CHANGE POINT ESTIMATION IN SEGMENTED REGRESSION MODEL USING RSTAN: AN APPLICATION TO EXERCISE PHYSIOLOGY DATA

Silvio Cabral Patrício[1]

Renato Valladares Panaro[2]

Adriana dos Santos Lima[3]

Patrícia Viana da Silva[4]

Resumo

Esse trabalho traz uma discussão a respeito do uso de dois métodos de estimação para o modelo de Regressão Segmentada com pontos de mudança desconhecidos. Os parâmetros serão estimados no contexto de inferência Bayesiana e comparados com o modelo frequentista desenvolvido por Muggeo (2003). As abordagens serão aplicadas a dados de fisiologia do exercício em que existe o interesse sobre os pontos de mudança no consumo de oxigênio (VO2) por atletas em função da velocidade na esteira. O desempenho dos métodos é investigado em modelos simulados considerando diferentes quantidades de pontos de mudança.

Palavras-chave: regressão segmentada, pontos de mudança, inferência Bayesiana, NUTS, Rstan.

Abstract

This work brings a discussion about the use of two estimation methods for the Segmented Regression model with unknown points of change. The parameters will be estimated in the context of Bayesian inference and compared with the frequentist model developed by Muggeo (2003). The approaches will be applied to physiological data of the exercise in which there is interest in the points of change in oxygen consumption (VO2) by athletes as a

[1] Universidade Federal de Minas Gerais – UFMG, silviocp@ufmg.br

[2] Universidade Federal de Minas Gerais – UFMG, renato@ufmg.br

[3] Universidade Federal de Minas Gerais – UFMG, adrianalima@ufmg.br

[4] Universidade Federal de Minas Gerais – UFMG, Universidade Federal de Uberlândia – UFU, patriciaviana@ufu.br

function of the speed on the treadmill. The performance of the methods is investigated in simulated models considering different amounts of change points.

Keywords: regression segmented, change point, Bayesian inference, NUTS, Rstan.

Introduction

In many practical applications, it is necessary to investigate the effect of one or more characteristics on a variable of interest (response) which requires linear multiple regression models. For this, it is assumed that the response variable depends linearly on covariates (characteristics) in relation to unknown parameters. However, the data may violate the linear relationship assumption in a number of ways. For these cases, the usual multiple linear regression should not be used. See Quandt (1958), Robson (1964). A common situation of breaking the assumption is that in which abrupt changes occur in the linear relationship with the response variable after certain values (change points) in the covariables, while it remains linear between the consecutive thresholds. The behavior of this type of data can be modeled using the so-called Segmented Regression Model.

In Segmented Regression, the points of change are estimated and are not previously fixed (Muggeo, 2003). Moreover, the regression parameters obtained are directly interpretable, which does not occur in other approaches for nonlinear effects such as polynomial regression, splines, and non-parametric smoothing. Nevertheless, standard inference based on likelihood is complicated because the likelihood function is only differentiated by parts and the classical conditions of regularity are not satisfied (Feder, 1975, Küchenhoff and Carroll, 1997, Seber & Wild, 1989). Also, change points are non-linear parameters and standard maximization procedures cannot be used (Seber & Wild, 1989).

In this work, we will use the Muggeo (2003) approach to describe the model, which reduces the problem to an usable linear structure and facilitates the estimation of the points of change and can be applied in different practical situations. To illustrate, we consider the circumstance that appears in the field of exercise physiology. In ramp test protocols, oxygen consumption (VO2) is assessed breath-to-breath, increasing the exercise load at a constant rate and at equidistant time intervals for one athlete. Consistent with the existing literature, there are up to two points of change (up to three straight segments) and this experiment produces behaviors that present one or two straight segments with a positive slope and a final segment with a zero slope (Abreu *et al.*, 2017). Hill and Lupton (1925) described that

oxygen consumption increases linearly with increasing velocity until reaching a maximum value, this justifies the use of a linear model.

