xz})(m^+_{xz} - \bar{m}_{xz})^T\\ &+ (1-\alpha_+)(m^-_{xz} - \bar{m}_{xz})(m^-_{xz} - \bar{m}_{xz})^T\,. \end{aligned}$$

This is also generalized into more components if required, see the literature on mixture models for a detailed proof. The correlation is not a simple function from the two components such as a weighted value: this may explain why it is increased here such that one needs to be very careful with the underlying hypothesis of linearity in regression as illustrated with the numerical examples in the dedicated (sub)section.

2.3 Correlations depending on two distances

The equality just above extends the Huygens one from a variance into a covariance. Nextafter, for empirical statistics from the sample, it is denoted s for the observed data sample, with:

$$s = \{(z_i, x_i); 1 \leq i \leq n\}\,, \text{ or } s = \{(y_i, x_i); 1 \leq i \leq n\}\,.$$

Here $x_i \in \mathbb{R}^p$ including an additional component equal to one for continuous outcomes but implicit nextafer for a lighter notation, while z_i and y_i are respectively continuous or discrete scalars. The number of observations is n hence $1 \leq i \leq n$ for the whole sample.

It is also directly found the resulting sample covariances by adding and substracting the subsamples means in the former squares after rewriting the sum on the sample s as two sums on each subsample s_ℓ, with finally:

$$\begin{aligned} \mathrm{cov}(x,z) &= \alpha_+\left\{c^+_{xz} + (m^+_x - \bar{m}_x)(m^+_z - \bar{m}_z)\right\}\\ &\quad +(1-\alpha_+)\left\{c^-_{xz} + (m^-_x - \bar{m}_x)(m^-_z - \bar{m}_z)\right\}\,. \end{aligned}$$

The same expression holds for $\mathrm{cov}(x,x)$ and $\mathrm{cov}(z,z)$ while for standardized columns it is clear that $m_x = \bar{m}_x = 0$ and $c_{xx} = 1$ which leads to even simpler expressions when c^-_{xz} is near or equal to zero. When also $\alpha_+ = \alpha_- = 0.5$ while one denotes $\Delta_x \approx m^+_x - m_x = m_x - m^-_x$ and $\Delta_z \approx m^+_z - m_z = m_z - m^-_z$, this possibly leads to get the between and within components as usually in variance decomposition, such that it is obtained:

$$\rho_{xz}(\Delta_x, \Delta_z) = \frac{c^-_{xz} + c^+_{xz} + 2\Delta_x\Delta_z}{\sqrt{c^+_{xx} + c^-_{xx} + 2\Delta^2_x}\sqrt{c^+_{zz} + c^-_{zz} + 2\Delta^2_z}}\,.$$

As there is no clear correspondence between ρ_{xz} and ρ^+_{xz}, ρ^-_{xz}, in the general case, it is checked further numerically how the correlations may be compared. Such case is met for some variables of the dataset of the example given in a subsection before and presented further in the experiment part, see Table 1 for the numerical values. The correlations and the different statistics for their computation from the whole sample and the two subsamples are found in this table. For this data, the proposed observation on correlations is true for the variable "bord", because the correlations are very smaller for the subsamples s_2 with lower outcomes, while for the other subsample s_1 with larger outcomes it is nearer to the one from the whole sample. Hence for this dataset, it is computed the correlation ρ_{xz} as a function of the quantities Δ_x and Δ_z.

Table 1: All statistics for each variable of D1. The values not included in the table are for the outcome for better rendering, $c_{zz} = 1201.45$, $c^+_{zz} = 533.90$ and $c^-_{zz} = 271.50$, $m_z = 26.94$, $m^+_z = 55.92$ and $m^-_z = -0.72$.

	hstrs	financ	fear	angry	happy	bord
m_x	0.00	-0.00	-0.00	0.00	0.00	-0.00
c_{xx}	1.00	1.00	1.00	1.00	1.00	1.00
c_{xz}	7.74	3.69	-7.51	-10.08	12.84	-16.28
ρ_{xz}	0.22	0.11	-0.22	-0.29	0.37	-0.47
α^+	0.49	0.49	0.49	0.49	0.49	0.49
$\bar{m}_x$	0.00	-0.00	-0.00	0.00	0.00	-0.00
$\bar{m}_z$	26.94	26.94	26.94	26.94	26.94	26.94
m^+_x	0.21	0.09	-0.18	-0.21	0.28	-0.40
c^+_{xx}	0.97	0.95	1.23	1.10	1.16	1.11
c^+_{xz}	4.47	2.10	-5.03	-6.82	8.52	-8.39
ρ^+_{xz}	0.20	0.09	-0.20	-0.28	0.34	-0.34
m^-_x	-0.20	-0.08	0.17	0.20	-0.27	0.38
c^-_{xx}	0.95	1.03	0.72	0.82	0.70	0.60
c^-_{xz}	-0.47	0.59	-0.16	-1.62	1.75	-2.42
ρ^-_{xz}	-0.03	0.04	-0.01	-0.11	0.13	-0.19

It is seen graphically in Figure 1 that the final correlation as a function of the two distances is dramatically changed in

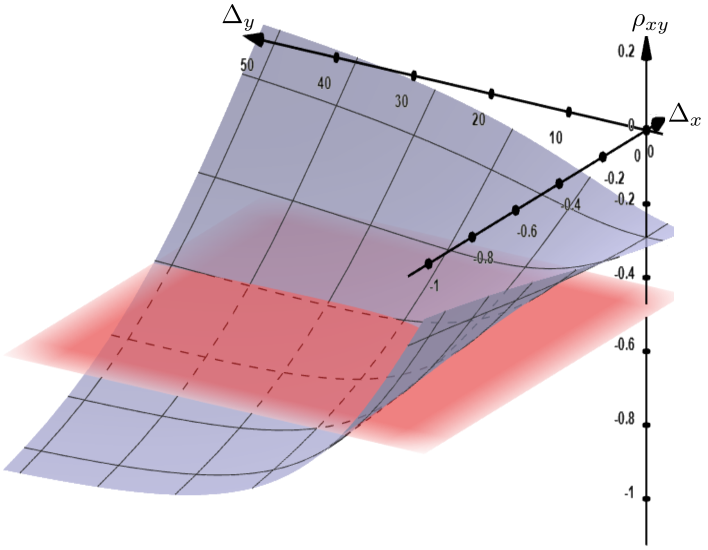

Figure 1: The correlations -here for the variable "bord"- are -for the whole sample- a function of the distances between the two clusters while the other data statistics are supposed constant. In this output from the sofware Geogebra, the three axis (X, Y, Z) are for the quantities Δ_x, Δ_y and ρ_{xy} respectively, by visualizing the bivariate function $(\Delta_x, \Delta_y) \rightarrow \rho_{xy}$. The horizontal plane is for the correlation equal to -0.47 in order to retrieve visually the value in Table 1 when $\Delta_x \approx -0.38 = \Delta_x^s$ and $\Delta_y \approx 28.32 = \Delta_y^s$ from the function defined just before.

comparison to the values from the two former correlations. This illustrates and validates our proposal of a segmented correlation herein.

