WEIGHTED CROSS VALIDATION IN THE SELECTION OF ROBUST REGRESSION MODEL WITH CHANGE-POINT FOR TELEVISION RATING FORECAST

VALIDAÇÃO CRUZADA PONDERADA EM SELEÇÃO DE MODELO DE REGRESSÃO ROBUSTO COM PONTO DE MUDANÇA PARA PREVISÃO DE AUDIÊNCIAS DE TELEVISÃO

Carlos Alberto Huaira-Contreras1
Carlos Cristiano Hasenclever Borges2
Camila Borelli Zeller3
Amanda Romanelli4

Abstract: The paper proposes a weighted cross-validation (WCV) algorithm to select a linear regression model with change-point under a scale mixtures of normal (SMN) distribution that yields the best prediction results. SMN distributions are used to construct robust regression models to the influence of outliers on the parameter estimation process. Thus, we relaxed the usual assumption of normality of the regression models and considered that the random errors follow a SMN distribution, specifically the Student-t distribution. In addition, we consider the fact that the parameters of the regression model can change from a specific and unknown point, called change-point. In this context, the estimations of the model parameters, which include the change-point, are obtained via the EM-type algorithm (Expectation-Maximization). The WCV method is used in the selection of the model that presents greater robustness and that offers a smaller prediction error, considering that the weighting values come from step E of the EM-type algorithm. Finally, numerical examples considering simulated and real data (data from television audiences) are presented to illustrate the proposed methodology.

Keywords: Cross-validation. Regression model. EM-type algorithm. Student-t distribution. Change-point.

1Mestre, Universidade Federal de Juiz de Fora, huairac@hotmail.com.
2Doutor, Universidade Federal de Juiz de Fora, cchborges@ice.ufjf.br.
3Doutora, Universidade Federal de Juiz de Fora, camila.zeller@ufjf.edu.br.
4Bachaler, Universidade Federal de Juiz de Fora, amandaromanelli@ice.ufjf.br.
Resumo: O trabalho propõe um algoritmo de validação cruzada (VC) ponderada para selecionar um modelo de regressão linear com ponto de mudança contínuo sob uma distribuição Mistura de Escala Normal (MEN) que produza os melhores resultados de predição. As distribuições MEN são utilizadas para construir modelos de regressão robustos à influência de observações aberrantes (outliers) sobre o processo de estimação dos parâmetros. Assim, relaxamos a suposição de normalidade, usual no contexto de regressão, considerando que os erros aleatórios seguem uma distribuição MEN, especificamente a distribuição t-Student. Além disso, consideramos o fato de que os parâmetros do modelo de regressão podem se alterar a partir de um ponto específico e desconhecido, que é denominado ponto de mudança. Neste contexto, as estimações dos parâmetros do modelo, que inclui o ponto de mudança, são obtidas via algoritmo tipo-EM (Esperança-Maximização). O método de validação cruzada ponderado é utilizado na seleção do modelo que apresente maior robustez e que ofereça um menor erro de predição, considerando que os valores de ponderação são provenientes da etapa E do algoritmo tipo-EM. Finalmente, exemplos numéricos considerando dados simulados e reais (dados de audiências televisivas) são apresentados para ilustrar a metodologia proposta.

Palavras-chave: Validação cruzada. Modelo de regressão. Algoritmo tipo-EM. Distribuição t-Student. Ponto de mudança.
1 INTRODUCTION

Linear regression models are widely used to describe the average relationship between a response variable and one or more explanatory variables. Their applications are found in several scientific areas: economics, agriculture, biology, medical sciences, and others.

In regression procedures it is usually assumed that a model is valid for the entire data set. However, the data behavior may change from a specific value (time, indexed value or a specific value of the predictor variable, for example), generating a point of change in the model. In the context of linear models, problems of change-point with discontinuity, under normality of the data were widely treated (CHEN; GUPTA, 2001; CHEN; GUPTA, 2011). In the case of a continuous change-point, the works of Muggeo (2003) and Hofrichter (2007) can be cited as examples.