The first segment generally has a strong positive slope for the physiological adaptation at the beginning of the exercise, while the second segment (if any) is characterized by an attenuated positive slope due to the linear progression of VO2 consumption with respect to exercise load. In both cases, the last linear segment determines the plateau at which VO2 consumption stabilizes (Abreu *et al*., 2017).

The frequentist approach to the problem has already been addressed by Muggeo (2003), and is well grounded. The same does not happen with the Bayesian approach. In addition, a common situation is when there are not the two breakpoints in the observed data (BRINK-ELFEGOUN *et al*., 2007; LIMA-SILVA *et al*., 2006; MYERS *et al*.,1989), that is, when the athlete does not reach his plateau. This is difficult to identify and arises the need for a model that can address this problem one possibility is the Bayesian approach.

Objective

This work aims to estimate the parameters of the Segmented Regression using two methods: the frequentist model adopted by Muggeo (2003) and Bayesian inference approach. Both methods were applied to actual exercise physiology data and the performance is also evaluated through simulation of models with a varying number of change points. The Bayesian results are accessed through Rstan, a R package that allows interface with Stan. This software uses No-U-Turn Sampler (NUTS), an extension to Hamiltonian Monte Carlo that eliminates the need to set a number of steps (Hoffman, 2014). The performance of the estimation is also evaluated through simulation of models with a varying number of change points from Mean Absolute Percentage Error (MAPE).

Material and Method

In the literature, multiple linear regression is widely used to analyze the effect of p explanatory variables $\boldsymbol{X} = (X_1, \ldots, X_p)'$, on a response variable, Y. In his context, n observations of vector $(\boldsymbol{Y}, \boldsymbol{X})$ are taken and the following linear model is set:

$$y_i = \beta_0 + X'_i \beta_1 + \varepsilon_i, i = 1, \ldots, n, \tag{1}$$

where y_i and X_i represent, respectively, the values of Y and $\mathbf{X}$ in i-th replica (or subject), β_0 is the intercept, $\boldsymbol{\beta_1} = (\beta_{11}, \ldots, \beta_{1p})$ is the coefficient vector that describes the effect of $\mathbf{X}$ in Y,

3

and ε_i is a random error of i-th or subject, under the assumptions $\varepsilon_i \sim N(0, \sigma^2)$ and $Cov(\varepsilon_i, \varepsilon_i) = 0, i \neq j$, normally distributed and uncorrelated. This is used to fit $\mu = E[Y|X]$, the mean of Y, where $y_i = \mu + \varepsilon_i, i = 1, \dots, n$, and $\mu = \boldsymbol{\beta_0} + \boldsymbol{X'\beta_1}$.

Moreover, it is assumed that the effect of X on Y has a single linear behavior, if not, this linear model may not be able to represent the data. An example is the relationship between the Oxygen Consumption (VO2 in l/min) and the exercise intensity practiced with successive increments of load, Figure 1. Then the evolution of the oxygen consumption of a runner is a function of the speed of the conveyor (km/h), which is continuously increased with speed increments at equally spaced time intervals but in up to three different segments.
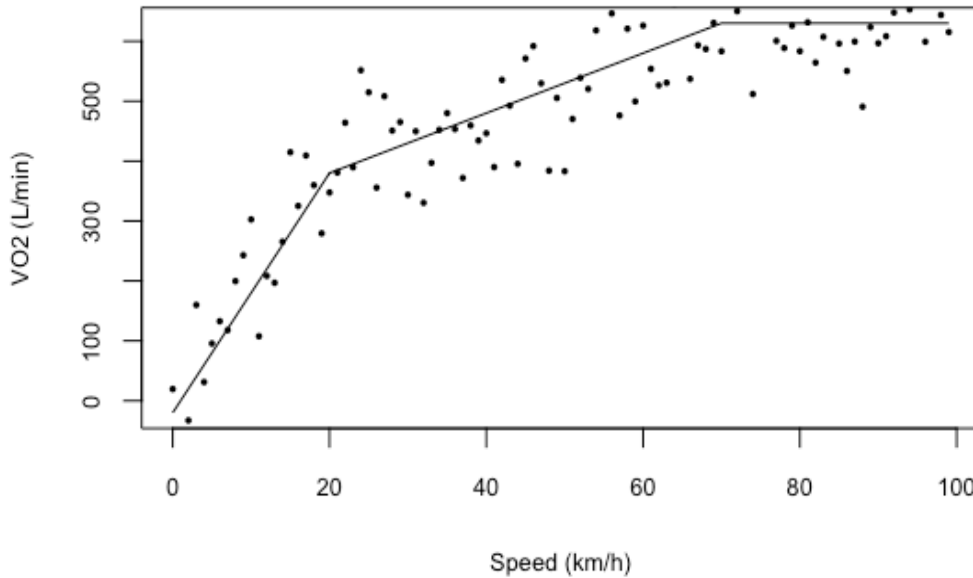