The expression above is a function of two parameters, the distances Δ_x and Δ_z between the two clusters from before and after the break at the threshold M_z. Note that $\rho_{xz}(\infty, \infty) = sgn(\Delta_x^s \Delta_x^s)$ for distances towards infinite with chosen signs same than constant values Δ_x^s and Δ_z^s obtained from the available sample, when the function $u- > sgn(u)$ gives one if u is strictly positive, minus one if u is strictly negative and zero otherwise. For null distances, it is obtained a weighted sum of the two subsample correlations ρ_{xz}^- and ρ_{xz}^+. The expression of ρ_{xz} as a function of the distance Δ_x and Δ_y suggests (see also Figure 1) a decreasing surface towards $sgn(\Delta_x^s \Delta_x^s)$, such that finally, the combination of the two samples may lead by continuity to a possible increase in absolute value of the correlation, as observed for the dataset D1 in the current example. A consequence of the observed nonlinearity is that the usual linear regression may be not anymore relevant as explained next section.

3 Stratifying regression models

In this section, the models are presented for continuous and discrete outcomes, thus it is discussed the shape of the noise and the parameters involved in these parametric models for the distribution of each component for the regression.

3.1 Mixture models from subsamples

The dichotomy - presented at the two previous sections - can be seen via a decision rule with a binay classifier $\mathcal{C}$ such that the dependent variable z_i is a function of the independent variables aggregated in the vector $\mathbf{x}_i$ as follows:

$$z_i \quad \approx \quad \mathbf{x}_i^T \left[\delta_{\{\mathcal{C}(\mathbf{x}_i;\beta_1,\beta_2,M_z)=0\}} \beta_1 + \delta_{\{\mathcal{C}(\mathbf{x}_i;\beta_1,\beta_2,M_z)=1\}} \beta_2 \right] .$$

Here $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are two vectors of regression coefficients instead of an unique one $\boldsymbol{\beta}$ in order to model a change in the prediction for smaller and larger outcomes which is likely to happen in non physical and non biological real data because proportionality is mostly natural. For prediction, the model can be seen as a classification followed by local regressions hence a clustering or more generally a partioning of the data sample is available. The classifier is such that its success or failure corresponds to the dichotomy for the outcome, $\{z_i \leq M_z\}$ and $\{z_i > M_z\}$, but is required only at an eventual prediction step because during training the position of z_i w.r.t. M_z is already known. The variation for the coefficients is a function of the outcome but not in a continuous way from an independent variable as in [16], but in a discrete way instead. Thus, the variation is for a partioning as in an usual mixture of regression but according to a stratification instead of a data clustering. Hence the nearly equivalent and induced way via a cut of the variables x_i is an approximation not considered herein while it remains appealing for classification purposes as a perspective.

The model above is for predicting purposes, but for explaining purposes [17], the classification rule is already known and not involved for new data. Thus this model can be understood as a mixture model where the mixing parameters are a function of only the outcome. Let define the mixing parameters:

$$\pi_{i\ell}(z_i) \propto \exp(-(z_i - m_\ell)^2/\sigma_\ell^2) \,,$$

with the means m_ℓ and variances σ_ℓ^2 such as defined in univariate gaussian distributions. Note that for the mixture, another variable than z_i may be preferred, but this choice remains herein without loss of generality. This leads to a more general model where a smooth break is introduced within the loglikelihood as follows:

$$\ell^{\pi}_{M}(\boldsymbol{\theta}) \;=\; \sum_{i\in s} \log \left[\sum_{\ell=1}^{2} \pi_{i\ell}(\tilde{z}_i)\, g_{\boldsymbol{\theta}_\ell}(z_i, x_i^T\boldsymbol{\beta}_\ell)\right] .$$

Note that for more constraints, one may prefer to replace the posterior probabilities at the maximisation step (from an usual mixture) by $\pi_{i\ell}$. Here $\tilde{z}_i$ may be equal to z_i because the outcome is already known for any fitting and explaining needs. But one may prefer the value $x_i^T\boldsymbol{\beta}_\ell$ which is less precise but available for new data, such that the model becomes closely related to mixture-of-experts (MoE) [18, 19, 20] in this particular case. Here the usual mixing parameter π_ℓ in a mixture model is replaced by a parameterized softmax function $\pi_{i\ell}$ for modeling a smooth break at the median of the outcome. Otherwise, a difference [21] with the MoE is the parameterization of $\pi_{i\ell}$ directly with the observed outcome instead of the usual independent variables. This is a generalization because eventually for $\tilde{z}_i$ a linear estimation of z_i is preferred to just the directly available observation. The amount of mixture depends mostly on the value of the outcome, which extends the idea of hard break to the loglikelihood $\ell^{\pi}_{M}(\boldsymbol{\theta})$ with a smooth break just above. When the coefficients $\pi_{i\ell}$ are equal to one or zero, it is retrieved the same model than just before which is considered nextafter:

$$\begin{aligned}\ell^{01}_{M}(\boldsymbol{\theta}) \;&=\; \sum_{i} \delta_{\{z_i\le M_z\}} \log g_{\boldsymbol{\theta}_1}(z_i, x_i^T\boldsymbol{\beta}_1)\\ &+\; \sum_{i} \delta_{\{z_i> M_z\}} \log g_{\boldsymbol{\theta}_2}(z_i, x_i^T\boldsymbol{\beta}_2) .\end{aligned}$$

For deciding if two groups are acceptable for the vectors of regression coefficients one must choose a method, as studied herein for the considered distribution. Then the target variable takes ordered integer values or continuous values with either an explicit model or either a latent one for the linear regression next after.