The assumption of normality error in a regression model is usually adopted in the literature. However, this assumption becomes unrealistic when the data follow a distribution with heavier tails, and even more, the coefficient estimates are sensitive to extreme observations. Thus, the development of flexible methodologies using non normal probability distributions, such as the SMN distribution class (ANDREWS; MALLOWS, 1974), could be proposed as an alternative. The estimation of parameters in this type of model is performed iteratively using EM-type algorithms, such as the EMC (Expectation-Conditional Maximization) algorithm, detailed in the works by Dempster, Laird e Rubin (1977) and Meng e Rubin (1993). Applications that use some distributions that make up SMN class in regression models can be found in Lange, Little e Taylor (1989), Lange e Sinsheimer (1993), Yamaguchi (1990), Rosa, Padovani e Gianola (2003) and Osorio (2006). Results on regression models with change-point discontinuous using robust distributions can be found in Huaira-Contreras (2014) and Young (2014).

Aiming to determine an prediction model based on a data set, the strategy of dividing this set in two parts is usually adopted, with the first group to adjust the model and the other to validate the potential of generalization of the model. Cross-validation (CV) to assess the predictive ability in a regression model is presented by Picard e Cook (1984). Shao (1993) proposes a method that uses CV to select a model that has a better prediction. Ronchetti, Field e Blanchard (1997) and Markatou, Afendras e Agostinelli (2018) extend this idea and propose WCV methods for selecting robust models for extreme observations. Following this approach, this work proposes a WCV method for selecting a regression model with a change-point continuous that considers distribution with heavier tails (SMN), as an alternative to normal distribution, in order to reduce the influence of outliers and allowing high level of predictivity. The presence of the change-point influences the planning matrix, and leads to a modification of
the algorithm proposed by Markatou, Afendras e Agostinelli (2018).

The paper is developed as follows: The regression model with a continuous change-point is specified in Section 2, some important concepts about the SMN distributions and the description of the EM-type algorithm used to obtain the maximum likelihood estimators are also presented; in Section 3, the WCV method and the respective algorithm, in the context of the proposed model, are described. Section 4 presents some simulation studies to evaluate the proposed methodology. An application on television audience data in some regions of Brazil is presented in Section 5. Finally, some observations and conclusions are presented in Section 6.

2 SPECIFICATION OF THE REGRESSION MODEL

According to Muggeo (2003) and Young (2014), the linear regression model with a continuous and unknown change-point, considering a sequence of observations \((x_i, Y_i), i = 1, \ldots, n\), can be specified as:

\[
Y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \gamma)_+ + \varepsilon_i,
\]

where \((x_i - \gamma)_+ = (x_i - \gamma)I_{\{x_i \geq \gamma\}}\), such that \(I_{\{\cdot\}}\) is the indicator function and \(\gamma\) is the change-point. Note that if \(x_i > \gamma\), then \((x_i - \gamma)_+ = (x_i - \gamma)\) and in turn, if \(x_i \leq \gamma\), then \((x_i - \gamma)_+ = 0\). In this way, the model defines a line to the left of the change-point given by \(Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\), while to the right of the change-point, we have \(Y_i = (\beta_0 - \beta_2 \gamma) + (\beta_1 + \beta_2) x_i + \varepsilon_i\). That is, \(\beta_1\) is the slope of the line to the left of the change-point, and \((\beta_1 + \beta_2)\) is the slope of the line to the right. Thus, \(\beta_2\) is the parameter that represents the “slope difference” of these lines. In addition, it is assumed that the explanatory variable is classified into ascending order, \(x_i \leq x_{i+1}, i = 1, \ldots, n - 1\). Now, the location of the change is no longer restricted to an observed \(x_i\). Instead, it can be any value within range \([a; b]\), where are the following points \(a = \min x_i = x_{(1)}\) and \(b = \max x_i = x_{(n)}\). In this work, we replace in equation (1) the usual assumption of normal distribution of random errors \(\varepsilon_i, i = 1, \ldots, n\), as

\[
\varepsilon_i \overset{iid}{\sim} t(0, \sigma^2, \nu)
\]

and so, we have

\[
Y_i \overset{iid}{\sim} t(\beta_0 + \beta_1 x_i + \beta_2 (x_i - \gamma)_+, \sigma^2, \nu), \quad i = 1, \ldots, n.
\]

Thus, the model proposed by the equations (1)-(2) considers that the probability distribution of errors belongs to the SMN distribution class. Specifically, in this work, we
consider the Student-t distribution, a particular case of this class of distributions. Furthermore, assuming \( \nu \) fixed in equation (3) (LUCAS, 1997), the parameter vector to be estimated is \( \theta = (\gamma, \beta_0, \beta_1, \beta_2, \sigma^2)^T \) and the maximum likelihood estimators for \( \theta \) are obtained via EM-type algorithm.