Figure 1: Oxygen Consumption, VO2 (l/min) by speed (km/h) in a treadmill, simulated data.

The general model for Y is given by

$$Y_i = \boldsymbol{\beta_0} + \boldsymbol{X'\beta_1} + \boldsymbol{\beta_2}(X_i - \varphi_1)I_{\{X > \varphi_1\}} - (\boldsymbol{\beta_1} + \boldsymbol{\beta_2})(X_i - \varphi_2)I_{\{X > \varphi_2\}} + \boldsymbol{\varepsilon}, \quad (2)$$

where β_0 is the intercept is the intercept of the first segment, β_1 is the slope of the of the first segment, β_2 is the increase between the slope of the first and the second segment, and ϕ_1 and ϕ_2 are the first and the second change point, respectively. Especially, for the first segment the model is given by

$$y_i = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} X_i{}' + \boldsymbol{\varepsilon_i}. \quad (3)$$

Similarly, the model for the second segment (if any) is given by

$$y_i = (\boldsymbol{\beta_0} - \boldsymbol{\beta_2}\varphi_1) + X_i{}'(\boldsymbol{\beta_1} + \boldsymbol{\beta_2}) + \boldsymbol{\varepsilon_i}. \quad (4)$$

And the model for the last segment, is given by a constant because of the null slope

$$y_i = \beta_0 + \beta_2 \varphi_1 - (\beta_1 + \beta_2)\varphi_2 + \varepsilon_i. \tag{5}$$

Also, there is the parameter σ^2 that must be estimated in the problem. Thus the likelihood function, $L(\cdot) = \prod_{i=1}^{n} f(y_i | \cdot)$, is

$$L(\beta_0, \beta_1, \beta_2, \varphi_1, \varphi_2, \sigma^2) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2} \right) \right] \tag{6}$$

where $\mu = \beta_0 + X'\beta_1 + \beta_2(X_i - \varphi_1)I_{\{X>\varphi_1\}} - (\beta_1 + \beta_2)(X_i - \varphi_2)I_{\{X>\varphi_2\}}$.

For the frequentist approach, the estimation is done via maximum likelihood and the estimators are obtained analytically via partial derivatives of $L(\cdot)$. The R package **segmented** (Muggeo, 2008) is used to get the estimates.

Bayesian estimators are obtained by mean of the posterior distributions (considering square loss). Therefore, it is necessary to specify a priori distributions for the six parameters of the model:

$$\beta_0 \sim N(m_0, s_0), \beta_1 \sim N(m_1, s_1), \beta_2 \sim N(m_2, s_2),$$
$$\varphi_1 \sim Unif(0, x_{(n)}), \varphi_2|\varphi_1 \sim Unif(\varphi_1, x_{(n)}), \tag{7}$$
$$\sigma^2 \sim Gamma(a, b)$$

where $x_{(n)}$ is the highest value of the covariate and $m_0, s_0, m_1, s_1, m_2, s_2, a$ e b are constants given.

After specifying the priori distributions the posterior distributions and estimates are found using numerical MCMC methods via R package (R Core Team, 2018) **Rstan** (Stan Development Team, 2018) for the exercise physiology data. The convergence criterion was evaluated by the **coda** R package (Plummer at al., 2006).