3.2 Linear model with subsamples

If there is a change, the sample may be divided into two subsamples (or more) with same models but different parameters. For each component with parameter $\boldsymbol{\theta}$, one writes the linear model for one observation, a p-dimensional vector, with the probability density function:

$$g_{\boldsymbol{\theta}}(z_i, x_i^T\boldsymbol{\beta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\tfrac{-1}{2\sigma^2}(y_i - x_i^T\boldsymbol{\beta})^2\right) .$$

When two components are fitted instead of one, this model is duplicated into two identical ones, except that the parameters are denoted for the first component $\boldsymbol{\beta}_1$ and σ_1 while for the second component $\boldsymbol{\beta}_2$ and σ_2, with eventually $\sigma_1 = \sigma_2 = \sigma$ for a common noise.
This is more formally written as follows.

- For the whole available sample s, an usual linear regression leads to the following likelihood, for a model named "type I" for one component only,

  $$\mathcal{L}(\boldsymbol{\theta}) \;= \prod_{i\in s} g_{\boldsymbol{\theta}}(z_i, x_i^T\boldsymbol{\beta}) .$$

  Here the noise is gaussian, with expectation 0 and standard error σ, denoted $\epsilon_i \sim \mathcal{N}(0,\sigma)$ for the observed pair (x_i, z_i). While $\boldsymbol{\beta}$ denotes the vector of regressions coefficients, $X = [x_1|\cdots|x_n]^T$ the design matrix and $z = [z_1, \cdots, z_n]^T$ the vector of target variables.
- When there is a change in the sample s, then let suppose two samples s_1 of size n_1 and s_2 of size n_2 such that $s = s_1 \cup s_2$. The two related noises are normally distributed as follows $\epsilon_{i\ell} \sim \mathcal{N}(0, \sigma_\ell)$. This induces the new likelihood, for a model named "type II" for two (or more) components,

  $$\tilde{\mathcal{L}}(\boldsymbol{\theta}) \;= \prod_{i\in s_1} g_{\boldsymbol{\theta}_1}(z_i, x_i^T\boldsymbol{\beta}_1) \prod_{i\in s_2} g_{\boldsymbol{\theta}_2}(z_i, x_i^T\boldsymbol{\beta}_2) .$$

To check if the linear regression should be replaced by two regressions, the models proposed often refer to a statistical test [22] in order to estimate different regressions coefficients before and after the break while checking if these coefficients are different or equal. It looks mandatory when some break happens in the linear trend to change the model into a new one which is more relevant. Such as this turns into testing if either the two vectors of parameters are equal and if either they are not equal. Approaches for comparing and selecting between the two models are via hypothesis testing or via model choice, with the second way chosen herein. With a discretization of the outcome, such break may be double checked further from a dedicated relevant distribution for the obtained integer outcomes as explained next subsection.

3.3 Logit ordered model with subsamples

A widely studied model [23] for ordinal outcomes supposes that there exists latent response variables z_i which are unobserved and that they are linear functions of the independent variables, with eventually two parts herein. Instead of z_i, it is measured the ordinal variables y_i such that:

$$y_i = k \text{ if } \gamma_{k-1} < z_i < \gamma_{k+1} .$$

It is supposed $\gamma_0 = -\infty$ and $\gamma_K = +\infty$ for notations reasons. The quantities γ_k define the bounds of the intervals where z_i belongs for each level of the discrete variable y_i. The later ones also may be seen as label classes except that there is an order such as observed in a likert scale in psychometry for instance. More generally such values are found after recoding of a variable such as age with 1 for young, 2 for middle age and 3 for old, even if the ordering may be not always kept. Thus, in order to keep the ordering in the model, it is proceed as follows. The outcomes y_i with integer values are changed into binary versions $(y_{i1}, \cdots, y_{iK})$ for notation reasons.

- For one sample, the likelihood of the model is written as follows.

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i\in s}\prod_{k=1}^{k=K} Pr\left(y_i = k; \boldsymbol{\theta}\right)\\ &= \prod_{i\in s}\prod_{k=1}^{k=K} Pr\left(\{\gamma_{k-1} < z_i\} \cap \{z_i \leq \gamma_k\}; \boldsymbol{\theta}\right)\\ &= \prod_{i\in s}\prod_{k=1}^{k=K} g_{\boldsymbol{\theta}}(y_i, x_i^T\boldsymbol{\beta}; \gamma_{k-1}, \gamma_k)^{y_{ik}}\,.\end{aligned}$$

  where,

$$\begin{aligned}g_{\boldsymbol{\theta}}(y_i, x_i^T\boldsymbol{\beta}; \gamma_{k-1}, \gamma_k) &= \phi_{\boldsymbol{\theta}}\left(\gamma_k - x_i^T\boldsymbol{\beta}\right)\\ &\quad -\phi_{\boldsymbol{\theta}}\left(\gamma_{k-1} - x_i^T\boldsymbol{\beta}\right)\quad .\end{aligned}$$

  Here, for this model named "type I", the parameter vector is just $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \gamma_1, \cdots, \gamma_{K-1})^T$. Example of function for $\phi_{\boldsymbol{\theta}}(.)$ are the sigmoid one $\frac{e^u}{1+e^u}$ or the cumulative distribution function of the centered and reduced norm law $N(0, 1)$ for instance. Such related models are presented in [24, 25] for instance for an alternative to the multinomial regression model where the categories have no ordering.

- For two samples, there is a break at B with $1 < B < K$. This leads to denote two versions of the likelihood in stake, as a product with two parts which are multiplied for the whole sample:

$$\begin{cases}\mathcal{L}_1(\boldsymbol{\theta}_1) = \prod_{i\in s_1}\prod_{k=1}^{k=B} g_{\boldsymbol{\theta}_1}(y_i, x_i^T\boldsymbol{\beta}_1; \gamma_{k-1}, \gamma_k)^{y_{ik}}\\ \mathcal{L}_2(\boldsymbol{\theta}_2) = \prod_{i\in s_2}\prod_{k=B+1}^{k=K} g_{\boldsymbol{\theta}_2}(y_i, x_i^T\boldsymbol{\beta}_2; \gamma_{k-1}, \gamma_k)^{y_{ik}}\,.\end{cases}$$

  Thus the new overall likelihood is as follows:

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}_1(\boldsymbol{\theta}_1)\mathcal{L}_2(\boldsymbol{\theta}_2)\,.$$

  These models are named "type II" or more precisely "type II$_{\mathrm{ncon}}$" and "type II$_{\mathrm{con}}$" respectively for the first non contiguous and the second contiguous. It is denoted $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \gamma_1, \cdots, \gamma_B)^T$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2^T, \gamma_{B+1}, \cdots, \gamma_{K-1})^T$ for the first case hence with non contiguous parameters. While, $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \gamma_1^{(1)}, \cdots, \gamma_B^{(1)})^T$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2^T, \gamma_B^{(2)}, \cdots, \gamma_{K-1}^{(2)})^T$ for the second case with also $\gamma_B^{(1)} = \gamma_B^{(2)}$ hence with contiguous parameters. Note that the second version keeps an order at the change of the regression coefficients and less parameters because the quantities γ_B are shared between the two components.