2.1 Scale Mixtures of Normal Distributions

The SMN symmetric distribution class (ANDREWS; MALLOWS, 1974) is constructed from the scale mix of a normal variable and a positive random variable, this is,

\[
Y = \mu + \kappa^{1/2}(U)Z, \tag{4}
\]

where \( \mu \) is a location parameter (\( \mu \in \mathbb{R} \)), independent of \( Z \), whose distribution is indexed by the \( \nu \) parameter that controls the tails of the SMN distributions. \( Y \) conditioned to \( U = u \) has normal distribution, this is, \( Y | U = u \sim N(\mu, \kappa(u)\sigma^2) \). Therefore, the pdf of \( Y \) is given by

\[
f(y) = \int_0^\infty \phi(y; \mu, \kappa(u)\sigma^2)dH(u; \nu), \tag{5}
\]

where \( \phi(.; \mu, \kappa(u)\sigma^2) \) denotes the pdf of a normal distribution with mean \( \mu \) and variance \( \kappa(u)\sigma^2 \). The normal distribution is a particular case of this class of distribution, obtained when \( H \) is a degenerate cdf and \( \kappa(u) = 1 \). Other families make up this class of distributions, for example, the Student-t.

The Student-t distribution with \( \nu \) degrees of freedom, denoted by \( t(\mu, \sigma^2, \nu) \), can be obtained from the definitions (4) or (5) considering \( \kappa(u) = u^{-1} \), where \( U \) has a distribution \( \text{Gamma}(\nu/2, \nu/2) \), with \( u > 0 \) and \( \nu > 0 \). In this case, the pdf of \( Y \) is given by

\[
f(y) = \frac{\Gamma(1+\nu/2)}{\Gamma(\nu/2)\pi^{1/2}\nu^{-1/2}\sigma^{-1}} \left(1 + \frac{d}{\nu}\right)^{-((1+\nu)/2)}, \quad y \in \mathbb{R}, \tag{6}
\]

where \( d = \frac{(y - \mu)^2}{\sigma^2} \) is the Mahalanobis distance, useful in identifying outliers. This distribution has tails heavier than the normal distribution. In addition, the Student-t distribution tends to a normal distribution when \( \nu \to \infty \).
2.2 The EM Algorithm

Note that the model specified in equations (1)-(2) can be described hierarchically as

\[ Y_i \mid U_i = u_i \sim N \left( \beta_0 + \beta_1 x_i + \beta_2 (x_i - \gamma)_+, \frac{\sigma^2}{u_i} \right), \quad (7) \]

\[ U_i \sim \text{Gamma} \left( \frac{\nu}{2}, \frac{\nu}{2} \right), \quad i = 1, ..., n. \quad (8) \]

In this estimation process, consider \( y = (y_1, ..., y_n)^T \) the vector of observed responses to \( n \) sample units and \( u = (u_1, ..., u_n)^T \). So, under the hierarchical representation (7)-(8), it follows that the complete log-likelihood associated with \( y_c = (y^T, u^T) \) is given by

\[ \ell_c(\theta) \propto -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} u_i (y_i - \beta_0 + \beta_1 x_i + \beta_2 (x_i - \gamma)_+)^2. \quad (9) \]

