Using the R package chngpt

Youyi Fong

July 1, 2020

1 Types of threshold effects supported by the package

We refer to models with a single threshold as two-phase models, models with two thresholds as three-phase models, and models with more than two thresholds as multi-phase models.

1.1 Continuous two-phase models

The package support the following two-phase models that are continuous at the threshold.

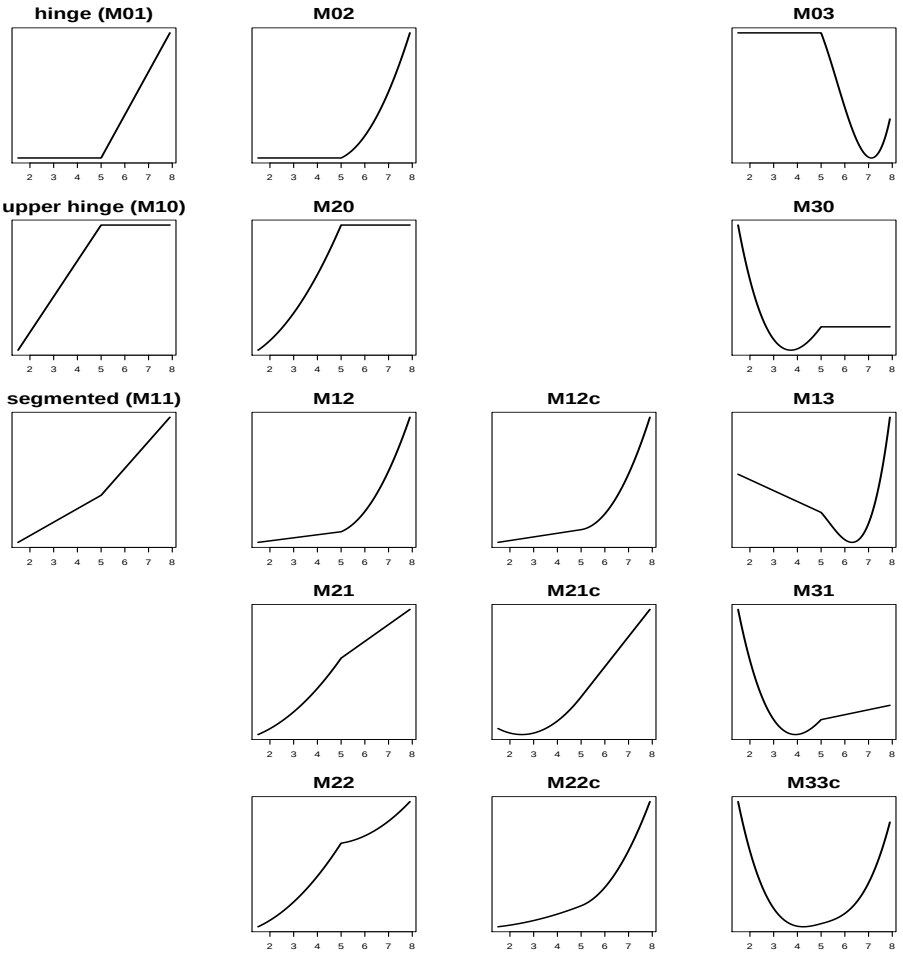

Figure 1.1: Types of continuous two-phase models supported in chngpt.

Piecewise linear two-phase models are studied in Fong et al. (2017b) and Elder and Fong (2019), two-phase polynomial models are studied in a manuscript under review. The two digits in the model names refer to the highest order of polynomials before and after the threshold, respectively. If the model name ends with 'c', the model is constrained and become smoother. The parameterization are adopted in the package:

$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x-e)_+ \quad \text{(hinge, M01)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x-e)_+ + \beta_2 (x-e)_+^2 \quad \text{(M02)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x-e)_+ + \beta_2 (x-e)_+^2 + \beta_3 (x-e)_+^3 \quad \text{(M03)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x-e)_- \quad \text{(upper hinge, M10)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x-e)_- + \beta_2 (x-e)_-^2 \quad \text{(M20)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x-e)_- + \beta_2 (x-e)_-^2 + \beta_3 (x-e)_-^3 \quad \text{(M30)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \gamma x + \beta_1 (x-e)_+ \quad \text{(segmented, M11)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \gamma x + \beta_1 (x-e)_+ + \beta_2 (x-e)_+^2 \quad \text{(M12)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \gamma x + \beta_1 (x-e)_+ + \beta_2 (x-e)_+^2 + \beta_3 (x-e)_+^3 \quad \text{(M13)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \gamma x + \beta_1 (x-e)_- + \beta_2 (x-e)_-^2 \quad \text{(M21)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \gamma x + \beta_1 (x-e)_- + \beta_2 (x-e)_-^2 + \beta_3 (x-e)_-^3 \quad \text{(M31)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_{1,-} (x-e)_- + \beta_{1,+} (x-e)_+ + \beta_{2,-} (x-e)_-^2 + \beta_{2,+} (x-e)_+^2 \quad \text{(M22)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \gamma x + \beta_{2,-} (x-e)_-^2 + \beta_{2,+} (x-e)_+^2 \quad \text{(M22c)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \gamma x + \beta_2 (x-e)^2 + \beta_{3,-} (x-e)_-^3 + \beta_{3,+} (x-e)_+^3 \quad \text{(M33c)}$$