To our knowledge these models are new approaches because there was not such proposed dichotomy, see for instance [26, 27, 28]. Nextafter, the models with subsamples are fitted with data in order to compare the results from one and two components, just after discussing the optimization for maximizing the loglikelihoods.

4 Summary and training algorithms

After that the models are defined with one or two components, one looks for a solution for the unknown parameters:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_I &= \mathrm{argmax}_{\boldsymbol{\theta}} \log\mathcal{L}(\boldsymbol{\theta})\\ \hat{\boldsymbol{\theta}}_{II} &= \mathrm{argmax}_{\boldsymbol{\theta}} \log\tilde{\mathcal{L}}(\boldsymbol{\theta})\,.\end{aligned}$$

The procedure for fitting the regression models with linear settings are explained just below with also the first and second order derivatives of the optimized criteria from the loglikelihoods.

4.1 Summary of the proposed family

These models are stratified normal or ordinal regressions in order to consider an underlying clustering along a variable, here the outcome. This is because the regression should not keep the same for lower and larger values of the outcome: the corresponding observations may be very different for social or economical reasons. For instance, when modeling the age, the living beings might not have the same physiology and social situation when younger versus older or working versus not working, such that the regression needs to change and a stratification may be mandatory in order to avoid a too much general model which is mistaken. The required nonlinearity is not anymore induced from the independent variables but directly from the dependent one. This leads to the proposed family of (weighted) loglikelihoods summarized in Table 2, with additional models when more constraints are introduced.

Table 2: Proposed family of stratified models along the outcome.

Model	Criterion	Constraints
Type I	$\log \mathcal{L}(\boldsymbol{\theta})$	
Type II_{ncon}	$\log \mathcal{L}_1(\boldsymbol{\theta}_1) + \log \mathcal{L}_2(\boldsymbol{\theta}_2)$	
Type $\text{II}^{\text{c}}_{\text{ncon}}$	$\log \mathcal{L}_1(\boldsymbol{\theta}_1) + \log \mathcal{L}_2(\boldsymbol{\theta}_2)$	$\hat{y}_1(\tilde{x}_0) = \hat{y}_2(\tilde{x}_0)$
Type II_{con}	$\log \mathcal{L}_1(\boldsymbol{\theta}_1) + \log \mathcal{L}_2(\boldsymbol{\theta}_2)$	$\gamma_B^{(1)} = \gamma_B^{(2)}$

An additional constraint here is possible for data where a value $\tilde{x}_0 \in s$ is available but may induce a bias when reducing the variance. Other additional constraints may also force that the prediction are ordored before and after the threshold M_z, for a few values of the vector of independent variables, say $\hat{y}_1(\tilde{x}_1) \leq \hat{y}_2(\tilde{x}_2)$ for some $\tilde{x}_1 \in s_1$ and $\tilde{x}_2 \in s_2$. The estimation of the parameters is explained next subsection for the models without additional constraints. A computer library will be available for fitting the different proposed models.

4.2 Optimization for the linear model

The models are as given previously in the previous subsection with one or two separated vectors of regression coefficients. For the case linear eventually latent, in each component having its own coefficients vector, it is supposed:

$$z_i \approx \boldsymbol{\beta}^T \mathbf{x}_i \,.$$

For the optimization of the parameters and increasing the likelihood, let denote the vector of first derivative of the log-likelihood $\boldsymbol{\nabla} \log \mathcal{L}(\boldsymbol{\theta})$ and the hessian matrix $\mathbf{H}_{\boldsymbol{\theta}}$ aggregating the second order derivatives. For instance, the Newton-Raphson algorithm repeats its iterations until convergence to a stable value when numbered (m) as below,

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} - \mathbf{H}^{-1}_{\boldsymbol{\theta}^{(m)}} \boldsymbol{\nabla} \log \mathcal{L}(\boldsymbol{\theta}^{(m)}) \,.$$

At the last iteration, one gets the maximum likelihood solution respectively denoted $\hat{\boldsymbol{\theta}}_I$ and $\hat{\boldsymbol{\theta}}_{II}$ for the one and two component(s) model(s). This algorithm and the computation of the hessian matrix is implemented eventually via a numerical solver from the computational library *scipy* and the module *optimize* or a dedicated libraries such as *statsmodels* or *scikit-learn* in Python, otherwise with a library in R for instance. Note that existing solvers for the ordinal regression may lead to different solutions because they may include or not constraints for the order of the interval bounds.

5 Experiments

In this section, the proposed models are implemented and compared with diverse numerical statistics for five real datasets in a descriptive purpose. Before the resulting interpretation, one decides if one or two components should be kept according to the computed criteria from each model, see next subsection. After the comparisons, an application with a medical dataset illustrates further our proposal with a discussion of the resulting regression coefficients.

5.1 Experimental settings

The output $\text{stat}^+_{\mathbf{y}\hat{\mathbf{y}}}$ and $\text{stat}^-_{\mathbf{y}\hat{\mathbf{y}}}$ are respectively the square root of the mean squared error (mse) for continuous outcomes or the mean absolute error (mae) for integer outcomes, both computed for each subsample and the whole sample, for comparisons. Thus, this is a sign plus for $y > M_y$ and a sign minus for $y \leq M_y$ where M_y is the median for each group (at the left or at the right of M_y) when computing $(y_i - \hat{y}_i)^2$ or $|y_i - \hat{y}_i|$ for continuous or integer outcomes. For a selection of the best model, an usual approach compares the values of the Bayesian information criterion (bic) or the Akaike information criterion (aic), when m is the number of parameters and n the sample size one writes that for instance for the one component model:

$$\begin{aligned} BIC(\mathbf{y}_s, \mathbf{x}_s) &= -2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}) + m \log(n) \\ AIC(\mathbf{y}_s, \mathbf{x}_s) &= -2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}) + 2m \,. \end{aligned}$$

Here $\mathbf{y}_s$ and $\mathbf{x}_s$ are replaced by $\mathbf{y}_{s_\ell}$ and $\mathbf{x}_{s_\ell}$ to refer to the same indicator but restricted to the subsample s_ℓ. These indicators behave nearly [29] as a cross-validation for enough large samples, hence they are informative for model selection such that the model with the lowest indicator is to be preferred between $\mathcal{L}(\hat{\boldsymbol{\theta}})$ and $\mathcal{L}_1(\hat{\boldsymbol{\theta}}_1)\mathcal{L}_2(\hat{\boldsymbol{\theta}}_2)$. Two different models for two subsamples may decrease the generalization thus one has to check if the regression coefficients or the corresponding models are statistically different. Nextafter, any values named BIC, AIC and LGL are respectively for bic and aic, and the loglikelihood at the optimum $\hat{\boldsymbol{\theta}}_I$ and $\hat{\boldsymbol{\theta}}_{II}$. This allows a ratio likelihood test for instance in the case of the linear regression but this is not discussed here further. The models with non constraints are computed with *statsmodels* and are compared next subsection just before an application.