Let \( \hat{\theta}^{(t)} \equiv (\hat{\gamma}^{(t)}, \hat{\beta}_0^{(t)}, \hat{\beta}_1^{(t)}, \hat{\beta}_2^{(t)}, \hat{\sigma}^2(t)) \) the estimate of \( \theta \) in the \( t \)-th iteration, the conditional expectation of the complete log-likelihood function, \( Q(\theta | \hat{\theta}^{(t)}) = E[\ell_c(\theta | y_c) | \hat{\theta}^{(t)}], y] \), is given by

\[ Q(\theta | \hat{\theta}^{(t)}) = -\frac{n}{2} \log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2(t)} (y - X_{\gamma(t)} \hat{\beta}^{(t)})^T D(\hat{\beta}^{(t)}) (y - X_{\gamma(t)} \hat{\beta}^{(t)}), \quad (10) \]

where \( \hat{\beta}^{(t)} = (\hat{\beta}_0^{(t)}, \hat{\beta}_1^{(t)}, \hat{\beta}_2^{(t)})^T \), \( X_{\gamma(t)} = [1_n^T x_\gamma^T] \), such that \( 1_n \) is a all-one vector of dimension \( 1 \times n \), \( x = (x_1, ..., x_n) \), \( x_{\gamma(t)} = ((x_1 - \hat{\gamma}^{(t)})_+, ..., (x_n - \hat{\gamma}^{(t)})_+) \) and \( D(\hat{\beta}^{(t)}) = \text{diag}(\hat{\sigma}_1^{(t)}, ..., \hat{\sigma}_n^{(t)}) \).

Assuming \( \theta = (\theta_1, \theta_2^T)^T \), where \( \theta_1 = \gamma \) and \( \theta_2 = (\beta^T, \sigma^2)^T \), the \((t+1)\)-th iteration of the EM-type algorithm is described as following.

2.2.1 Description of the EM-type Algorithm

The EM-type algorithm is composed by well defined steps: Expectation (E) and Maximization (M), in this propose the M step is divided in two conditional steps CM1 and CM2.

**STEP E:** From the estimates \( \hat{\theta}^{(t)} \) and a fixed value \( \nu \), the weights \( \hat{u}_i^{(t)} \) are obtained through the following conditional expected:

\[ E[U_i | y_i, \hat{\theta}^{(t)}] = \hat{u}_i^{(t)} = \frac{\nu + 1}{\nu + \hat{d}_i^{(t)}}, \quad \text{with} \quad \hat{d}_i^{(t)} = \frac{(y_i - X_{\gamma(t)} \hat{\beta}^{(t)})^2}{\hat{\sigma}^2(t)}, \quad i = 1, ..., n. \quad (11) \]
STEP CM1: In this step, the change-point is estimated, where we assume that the change-point must occur within the domain of $x$, and we calculate

$$\hat{\theta}_1^{(t+1)} = \text{argmax}_{\theta_1} Q(\theta | \hat{\theta}_1^{(t)})$$, with $\theta_2$ fixed as $\hat{\theta}_2^{(t)}$.

STEP CM2: Calculate $\hat{\theta}_2^{(t+1)} = \text{argmax}_{\theta_2} Q(\theta | \hat{\theta}_1^{(t)})$, with $\theta_1$ fixed as $\hat{\theta}_1^{(t)}$, to get

$$\hat{\beta}^{(t+1)} = (X_{\hat{\theta}_1^{(t)}}^T D(\hat{u}^{(t)}) X_{\hat{\theta}_1^{(t)}})^{-1} X_{\hat{\theta}_1^{(t)}}^T D(\hat{u}^{(t)}) y,$$

and

$$\hat{\sigma}^2(t+1) = \frac{1}{n} (y - X_{\hat{\theta}_1^{(t)}} \hat{\beta}^{(t)})^T D(\hat{u}^{(t)}) (y - X_{\hat{\theta}_1^{(t)}} \hat{\beta}^{(t)}).$$

Note that step E of the algorithm shows that the higher the value of $\hat{d}_i$, the lower the value of $\hat{u}_i$. Thus, the estimation procedure tends to give less weight to the atypical observations in the direction of the Mahalanobis distance. Thus, when using distributions with tails heavier than the normal distribution, the EM algorithm accommodates atypical observations, giving them less weight in the estimation process. When $u_i = 1$, $i = 1, ..., n$, the results described above coincide with the maximum likelihood estimates of the linear regression model with change-point under normal errors.