One of the motivations for using this model is to apply it to exercise physiology data. The specific objective is to determine velocity upon reaching the VO2 plateau. A common phenomenon is an observed athlete does not reach the plateau, which would imply in changing the number of change points of the model, which is at most two. Therefore, we will study the behavior of the parameters of the model (2) in the scenarios where we have two, one or no point of change. The performance of the model proposed by Muggeo (2003) and Bayesian, proposed in this article, will be compared.

Results and Discussion

The results of the analyzes and simulations were obtained by means of R packages. Exercise physiology data from a ramp test with an athlete in which the oxygen consumption (VO2) was analyzed as a function of the exercise load (speed), which undergo increments at equidistant intervals of time.

Application to actual data

In Figure 2, we have the observed data refer to an athlete submitted to cardiopulmonary exercise test, an exam performed to identify VO2 plateau. The test is performed on a treadmill or exercise bicycle, where the load is increased in linear function from time to voluntary exhaustion of the athlete. 263 measurements of VO2 consumption (l/min) and speed controlled in the experimente of one athlete were obtained. The points of change are not quickly identified visually, but those responsible for the experiment believe that the athlete reached the plateau and VO2 consumption presents the three segments present in what is expected of the model containing two points of change.
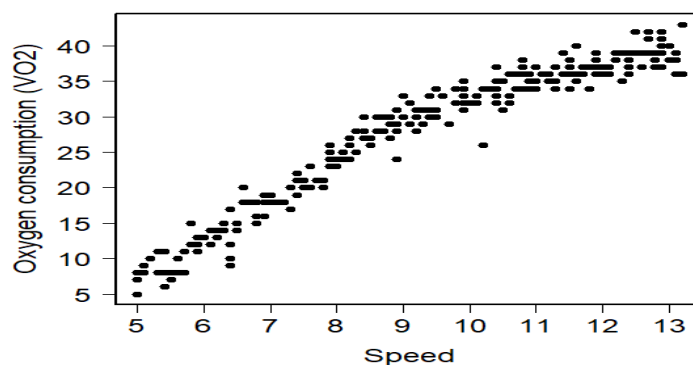

Figure 2: Scatter plot of speed (km/h) × VO2 (l/min).

Table 1 and Figure 3 presents the results of the estimates for the parameters of the model with two points of change for the Bayesian and Muggeo model. Both present estimates and similar variations for all parameters. The estimates of the standard deviation of the frequentist model presented are obtained via bootstrap, since the parameters φ_1 and φ_2 can not be estimated directly. This problem does not occur with the Bayesian model.

Table 1: Estimated parameters for the data.

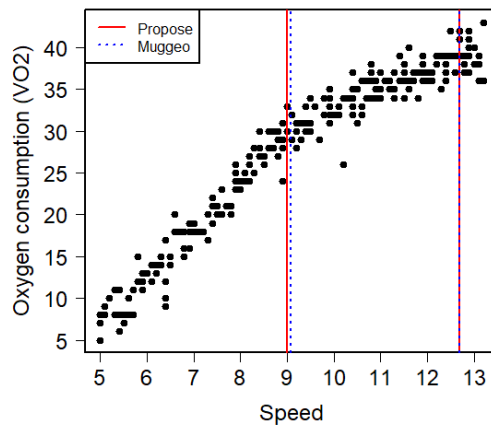| Method		β_0	β_1	β_2	φ_1	φ_2	σ
Muggeo	Estimative	-23.53	5.97	-3.53	9.09	12.85	1.70
	S.d.	0.38	0.14	0.20	0.13	0.20	0.25
Bayesian	Estimative	-23.48	5.96	-3.51	9.00	12.69	1.73
	S.d.	1.01	0.14	0.20	0.14	0.20	0.08

Figure 3: Comparation between Bayesian and Muggeo estimations.

Next, in Table 2, the CI, confidence interval of 95%, and the HPD credibility interval of 95% probability for the parameters for the Bayesian are presented.

Table 2: Confidence Interval for Muggeo model and High Posterior Density Credibility Interval for Bayesian model.