Table 3: Output from the linear and ordinal regressions with one and two groups, named after table 2, in an unsupervised setting for the whole sample and the two subsamples.

		Linear regression model			Ordinal regression model		
		Type I	Type $\mathrm{II}_{\mathrm{ncon}}$		Type I	Type $\mathrm{II}_{\mathrm{ncon}}$	
		All z	$z > M_z$	$z \leq M_z$	All y	$y > M_y$	$y \leq M_y$
covid-19	LGL	-20886.83	-9461.39	-9356.11	-8979.19	-3205.33	-3211.28
	BIC	41832.32	18976.43	18766.19	18084.09	6487.30	6499.66
	AIC	41787.66	18936.79	18726.22	17988.38	6430.66	6442.55
	$\mathrm{stat}^{+}_{\mathbf{y\hat{y}}}$	30.53	20.55		3.88	1.61	
	$\mathrm{stat}^{-}_{\mathbf{y\hat{y}}}$	27.66		16.03	2.15		1.73
pre-diabet	LGL	-12030.81	-4908.17	-5528.07	-6796.29	-2334.08	-2476.72
	BIC	24101.76	9852.9	11092.9	13696.91	4726.67	5012.24
	AIC	24071.63	9826.33	11066.15	13618.57	4684.15	4969.43]
	$\mathrm{stat}^{+}_{\mathbf{y\hat{y}}}$	11.62	6.36		2.89	1.76	
	$\mathrm{stat}^{-}_{\mathbf{y\hat{y}}}$	13.02		8.41	3.17		2.13
housing	LGL	-22623.77	-11772.18	-1761.74	-36946.16	-14191.47	-14106.19
	BIC	45336.96	23627.54	3606.66	74061.22	28493.83	28323.28
	AIC	45265.54	23562.37	3541.48	73926.33	28406.94	28236.38
	$\mathrm{stat}^{+}_{\mathbf{y\hat{y}}}$	0.86	0.76		1.75	1.33	
	$\mathrm{stat}^{-}_{\mathbf{y\hat{y}}}$	0.56		0.29	1.38		1.25
pisa-2009	LGL	-30451.08	-14017.60	-14152.05	-11234.16	-4107.39	-4036.64
	BIC	61081.98	28200.45	28469.37	22716.63	8403.65	8262.15
	AIC	60944.16	28077.19	28346.11	22526.31	8262.79	8121.28
	$\mathrm{stat}^{+}_{\mathbf{y\hat{y}}}$	79.89	51.39		2.64	1.74	
	$\mathrm{stat}^{-}_{\mathbf{y\hat{y}}}$	83.00		53.99	2.60		1.68
life-expectancy	LGL	-8253.12	-3382.65	-4090.24	-4311.36	-1774.70	-1476.69
	BIC	16641.93	6889.08	8304.53	8822.27	3695.03	3099.30
	AIC	16540.23	6799.29	8214.49	8672.71	3589.40	2993.37
	$\mathrm{stat}^{+}_{\mathbf{y\hat{y}}}$	3.58	2.48		1.47	1.20	
	$\mathrm{stat}^{-}_{\mathbf{y\hat{y}}}$	4.47		3.87	1.07		1.04

5.2 Comparison of the linear models

The five studied datasets (covid-19[1], pre-diabet[2], life-expectancy[3], pisa-2009[4] and housing[5]) are described in the Table 4, after filtering eventual rows with missing outcomes. The outcome for the five datasets including the medical survey covid-19 is discretized into an ordinal variables with ten ordered categories. The regressions are computed for the continuous outcome and the discretized outcome. For each type of outcome, there is three cases: the usual full model for the whole sample plus the two separated models without shared parameters for the subsamples according to the position of the outcome z_i w.r.t. its median value M_z. The proportionality factors for the regression coefficients between the linear and discrete models are given in Table 5.

It is informative to notice that for the full model and for the model from the data where the outcome is larger than its median, the coefficients from the linear model and the ordinal model are proportional almost exactly for most of the datasets. For instance, the respective proportional factors are equal respectively to 28.80 and 20.33 for the survey dataset. The variance is high often for the subsample where the outcome is lower than the median for three datasets out of the five considered ones. This is observed for the survey

[1] "https://www.ncbi.nlm.nih.gov", "PMC7416923"

[2] "https://github.com", "MLDataR"

[3] "https://www.kaggle.com", "life-expectancy"

[4] "https://www.kaggle.com", "pisa-test-scores"

[5] "https://scikit-learn.org", "california_housing"

Table 4: For each dataset, the name, number of rows, number of variables kept and number of classes after discretizing the continuous outcome.

	Name	n	p	# classes
D1	covid-19	4361	6	10
D2	pre-diabet	3059	4	10
D3	life-expectancy	2928	16	10
D4	pisa-2009	5233	20	10
D5	housing	20640	8	10

Table 5: For each dataset, the average (and standard-deviation) factor of proportionality between the coefficients from the linear model for the continuous outcome and the ordered logit models for the discretized outcome.

Name	All z	$z > M_z$	$z \leq M_z$
D1	28.80 (2.28)	20.33 (3.57)	4.89 (34.07)
D2	10.57 (3.71)	8.34 (4.33)	7.84 (1.42)
D3	4.86 (6.48)	2.56 (1.13)	6.04 (7.91)
D4	76.96 (13.21)	62.69 (28.78)	20.81 (109.12)
D5	0.49 (0.29)	0.74 (0.18)	0.26 (0.03)

dataset with the third model where there is no proportionaly between the vectors of regression coefficients from the continuous and discretized outcomes, which may confirm that this model is less relevant. This makes sense that the models for continuous and discretized outcomes are related in some ways but a more formal proof would be required here, this is left to the reader but may be difficult to show. Similarly when the discrete outcomes are used in a multivariate regression instead of an ordinal one. A more robust model would be interesting to test here in order to check if same proportionality is met again when the fitting is more cautious.