where e denote the threshold parameter, x is the predictor with threshold effect, z denote a vector of additional predictors, and $(x-e)_+ = x - e$ if $x > e$ and 0 otherwise, and $(x-e)_- = x - e$ if $x \leq e$ and 0 otherwise.

1.2 Continuous three-phase models

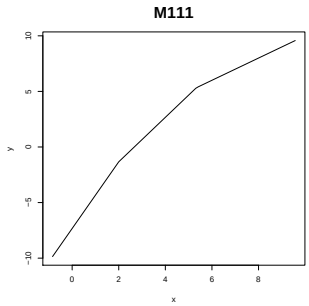

Figure 1.2: A three-phase segmented model.

$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x-e)_- + \beta_2 (x-f)_- + \beta_3 x \quad \text{(M111)}$$

1.3 Discontinuous two-phase models

The following discontinuous two-phase models are supported in the *chngpt* package:

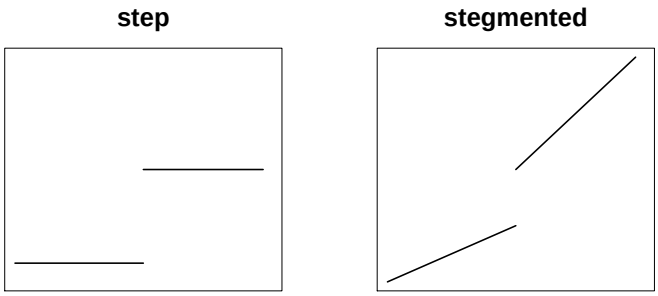

Figure 1.3: Types of discontinuous threshold effects supported in chngpt.

The models can be written as

$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 I\left(x > e\right) \qquad \text{(step)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 \left(x - e\right)_+ + \gamma x + \beta_2 I\left(x > e\right), \qquad \text{(stegmented)}$$

where e denote the threshold parameter, x is the predictor with threshold effect, z denote a vector of additional predictors, and

$$I\left(x > e\right) = \left\{ \begin{array}{lll} 1 & \text{if} & x > e \\ 0 & \text{if} & \text{otherwise} \end{array} \right. .$$

2 Estimation examples

Some general notes:

- The fitted model has a component named best.fit, which is the glm or coxph fit conditional on the estimated threshold parameter.
- The recommended ci.bootstrap.size is 1000 in real problems.

2.1 Continuous two-phase linear regression

For continuous two-phase linear regression, we have developed a grid search method for estimation that is super fast (Fong, 2019; Elder and Fong, 2019). Together with the observation that bootstrap confidence intervals have better coverage than robust analytical confidence intervals (Fong et al., 2017b) for continuous two-phase linear regression, internally we set the default estimation method to be fast grid search and the default variance method to be bootstrap.

2.1.1 Segmented model

To fit a segmented linear regression model, we call

```
fit=chngptm (formula.1=V3_BioV3B~1, formula.2=~NAb_score, dat.mtct.2, type="segmented", family="gaussian")
summary(fit)
```

```
Change point model type:  segmented

Coefficients:
                          est     p.value*    (lower    upper)
(Intercept)         -22.33152 1.593423e-08 -30.07675 -14.58628
NAb_score            67.23925 2.212981e-14  49.98398  84.49452
(NAb_score-chngpt)+ -64.83129 3.692679e-14 -81.61413 -48.04845

Threshold:
      est    (lower    upper)
0.4653923 0.4535000 0.4772845
```

In the output above, the row starting with (NAb_score-chngpt)+ corresponds to β_1 in equation (segmented, M11). In other words, it is the change in slope as the covariate NAb_score crosses the threshold. Note that we there is an asterisk next to p.value. This is because bootstrap procedures to generate confidence intervals do not readily lead to p values. The presented p values are approximations, obtained assuming that the bootstrap sampling distributions are normal.