	β_0		β_1		β_2		φ_1		φ_2		σ	
CI(95%)	-25,44	-21,61	5,69	6,23	-3,92	-3,13	8,84	9,34	12,47	13,24	1,57	1,86
HPD(95%)	-25.41	-21.52	5.69	6.24	-3.90	-3.13	8.67	9.25	12.28	13.09	1.58	1.88

Convergence was ascertained for the six parameters by using 2 chains with different starting values. After a burn-in period of 1,000 draws the sample consisted of 1,000 draws. See Figure 4.
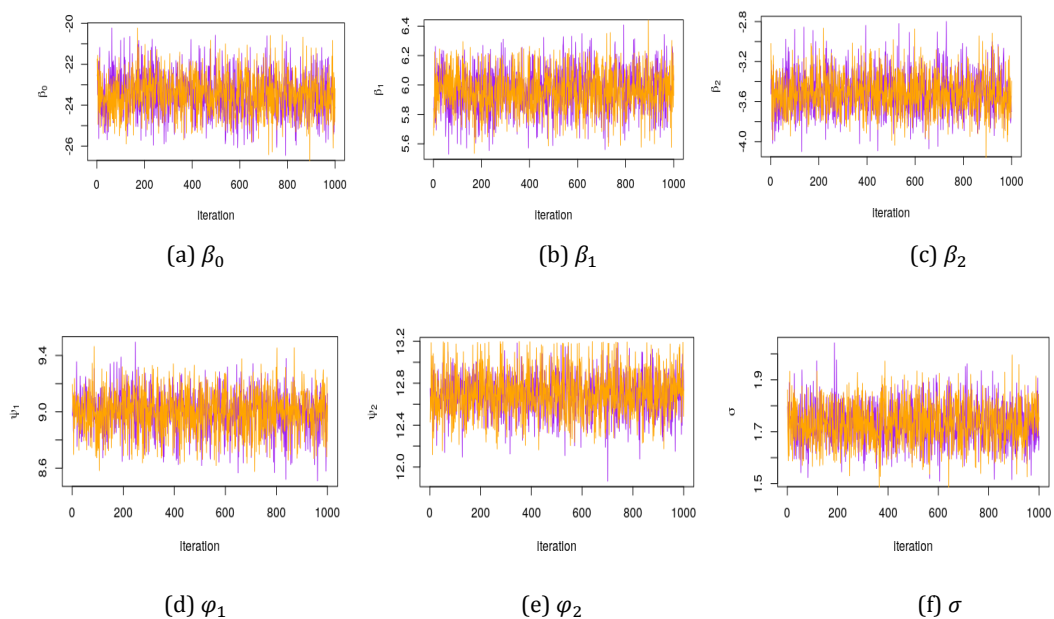

(a) β_0

(b) β_1

(c) β_2

(d) φ_1

(e) φ_2

(f) σ

Figure 4: Chain convergence for the parameters.

Simulated data

All the scenarios have the same settings. 100 replicated data sets were created of length 100; the speed was evaluate from 0 to 100; the parameters was fixed at $\beta_0 = -20$, $\beta_1 = 20$, $\beta_2 = 5$ and, where applicable, $\varphi_1 = 20$ and $\varphi_2 = 70$. To the Bayesian model the prior distribution for the parameters was: $\sigma \sim Gamma(0.01; 0.01)$; $\beta_i \sim N(0, 100^2)$, $i = 0, 1, 2$; $\varphi_1 \sim U(0, \ x_{(n)} = 100)$ and $\varphi_2 | \varphi_1 \sim U(\varphi_1, \ x_{(n)} = 100)$.

It can be seen from Table 3, that MAPE values indicate the Bayesian model with better results for practically all scenarios. Especially for parameters β_0, β_1 and β_2, much more advantageous for β_0, where the error can be ten times greater.

Table 3: MAPE (%) of the simulation results for multiples scenarios.