The different indicators for the linear and ordinal regressions are presented in Table 3. From both indicators mse and mae, there is a dramatic improvement between modeling the full sample or the two subsamples with smaller indicators for the mixture model. For all the data samples, the value of M_y in the case of the ordered logit model was chosen equal to the value 5.0 because the discretization of the outcomes variables from the five datasets were according to the percentiles into ten classes hence with five classes for each component. Note that the ordinal regressions in the clusters have only five choices for the predictions while the full regression has ten choices, this difference may also induce a reduction of the mae with an error computed with less alternative choices. This illustrates the need for a method of model selection in order to decide which model is the more relevant among the two considered ones.

The indicator AIC is directly compared for the two likelihoods $\mathcal{L}(\boldsymbol{\theta})$ and $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ with for all the five datasets:

$$AIC(\mathbf{y}_s, \mathbf{x}_s) > AIC(\mathbf{y}_{s_1}, \mathbf{x}_{s_1}) + AIC(\mathbf{y}_{s_2}, \mathbf{x}_{s_2})\,.$$

For the five datasets, the aic is smaller and the mse is smaller for the proposed model. This induces that despite that there are more parameters, the model with two components is likely to be preferred according to the aic. The reduction for the mean squared error confirms and shows a dramactic decrease such that the new pairs of coefficients are more informative in order to explain the outcomes. According to these conducted numerical results, the proposed models with two components along the outcome lead to a better fit, while the model with one component, the usual regression, is a less relevant choice. This may be explained by several possible reasons: some nonlinear relations and spurious linear correlations between dependent and independent variables, an undetected heteroscedasticity of the noise with a varying variance instead of being constant, or the models for the lower and larger values of the outcome are different.

5.3 Application with medical data

The proposed approach is illustrated with a medical dataset "covid-19" from a psychological survey about the worldwide lockdown for covid which happened in the year 2020. This dataset was chosen because it is able to demonstrate the improvement when one considers two samples instead of one sample for the regression due to the correlations which may be spurious on the whole sample when looking at the bivariate empirical densities at the top right of the Figure 2. The correlations have not the same shape before and after the median, confirming different values in the correlations in Table 1. After checking, the issue does not look like coming from some bias after the truncation such that only the Accam razor may justify an usual regression as a first approach by choosing the linear model. The variables are for "y" the felt difference for the passage of time (during and before the lockdown) while for "x" are kept: boredom (occupied vs. bored), happiness (sad vs. happy), anger (peaceful vs. angry), fear of death, home stress and financial concern.

From a one component linear regression, the output is as follows in the Table 6 after imputation of missing "x" values with the median, removal of three rows for missing outcomes and centering plus reducing the design matrix. The linear regression is computed for the full sample, the subsample when

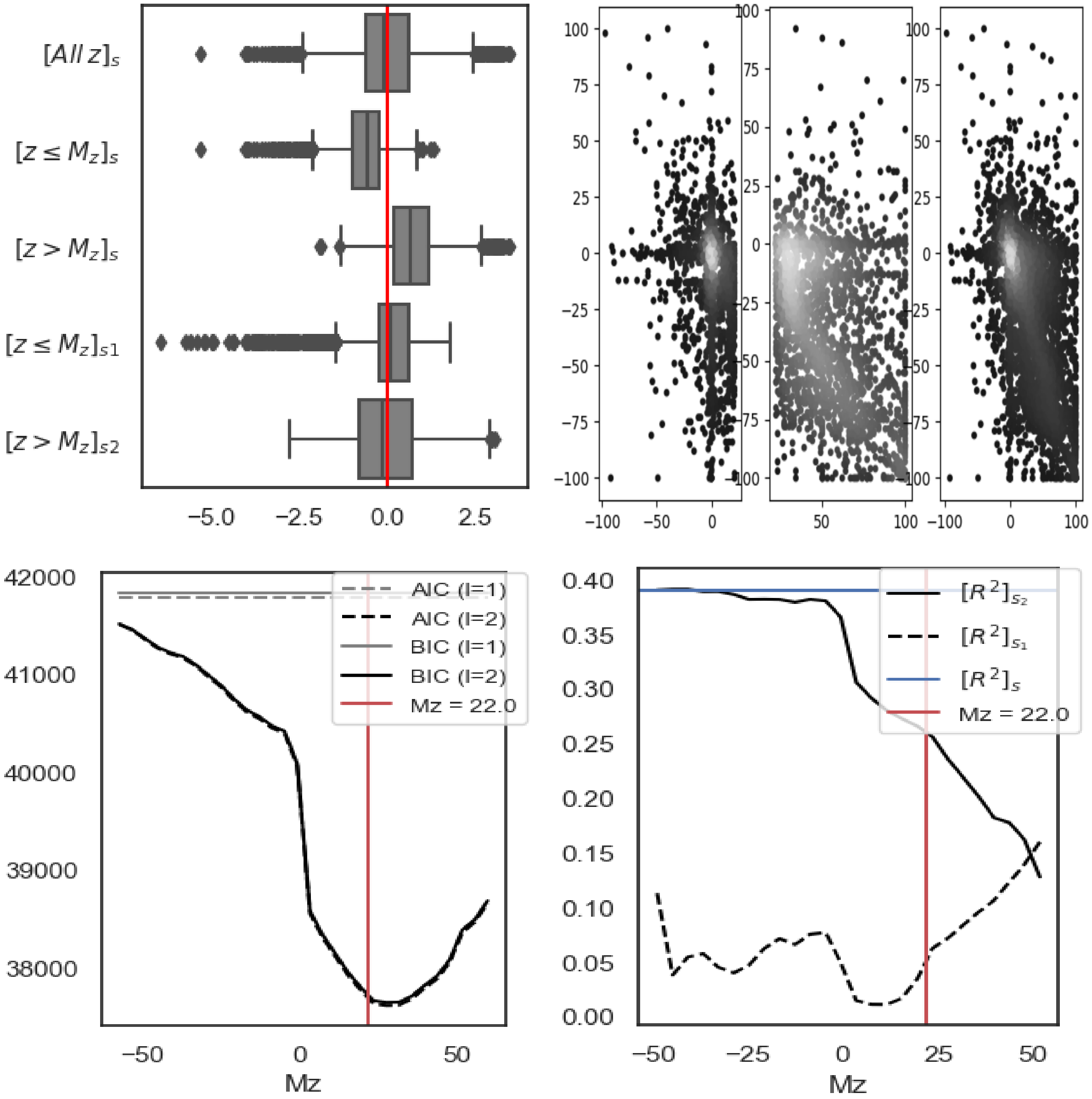

Figure 2: For left to right, first row to second row, for the sample D1. a) From the top, three boxplots for the residuals from the regression on the whole sample s and after from each subsample s_ℓ. Followed by two boxplots for the residuals from the regression fitted on each subsample separately. b) Scatter plot with bivariate densities between time and the main variable bored, for the full sample and the two sub-samples from the left and right to the median value of the outcome. c) Plot of the bic and aic for the regression with one and two components as a function of M_z. d) Plot of the R^2 with weigths from a robust regression as a function of M_z.