Steps E and M must be repeated alternately until convergence is achieved. The criterion of convergence used is $||\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}|| < \epsilon$, where $||*||$ indicates the norm of the vector and $\epsilon > 0$.

### 3 WEIGHTED CROSS-VALIDATION METHOD

In this work, the suggested WCV algorithm is based on the proposal by Markatou, Afendras e Agostinelli (2018) that presents a WCV algorithm to select the best set of covariates in a multiple regression model and considers a nonparametric method to determine the weights of the observations in the data set. We consider a regression model with change-point and a parametric method to define the weighting.

Two main characteristics important for the proposed algorithm are emphasized. First, good generalizability, that is, the performance of the algorithm on the training set accurately reflects the performance on the validation set an important property in models used for forecasting. Second, stability performance in the sense that the removal of any instance in the training set usually results in a small change in the parameter estimation. The theoretical framework and discussions on the concepts of generalization and stability are presented in Bousquet e Elisseeff (2002).

It is expected that in the WCV method proposed in this work, outliers have little influence,
because these observations will be weighted by the values $\hat{u}_i, i = 1, \ldots, n$, coming from step E of the EM-type algorithm, expressed in equation (11). In other words, in the training sample, are obtained parameter estimates robust to extreme observations and later, these estimates are used to calculate the weights $\hat{u}_i, i = 1, \ldots, n$ in the validation sample. Thus, the criterion for choosing the best model via the mean of the prediction error also considers a limited influence from the outliers.

3.1 Description of the Algorithm

The proposed CV method is described in the following steps:

1. Estimate $\theta = \theta(\nu) = (\gamma, \beta_0, \beta_1, \beta_2, \sigma^2)^T$ considering the total data set via the proposed EM-type algorithm for all $\nu \in A$, such that $A = \{1 \leq \nu \leq \nu_{max}, \nu \in \mathbb{N}\}$.

2. Select $\theta(\nu)$ estimated such that the log-likelihood function is maximum, among the estimates made in the previous step.

3. Build the weighting matrix $\hat{W} = D(\hat{u}) = \text{Diag}(\hat{u}_1, \ldots, \hat{u}_n)$ on the total data set, considering $\theta(\nu)$ selected in the previous step.

4. Divide randomly the total data set into two parts for training ($c_t$) and validation ($c_v$) of sizes $n_t$ and $n_v$, such that $n_t + n_v = n$.

5. Partition the weighting matrix $\hat{W}$ obtained in step 3, in $\hat{W}_{c_t}$ and $\hat{W}_{c_v}$ that match the test sets ($c_t$) and validation ($c_v$).

6. Estimate $\theta_{c_t} = \theta(\nu)_{c_t} = (\beta_0, \beta_1, \beta_2, \sigma^2)^T$ via formulas of step CM2 of the EM type algorithm, considering the training data set $X_{n_t}$ and $y_{n_t}$, $\gamma = \gamma(\nu)$ obtained in step 1 and $\hat{W}_{c_t}$ from the previous step.

7. Estimate the average forecast error considering the validation set from $\hat{\Upsilon}_\nu(c_v) = (\sum_{i=1}^{n_v} \hat{u}_i)^{-1/2} \| \hat{W}_{c_v}^{1/2} y_{n_v} - \hat{W}_{c_v}^{1/2} X_{n_v} \hat{\beta}(\nu)_{c_t} \|^2$.

8. Repeat steps 6 and 7 for all $\nu \in A$, such that $A = \{1 \leq \nu \leq \nu_{max}, \nu \in \mathbb{N}\}$.

9. Repeat steps 4, 5 and 8 for all $N$ random subsamples of size $n_v$ defined and calculate $\hat{\Upsilon}_\nu^{Total}(c_v) = \frac{1}{N} \sum_{k=1}^{N} \hat{\Upsilon}_\nu(c_v)$.

10. Select the model with the smallest $\hat{\Upsilon}_\nu^{Total}(c_v)$.

The following considerations are applied to the application of the proposed CV method.
Unlike the proposal of Markatou, Afendras e Agostinelli (2018), the calculation of weights is performed for all \( \nu \in A \) as part of the EM-type algorithm and then fixed when selecting an estimate (steps 1 and 2), these are associated with the Mahalanobis distance and consequently with these estimates.