	Nº Change Points	Method	β_0	β_1	β_2	φ_1	φ_2	σ
$\sigma = 5$	K = 0	Muggeo	15.43	2.36	-	-	-	5.47
		Bayesian	3.73	0.18	-	-	-	5.46
	K = 1	Muggeo	98.58	6.67	8.23	0.71	-	6.13
		Bayesian	8.63	0.67	0.91	0.68	-	5.31
	K = 2	Muggeo	163.45	11.89	14.03	0.75	0.39	6.65
		Bayesian	8.23	0.72	0.97	0.67	0.42	6.15
$\sigma = 25$	K = 0	Muggeo	99.21	8.12	-	-	-	5.73
		Bayesian	18.38	0.38	-	-	-	4.78
	K = 1	Muggeo	125.76	9.37	11.77	4.24	-	5.78
		Bayesian	40.38	3.68	5.07	3.15	-	5.24
	K = 2	Muggeo	213.42	14.32	14.77	3.79	1.95	5.68
		Bayesian	47.64	4.21	5.55	4.30	1.94	5.48
$\sigma = 50$	K = 0	Muggeo	155.63	16.13	-	-	-	6.44
		Bayesian	38.16	0.75	-	-	-	6.13
	K = 1	Muggeo	218.67	17.95	22.99	8.07	-	6.01
		Bayesian	77.79	8.44	11.07	7.69	-	4.98
	K = 2	Muggeo	234.44	18.72	20.75	7.36	4.19	5.96
		Bayesian	83.35	8.44	11.05	7.46	3.94	6.46

In addition, it was found that the model for not being prepared to actually identify the number of change points estimated all parameters. However, again the Bayesian model presents advantage because it estimates the points of change non-existent always at the end of the scale of values of the covariable, while the estimates of Muggeo have a large variation in any value of the scale of the covariate, including the beginning. See Table 4.

Table 4: Summary of the estimates of change points in scenarios with less then three segments.

k	Parameter	$\sigma = 5$			$\sigma = 25$			$\sigma = 50$		
		1º Q	Median	3º Q	1º Q	Median	3º Q	1º Q	Median	3º Q
0	φ_1	98.57	98.98	99.21	96.89	97.68	98.10	95.31	96.29	97.07
	φ_2	99.52	99.60	99.70	98.64	98.95	99.11	97.86	98.32	98.68
1	φ_2	98.98	99.25	99.41	96.49	97.20	97.82	93.31	94.95	96.20

Conclusions

Both models are able to estimate the parameters from the application data. Estimation values were close, including CI and HPD. For the simulated data, the Muggeo model suffers with extremely large values of MAPE and gives advantage to the Bayesian model which, in addition to presenting smaller relative errors, does not estimate the points of change that do not exist in any location of the scale of values of the covariate, takes them to the end of the scale. Thus, it provides a mechanism to identify cases in which the athlete did not reach the plateau.

The test used in the application is performed on a treadmill or exercise bicycle, where the load is increased in linear function from time to voluntary exhaustion of the athlete. Therefore, there is correlation in the data, usually an AR (1), and we are approaching this problem in a new work also with Bayesian approach.

References

ABREU, V. C., CARLETTI, L., SARNAGLIA, A. J. Q., and PEREZ, A. J. (2017). **Systematic review of criteria for identification of plateau maximum oxigen consumption**. Submitted to J. Phys. Educ.

BRINK-ELFEGOUN, T.; HOLMBERG, H.; EKBLOM, M. N.; EKBLOM, B. **Neuromuscular and circulatory adaptation during combined arm and leg exercise with different maximal workloads**. European Journal of Applied Physiology, Berlim, 101(5), 603-611. 2007.

FEDER, P. I. **On asymptotic distribution theory in segmented regression problems–identified case**. The Annals of Statistics, 3 (1), 49–83. 1975.

HILL, A. V; LUPTON, H. **Muscular exercise, lactic acid, and the supply and utilization of oxygen**. Ergebnisse der Physiologie, 24(1), 43-51. 1925.

HOFFMAN, M. D. and GELMAN, A. **The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.** Journal of Machine Learning Research 15 (1), 1593-1623. 2014.

KÜCHENHOFF, H. and CARROLL, R. J. **Segmented regression with errors in predictors**: semi-parametric and parametric methods. Statist. Med., 16: 169-188. 1997.