the outcome is less or equal to its median, and the subsample when the outcome is more than its median. The quantile regression is also given in Table 7 as obtained directly from the package *statsmodels* because the different fits were obtained separately without constraints for each subsample and for the whole sample.

Here $M_z = 22.0$ is the median of the outcome or "felt difference of time". The regression models for each subsample performs clearly better than the unique regression for the whole sample. This suggests also for instance that the parameters for "financ" and "fear" are not reliable because the standard-deviations have high values in comparison to

Table 6: Coefficients $\boldsymbol{\beta}$ and their standard-deviations from a linear regression for the sample D1 and its two subsamples.

	Type I		Type II_{ncon}			
	All z		$z > M_z$		$z \leq M_z$	
x_j	coefs	std	coefs	std	coefs	std
const	26.94	0.44	55.92	0.45	-0.72	0.34
hstrs	2.68	0.47	1.64	0.47	-1.05	0.35
financ	1.61	0.45	1.41	0.45	0.32	0.34
fear	-2.43	0.47	-1.68	0.48	0.30	0.35
angry	-2.51	0.52	-2.12	0.53	-1.14	0.39
happy	5.80	0.54	4.16	0.54	1.34	0.39
bored	-12.91	0.47	-5.84	0.47	-2.82	0.35

Table 7: Coefficients $\boldsymbol{\beta}$ and their standard-deviations from a quantile regression for the sample D1 and its two subsamples.

	Type I		Type II_{ncon}			
	All z		$z > M_z$		$z \leq M_z$	
x_j	coefs	std	coefs	std	coefs	std
const	24.63	0.02	53.87	0.02	0.39	0.02
hstrs	2.72	0.02	1.81	0.02	-0.03	0.02
financ	1.22	0.02	2.59	0.02	0.00	0.02
fear	-2.28	0.02	-2.16	0.02	0.00	0.02
angry	-2.53	0.02	-2.69	0.03	-0.01	0.02
happy	5.79	0.02	5.21	0.03	0.29	0.02
bord	-17.75	0.02	-7.42	0.02	-0.61	0.02

the estimated coefficients, while there is a change of sign for "hstrs", such that less variables may be kept here (as suggested also by a quantile regression) for further analysis. This result combinated with the reduced mean squared error and the information criterion confirms that the one component linear regression may be not enough relevant for explaining the lower values of the outcome. Graphically in Figure 2, it is checked the residuals divided by their standard deviations, for the two groups of observations when one regression or two are fitted, such that considering two components allows a better centering but also adds outliers in the noise for one of the two components such that the gaussian distribution may be replaced for a better fit.

The output in Figure 2 suggests visually that the models are different for each part of the sample before and after the median. According to the scatter plots, the relations are not completely linear such as the Pearson correlations are not able to explain fully the links between the independent variables and the outcome. Scatter plots with bivariate densities allow a better overview of the linear or nonlinear correlation. The curves in Figure 2, at the left and below, shows the value of the bic and aic when the value of the break changes, such that the median is not far of the optimal threshold value for this dataset. The curves in Figure 2, at the left and below, shows the determination coefficient of Pearson or R^2 which also suggests a change. Note that an usual mixture of linear regressions lead to only 40926.73 for aic and 41035.2 for bic, which confirms the interest of our proposal for data analysis. For this dataset a mixture of freed quantile regressions with two components in Table 7 remains competitive with an error slightly smaller but with an alternative and more tricky interpretation. As explained above, it is expected the coefficients to be almost equal for the two first models in the tables, but different for the last model. The one component model remains able to provide informative coefficients for explaining purposes but biased to an unknown amount. This result is confirmed by the curve for a robust coefficient of determination R^2 where only the two components have very different values for this indicator. The regression model for s_1, the lower values of the outcome, remains weak such that the final result may be that only the component for s_2 for this dataset is relevant for the analysis for an explaining purpose, as in an usual two-parts model [30]. One may keep only its corresponding component (or eventually the one for $M_z > 0.0$), instead of both components otherwise both components are required for a prediction purpose. Reproducing the study with new data would allow to check further these numerical results, but lockdowns are rare events. More research may be wanted in future for further validation of the proposed model for this dataset: biases study via monte-carlo simulations, distributional study of the residuals via statistical tests, generalized error via cross-validation or additional comparison with existing (mixture) models for instance. Note also that a truncated or a censored distribution plus an additional transformations may fit even better the data.

6 Conclusion and perspectives

In this paper, it is discussed the multiple regression model and the ordered logit model in order to separate into two components such models along the outcome. The dichotomy is discussed and tested with real data with continuous and discretized outcomes. The model may be not the same for lower and upper values of the outcome (and more generally ranges of values), while the usual linear regression model makes this hypothesis. A model for intervals of outcome may be able to simultanously allow a better fit for the available data sample and to detect issues with the linear model. To our knowledge, the contribution is novel for dealing with regression:

- The correlation between two variables may increase when a sample aggregates two subsamples: from two weak to one moderate.
- A smooth mixing of components along the outcome leads to a mixture of regressions which is able to improve the fit.
- A discretization of the outcome allows to fit an ordinal regression with related proportional regression coefficients.
- The curves with a varying window along the outcome confirms graphically the varying quality of the fit along the outcome.

Such parameterizations for the ordinal regression and even the gaussian one were not studied before because the mixture of regressions finds a clustering of the independent and dependent variable together thus not along the outcome alone. A stratification of the outcome for regression is able to improve dramatically the results for explaining (and eventually predicting with the knowledge of the ranges or an approximation via the induced cut of the explaining variables) an outcome from some datasets. In short, this research suggests that in some cases, the interpretation from the regression may be biased with a part of the outcome more explained than another part. Possible alternatives to this lack of non-linearity may be a two-parts model, an usual mixture model of regressions, or as proposed herein to simply cluster with regressions along the outcome.