The structure of matrix \( X \) is dependent on the parameter associated with the change-point, \( \gamma \), thus, this structure is fixed for the final steps of the algorithm to optimize the other parameters by evaluating the forecast.

The selection of models is made on the set \( A = \{1 \leq \nu \leq \nu_{\text{max}}, \nu \in \mathbb{N}\} \) where \( \nu \) is the degree of freedom of the evaluated Student-t. Integer values between 1 and \( \nu_{\text{max}} = 30 \) are evaluated. Remember that when \( \nu \to \infty \), normal distribution is considered an alternative.

The CV Monte Carlo method, proposed by Shao (1993), is considered. It uses \( N \) subsamples as an alternative to making the \( \binom{n}{n_{\nu}} \) combinations with size \( n_{\nu} \) without replacement.

Following the suggestion of Ronchetti, Field e Blanchard (1997), \( n_{t} \approx n^{3/4} \) is considered.

## 4 SIMULATION STUDIES

The objective is to verify if we can recover the values of the real parameters when we use the proposed EM-type algorithm, adjusting the Student-t linear regression model with change-point, without discontinuity, to the data that were artificially generated. In this case, we generated 500 samples of size \( n = 25, 50, 100 \) and 200 from Student-t linear regression model with continuous change-point, considering \( \gamma = 4, 6 \) and 8. The explanatory variable used was \( x_{i} \sim U(1, 10) \) and random errors \( \epsilon_{i} \sim t(0, 1, 3) \), with the following values for the parameters: \( \beta_{0} = 2, \beta_{1} = -1 \) and \( \beta_{2} = 3 \).

Figure 1 shows the boxplots of the parameter estimates of the Student-t linear regression model with simulated change-point in the position \( \gamma = 8 \). For the other cases, the results are similar, so they will not be shown here to save space. It is observed that, in general, the bias and variability of the parameter estimates decrease when the sample size increases, as expected. That is, this agrees with the properties asymptotic effects of maximum likelihood estimators: asymptotic consistency and efficiency. In addition, the mean, median values and the corresponding estimates deviations standard obtained via the EM-type algorithm were calculated, the results are shown in Table 1, considering \( n = 100 \) and 200. Note that the proposed algorithm is efficient for the estimation of the parameters to the Student-t linear
regression, in the context of change-point, considering different shift point positions and sample sizes.

Figure 1: (a)-(e) *Boxplots* of the parameter estimates (line indicates the true values of the parameters) of the Student-t linear regression model with simulated change-point at \( \gamma = 8 \), considering \( n = 25, 50, 100 \) e 500. Labels are on the panel (a).

Source: Authors

5 APPLICATION IN TELEVISION AUDIENCE DATA

A study to measure television audiences using two different measuring instruments was conducted by means of data collected during several days. In a specific period \( i \) two audience measurements are obtained \((\text{Aud}_1i, \text{Aud}_2i), i = 1, \ldots, n\), where \( \text{Aud}_1i \) is the audience measured by instrument 1 and \( \text{Aud}_2i \) is the audience measured by instrument 2. The number of periods evaluated \((n)\) during the study was 2660 periods.

With this data set, a regression model that represents the relationship between audience values...
Table 1: Mean, Median (Med) and standard deviations (SD) of estimates obtained via EM algorithm based on 500 samples from the linear Student-t regression, with continuous change-point. The true values of parameters are in parentheses. Labels - Case 1: $\gamma = 4$, Case 2: $\gamma = 6$ and Case 3: $\gamma = 8$.