LIMA-SILVA, A. E.; BERTUZZI, R.C.M.; GEVAERD, M.S.; OLIVEIRA, F.R. **Identificação do platô do consumo de oxigênio com diferentes tempos de análise**. Revista Brasileira de Ciência e Movimento, 14(1), 7-12. 2006.

MYERS, J.; BUCHANAN, N.; WALSH, D.; KRAEMER, M.; McAULEY, P.; HAMILTON- WESSLER, M.; FROELICHER, V. F. **Comparison of the ramp versus standard exercise protocols**. Journal of the American College of Cardiology, 17(6), 1334-1342. 1991.

MUGGEO, V. M. R. **Estimating regression models with unknown break-points.** Statistics in Medicine, 22 (19), 3055-3071. 2003.

MUGGEO, V. M. R. **segmented:** an R Package to Fit Regression Models with Broken-Line Relationships. R News, 8/1, 20-25. 2008.

PLUMMER, M.; BEST, N.; COWLES, K. and VINES, K. **CODA: C**onvergence Diagnosis and Output Analysis for MCMC**,** R News, 6, 7-11. 2006.

QUANDT, R. E. **The estimation of the parameters of a linear regression system obeying two separate regimes**. Journal of the american statistical association, 53(284), 873–880. 1958.

ROBSON, D. E. **Estimates for the points of intersection of two polynomial regressions**. Journal of the American Statistical Association, 59(305), 214–224. 1964.

R Core Team. **R:** A language and environment for statistical computing**.** R Foundation for Statistical Computing, Vienna, Austria. 2018.

SEBER, G. A. F. and WILD, C. J. **Nonlinear Regression**. New York: Wiley, 768 pages. 1989.

Stan Development Team. **RStan:** the R interface to Stan. R package version 2.18.2. http://mc-stan.org/. 2018.

Appendix A – Stan code

```
data{
  // setting observed data:
    int<lower=1> n;            // sample length
    vector[n] x;       // independent variable
    vector[n] y;       // dependent variable
    real tau;               // max(x)


    //hyperparameters:
      real<lower=0> a_sigma;
      real<lower=0> b_sigma;
      real<lower=0> mu_beta;
      real<lower=0> s_beta;
}


parameters{
  vector[3] beta;               // beta: vector of length 3
  real<lower=0> psi2;              // psi2: positive real number
  real<lower=0, upper = psi2> psi1;   // psi1: real number, less than psi2
  real<lower=0> sigma;              // sigma: positive real number
}
```

```
model{
  // likehood
  for(i in 1:n){
    if(x[i] < psi1){
      y[i] ~ normal( beta[1] + beta[2]*x[i], sigma);
    }
    else if(x[i] < psi2){
      y[i] ~ normal( beta[1] + beta[2]*x[i] + beta[3]*(x[i]-psi1), sigma);
    }
    else{
      y[i] ~ normal( beta[1] + beta[2]*x[i] + beta[3]*(x[i]-psi1)
             - (beta[2]+beta[3])*(x[i]-psi2), sigma);
    }
  }
    // priori especification
    beta ~ normal(mu_beta,s_beta);
    sigma ~ gamma(a_sigma, b_sigma);
    psi1 ~ uniform(0, tau);
    psi2 ~ uniform(psi1, tau);
}
```