Future perspectives for the reader is to look for a formal proof of the increased correlation from underlying weak ones, the number and the choice of the strata for discretizing the outcome, for a more robust or smoother setting, for the optimal value(s) of M_z and $\tilde{x}_0$, and the consequence on the variance of the coefficients. As a complement to the graphical outputs, one may also show the curves for the coefficients of regressions from each component in order to check their trend when the threshold M_z varies. Other curves may be the distance between the regressions coefficients from the full model and each submodel or any other indicator such as the varying coefficient of determinations or the correlations for instance with local values within ranges from the outcome. This result also suggests the need for local indicators within an interval from a moving window of the outcome in order to check further the validity of a one component linear regression and to understand for which values or intervals of the outcome, the model remains relevant or not relevant. Other concerns may be the behavior of the involved criteria when the noises are not full gaussian or independent, with alternative distributions for the residuals including varying variance along the outcome or by adding priors for a bayesian estimation. Comparing with the existing piecewise models, fitting hybrid models such that mixtures with and without regression, proceeding to their validation with statistical tests in order to validate further the choice of several components, and solving for their eventual bias, are finally of main interest in order to choose and fit an optimal model.

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning: data mining, inference and prediction*, Springer, 2 edition, 2009.

[2] Gary King and Margaret Roberts, “How robust standard errors expose methodological problems they do not fix, and what to do about it,” *Political Analysis*, vol. 23, pp. 159–179, 04 2014.

[3] BS Everitt, “An introduction to finite mixture distributions,” *Statistical Methods in Medical Research*, vol. 5, no. 2, pp. 107–127, 1996.

[4] Geoffrey J. McLachlan, Sharon X. Lee, and Suren I. Rathnayake, “Finite mixture models,” *Annual Review of Statistics and Its Application*, vol. 6, no. 1, pp. 355–378, 2019.

[5] Jian Liu, Shiying Wu, and James V. Zidek, “On segmented multivariate regression,” *Statistica Sinica*, vol. 7, no. 2, pp. 497–525, 1997.

[6] Vito M. R. Muggeo, “Estimating regression models with unknown break-points,” *Statistics in Medicine*, vol. 22, no. 19, pp. 3055–3071, 2003.

[7] Ioannis Gkioulekas and Lazaros G. Papageorgiou, “Piecewise regression analysis through information criteria using mathematical programming,” *Expert Systems with Applications*, vol. 121, pp. 362–372, 2019.

[8] Odile Sauzet, Oliver Razum, Teresia Widera, and Patrick Brzoska, “Two-part models and quantile regression for the analysis of survey data with a spike. the example of satisfaction with health care,” *Frontiers in Public Health*, vol. 7, 2019.

[9] G. Udny Yule, “Notes on the theory of association of attributes in statistics,” *Biometrika*, vol. 2, no. 2, pp. 121–134, 1903.

[10] Joseph Berkson, “Limitations of the application of fourfold table analysis to hospital data,” *Biometrics Bulletin*, vol. 2, no. 3, pp. 47–53, 1946.

[11] E. H. Simpson, “The Interpretation of Interaction in Contingency Tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 13, no. 2, pp. 238–241, 1951.

[12] Aiyou Chen, Thomas Bengtsson, and Tin Kam Ho, "A regression paradox for linear models: Sufficient conditions and relation to simpson's paradox," *The American Statistician*, vol. 63, no. 3, pp. 218–225, 2009.

[13] Rogier Kievit, Willem Frankenhuis, Lourens Waldorp, and Denny Borsboom, "Simpson s paradox in psychological science: A practical guide," *Frontiers in psychology*, vol. 4, pp. 513, 08 2013.

[14] Aris Spanos, "Yule-simpson's paradox: the probabilistic versus the empirical conundrum," *Statistical Methods & Applications*, vol. 30, 07 2020.

[15] Bruce Ratner, "The correlation coefficient: Its values range between +1/-1, or do they?," *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 17, 06 2009.

[16] Trevor Hastie and Robert Tibshirani, "Varying-coefficient models," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 55, no. 4, pp. 757–779, 1993.

[17] Galit Shmueli, "To Explain or to Predict?," *Statistical Science*, vol. 25, no. 3, pp. 289 – 310, 2010.

[18] Robert Jacobs, Michael Jordan, Steven Nowlan, and Geoffrey Hinton, "Adaptive mixture of local expert," *Neural Computation*, vol. 3, pp. 78–88, 02 1991.

[19] Seniha Esen Yuksel, Joseph N. Wilson, and Paul D. Gader, "Twenty years of mixture of experts," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1177–1193, 2012.

[20] Isobel Claire Gormley, Thomas Brendan Murphy, and Adrian E. Raftery, "Model-based clustering," *Annual Review of Statistics and Its Application*, vol. 10, no. 1, pp. 573–595, 2023.

[21] Freek Stulp and Olivier Sigaud, "Many regression algorithms, one unified model: A review," *Neural Networks*, vol. 69, pp. 60–79, 2015.

[22] Gregory C. Chow, "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, vol. 28, no. 3, pp. 591–605, 1960.

[23] Peter McCullagh, "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 109–127, 1980.

[24] C V Ananth and D G Kleinbaum, "Regression models for ordinal responses: a review of methods and applications.," *International Journal of Epidemiology*, vol. 26, no. 6, pp. 1323–1333, 12 1997.

[25] Andrew S. Fullerton, "A conceptual framework for ordered logistic regression models," *Sociological Methods & Research*, vol. 38, no. 2, pp. 306–347, 2009.

[26] Gerhard Tutz, "Ordinal regression: A review and a taxonomy of models," *WIREs Computational Statistics*, vol. 14, no. 2, pp. e1545, 2022.

[27] Dongying Zhan and Derek Young, "Finite mixtures of mean-parameterized conway-maxwell-poisson regressions," *Journal of Statistical Theory and Practice*, vol. 18, 01 2024.

[28] Ryan Ho Leung Ip and Ka Yui Karl Wu, "A mixture distribution for modelling bivariate ordinal data," *Statistical Papers*, pp. 1–36, 05 2024.

[29] M. Stone, "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 44–47, 1977.

[30] Siu Fai Leung and Shihti Yu, "On the choice between sample selection and two-part models," *Journal of Econometrics*, vol. 72, no. 1, pp. 197–229, 1996.