| Parameters | n=100 | Case 1 | Case 2 | Case 3 | n=200 | Case 1 | Case 2 | Case 3 |
|------------|-------|--------|--------|--------|-------|--------|--------|--------|
| $\beta_0(2)$ | Mean | SD | Med | Mean | SD | Med | Mean | SD | Med |
|             | 1.890 | 0.766 | 1.898 | 2.009 | 0.480 | 2.019 | 2.077 | 0.350 | 2.087 |
| $\beta_1(-1)$ | Mean | SD | Med | Mean | SD | Med | Mean | SD | Med |
|             | -0.952 | 0.292 | -0.953 | -1.017 | 0.127 | -1.005 | -1.024 | 0.078 | -1.021 |
| $\beta_2(3)$ | Mean | SD | Med | Mean | SD | Med | Mean | SD | Med |
|             | 2.972 | 0.298 | 2.971 | 2.995 | 0.224 | 2.981 | 2.873 | 0.620 | 2.912 |
| $\sigma^2(1)$ | Mean | SD | Med | Mean | SD | Med | Mean | SD | Med |
|             | 0.953 | 0.239 | 0.937 | 0.945 | 0.239 | 0.919 | 0.957 | 0.236 | 0.939 |
| $\nu (3)$ | Mean | SD | Med | Mean | SD | Med | Mean | SD | Med |
|             | 3.075 | 0.755 | 3.000 | 3.077 | 0.771 | 3.000 | 3.094 | 0.758 | 3.000 |
| $\gamma$ | Mean | SD | Med | Mean | SD | Med | Mean | SD | Med |
|             | 4.069 | 0.225 | 4.038 | 5.975 | 0.186 | 5.986 | 7.823 | 0.440 | 7.922 |

Source: Authors

obtained by the two instruments should be proposed. The model will be used in the complete audience database and to predict new audiences via instrument 2 when necessary.

The measures presented by the two instruments show relationships that change when an audience level is reached and the presence of measurements with extreme values is frequent in the data set. Thus, a regression model with a change-point that describes the relationship between the audience values obtained by instruments 1 and 2 is proposed as:

$$Aud_{2i} = \beta_0 + \beta_1 Aud_{1i} + \beta_2 (Aud_{1i} - \gamma)_+ + \varepsilon_i$$

(12)

It is considered $\varepsilon_i$ distributed as Student-t and normal to evaluate the weighting and non-weighting of outliers. The proposed model follows all the specifications described in Section 2 and the weighted cross-validation method, described in section 3, is used to obtain the parameters. Table 2 presents estimated parameter values in two situations: (i) via EM-type algorithm in the total data set and which maximizes the log-likelihood function, and (ii) via the cross-validation method that minimizes the prediction error, with $N = 200$ subsamples and $n_t = 370$.

It is observed that in the two evaluated situations, the Student-t model is always better than...
Table 2: Estimates of parameters (Est) and standard deviation of the estimates (SD(Est)) obtained via type-EM algorithm (EM) and by cross-validation (CV) for linear regression Student-t (t) and Normal (N), with a continuous change-point.

| Parameters | EM  | VC  |
|------------|-----|-----|
|           | t   | N   | t   | N   |
| β₀        | 6.166 | 6.743 | 6.863 | 2.468 | 6.796 | 2.378 |
| β₁        | 0.693 | 0.681 | 0.658 | 0.093 | 0.681 | 0.081 |
| β₂        | 0.142 | 0.150 | 0.173 | 0.106 | 0.149 | 0.096 |
| σ²        | 42.290 | 45.765 | 42.419 | 2.927 | 45.872 | 3.388 |
| γ         | 32.886 | 32.967 | 29.690 | 32.967 |       |       |
| ν         | 26   | 13   |       |       |       |       |
| log-likelihood | -8857.890 | -8859.192 | -8858.841 |       |       |       |
| Prediction error | 42.685 | 42.664 | 46.206 |       |       |       |

Source: Authors

the normal model. In situation (i), the highest log-likelihood value occurs when ν = 26 and in the situation (ii) the prediction error is minimal when ν = 13. The values of the parameters are similar and the cross-validation method allows the calculation of confidence intervals for the parameters directly using Est and SD(Est).

Figure 2 shows the model adjusted to the analyzed data. Note that the turning point is necessary for a better fit of the model. Additionally, the weights used in the robust validation method are presented, indicating that there is a large number of observations with small weights which characterizes extreme values, it is important to stand out that when considering the normal classic model the weights are not differentiated (these receive the value 1).