Appendix B – R code: Simulated Data

Bayesian model

```
source("https://github.com/rvpanaro/SER2019/blob/master/simulastan_Adriana_final.R"
)
theta = c( -20, 20, -15, 20, 70,  (sigma) ) # real parameters
MAPE = function(parametro){  # function to estimate the MAPE
  erro = apply(parametro, 1, function(x){
```

```
    out = (theta - x[-7])/theta

    return(abs(out))

  })

  final = apply(erro, 1, mean)

  return(final)

}


MAPE(param1)*100 # MAPE: simulation with 0 changepoint

MAPE(param2)*100 # MAPE: simulation with 1 changepoint

MAPE(param3)*100 # MAPE: simulation with 2 changepoint

c(sigma, fail1, fail2, fail3) # how many times we've divergente models

summary(param1[,4]) # distribution of psi1: simulation with 0 changepoint

summary(param1[,5]) # distribution of psi2: simulation with 0 changepoint

summary(param2[,5]) # distribution of psi2: simulation with 1 changepoint


tab <- rbind(MAPE(param1)*100, MAPE(param2)*100, MAPE(param3)*100)

colnames(tab) <- c("beta0", "beta1", "beta2", "psi1", "psi2", "sigma")

rownames(tab) <- c("k=0", "k=1", "k=2")

tab

write.csv(tab, "sigma50_sim.csv")
```

Muggeo model

```
source("https://github.com/rvpanaro/SER2019/blob/master/simula_muggeo.R")

K = out$K


  theta = c(-20, 20, -15, 20, 70, s)
```

```r
erro_relativo = t(apply(out[,-1], 1, function(x){

  (x-theta)/theta

})))
erro_medio = as.data.frame(t(apply(as.matrix(0:2), 1, function(k){

  idx = which(K==k)

  out = apply(abs(erro_relativo[idx,]), 2, mean)


  return(out)

})))


  # rownames(erro_medio) = 0:2

  out = data.frame(sigma = s ,k = 0:2, erro_medio)

  print(out)

  file = paste('sim_Sig_', s, '.csv', sep="")


  write.csv(out , file, row.names = F)

}
```

Appendix C – R code: Case study (exercise physiology)

```r
##------------------ Importing Dataset ------------------
rm(list=ls(all=TRUE)) # Clear the R workspace
dat <- read.csv('data\\atleta_12.csv', sep = '\t')
n <- nrow(dat)
y <- dat$vo2
x <- dat$vel


##------------------ Graphic ------------------
par( mfrow = c(1,1), lwd = 2, cex.lab = 1.5, cex.axis = 1.5, lab = c(8,6,5),
```

```
        mar = c(6,6,3,6), las = 1 )
plot(x, y, pch = 19, ylab = "Oxygen consumption (VO2)", xlab = "Speed")


##----------------- Stan -----------------
#  load rstan.
library(rstan)


# avoid recompilations
rstan_options(auto_write = TRUE)


# run different chains in parallel.
options(mc.cores = parallel::detectCores())


##----------------- Prepare data for Stan -----------------
# sigma ~ Gamma( a_sigma, b_sigma )
a_sigma = .01
b_sigma = .01


# beta ~ N( mu_beta, s_beta )
mu_beta = 0
s_beta = 100


# psi1 ~ U( 0 , tau )
# psi2 ~ U( psi1, tau )
tau = max(x)


data = list( n = n, x = x, y = y,
        a_sigma = a_sigma,  b_sigma = b_sigma,
        mu_beta = mu_beta,  s_beta = s_beta,
        tau = tau )
##----------------- Modeling -----------------
# parameters I want estimate
```

```r
pars = c("beta", "psi1", "psi2", "sigma")


## Bayesian
output = stan(file = "be_pwlm2.stan", data=data,
        pars=pars, chains = 2, verbose=FALSE)


print(output, pars=pars)          # print output
traceplot(output, pars=pars)      # chain convergence
chain = extract( output )  # extract chains
param = coda::as.mcmc( cbind( chain$beta, chain$psi1, chain$psi2, chain$sigma ) )
coda::HPDinterval(param)  # high posterior density credibility interval
##------------------ Graphic ------------------
par( mfrow = c(1,1), lwd = 2, cex.lab = 1.5, cex.axis = 1.5, lab = c(8,6,5),
        mar = c(6,6,3,6), las = 1 )
plot(x, y, pch = 19, ylab = "Oxygen consumption (VO2)", xlab = "Speed")
abline( v = c(9, 12.69), col = 2, lty = 1)
abline( v = c(9.09, 12.69), col = "blue", lty = 3)
legend('topleft', legend = c('Propose','Muggeo'), col = c(2, "blue"), lty = c(1,3), lwd=2,
bg='transparent')
```