6 CONCLUSIONS

In this work, a regression model with change-point robust to atypical or outliers values was developed, based on SMN distributions as an alternative to the normal distribution and that presents better results in the adjustment. The inclusion of the weighted cross-validation method allows us to adjust a model that improves the prediction results as well as presenting confidence intervals for the parameters estimated in the right way. Future work will be oriented towards the use of robust asymmetric distributions and the inclusion of more than one change-point in the model. On the other hand, it is possible to consider a method of selecting the best model from a combination of a log-likelihood function (Akaike Information Criterion or other) that assesses the estimation in the model and the forecast error obtained from the validation sample.

Revista Mundi, Engenharia e Gestão, Paranaguá, PR, v. 5, n. 2, p. 233-01, 233-15, 2020  DOI: 10.21575/25254782rmetg2020vol5n21179
Figure 2: (a) Scatter plot with Student-t linear regression model with the proposed change-point (b) Weighting values for audience values measured by instrument 2.

Source: Authors

ACKNOWLEDGMENTS

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001. The authors would like to thank Kantar Ibope Media for making data available for analysis. Camila B. Zeller thanks the support of FAPEMIG and CNPq.

REFERENCES

ANDREWS, D. F.; MALLOWS, C. L. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 36, n. 1, p. 99–102, 1974.

BOUSQUET, O.; ELISSEEFF, A. Stability and generalization. *Journal of machine learning research*, v. 2, n. Mar, p. 499–526, 2002.

CHEN, J.; GUPTA, A. K. On change point detection and estimation. *Communications in statistics-simulation and computation*, Taylor & Francis, v. 30, n. 3, p. 665–697, 2001.

CHEN, J.; GUPTA, A. K. *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. [S.l.]: Springer Science & Business Media, 2011.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977.
HOFRICHTER, J. Change point detection in generalized linear models. Dissertation zur erlangung des akademischen grades doktor der technischen wissenschaften, Graz University of Technology, 2007.

HUAIRA-CONTRERAS, C. A. Modelo de regressão linear mistura de escala normal com ponto de mudança: estimação e diagnóstico. Dissertação de Mestrado, IMECC-Universidade Estadual de Campinas, 2014.

LANGE, K.; SINSHEIMER, J. S. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, Taylor & Francis Group, v. 2, n. 2, p. 175–198, 1993.

LANGE, K. L.; LITTLE, R. J.; TAYLOR, J. M. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 84, n. 408, p. 881–896, 1989.

LUCAS, A. Robustness of the student t based m-estimator. *Communications in Statistics- Theory and Methods*, Taylor & Francis, v. 26, n. 5, p. 1165–1182, 1997.

MARKATOU, M.; AFENDRAS, G.; AGOSTINELLI, C. Weighted cross validation in model selection. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 10, n. 6, p. e1439, 2018.

MENG, X.-L.; RUBIN, D. B. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, Oxford University Press, v. 80, n. 2, p. 267–278, 1993.

MUGGEO, V. M. Estimating regression models with unknown break-points. *Statistics in medicine*, Wiley Online Library, v. 22, n. 19, p. 3055–3071, 2003.

OSORIO, F. Diagnóstico de influência em modelos elípticos com efeitos mistos. PhD dissertation, USP, São Paulo, 2006.

PICARD, R. R.; COOK, R. D. Cross-validation of regression models. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 79, n. 387, p. 575–583, 1984.

RONCHETTI, E.; FIELD, C.; BLANCHARD, W. Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, Taylor & Francis, v. 92, n. 439, p. 1017–1023, 1997.

ROSA, G.; PADOVANI, C. R.; GIANOLA, D. Robust linear mixed models with normal/independent distributions and bayesian mcmc implementation. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, Wiley Online Library, v. 45, n. 5, p. 573–590, 2003.

SHAO, J. Linear model selection by cross-validation. *Journal of the American statistical Association*, Taylor & Francis Group, v. 88, n. 422, p. 486–494, 1993.

YAMAGUCHI, K. Analysis of repeated measurements with multivariate t or contaminated multivariate normal errors. *Bulletin of the Computational Statistics of Japan*, v. 3, n. 1, p. 11–22, 1990.
YOUNG, D. S. Mixtures of regressions with changepoints. *Statistics and Computing*, Springer, v. 24, n. 2, p. 265–281, 2014.