To get an estimate of the slope after threshold, we call

```
lincomb(fit, comb=c(0,1,1), alpha=0.05)
```

```
       est          lb          ub
 2.40795883 -0.06780353  4.88372120
```

To perform a likelihood ratio test, we call

```
library(lmtest)
fit.0=lm(V3_BioV3B~1, dat.mtct.2)
lrtest(fit, fit.0)
```

```
Likelihood ratio test

Model 1: V3_BioV3B ~NAb_score + x.mod.e
Model 2: V3_BioV3B ~1
  #Df  LogLik Df Chisq Pr(>Chisq)
1   5 -354.95
```

```
2   2 -431.50 -3 153.1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Calling `plot(fit)` makes the following figure.

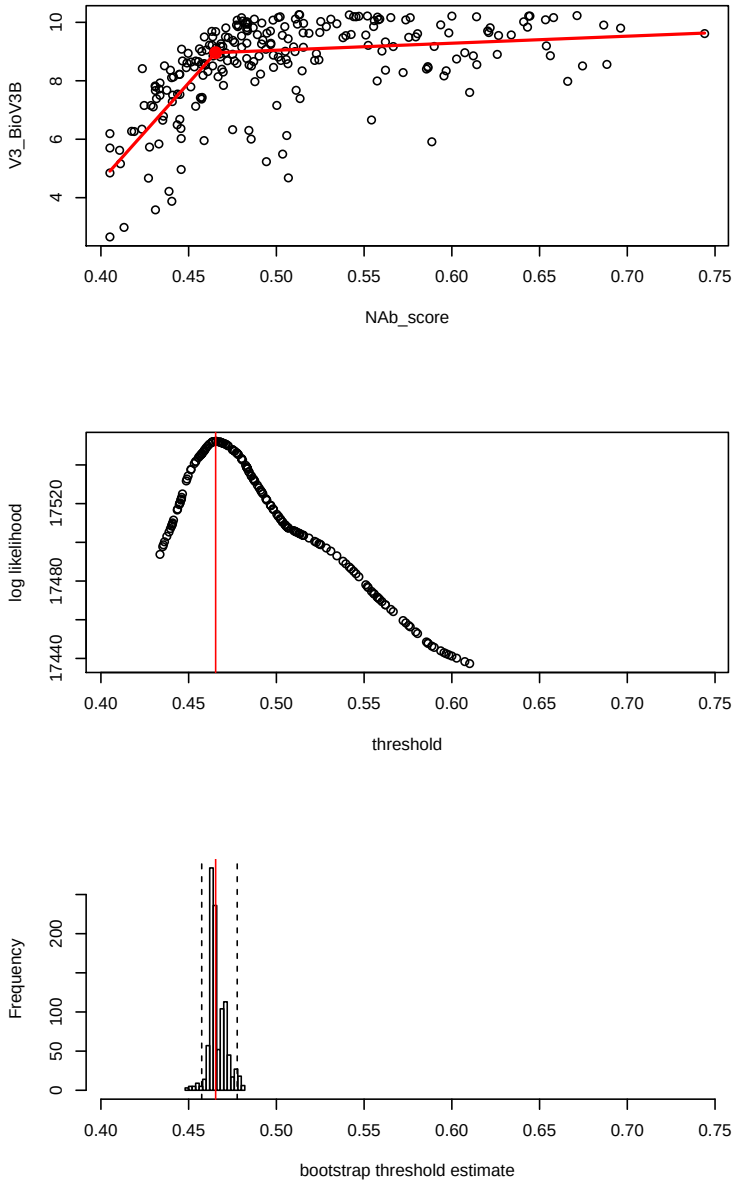

Figure 2.1: Scatterplot, profile likelihood plot, and bootstrap distribution of threshold estimates.

2.1.2 M02 model

To estimate a M02 linear regression model, we call

```
fit=chngptm(formula.1=pressure~1, formula.2=~temperature, data=pressure,
 type="M02", family="gaussian", var.type="bootstrap")
summary(fit)
```

```
Change point model threshold.type:  hingequad

Coefficients:
                                  est  p.value*       (lower     upper)
(Intercept)                8.278463507 0.4733673 -14.35129837 30.9082254
(temperature-chngpt)+      0.007124705 0.9944183  -2.00325636  1.9890069
I((temperature-chngpt)+^2) 0.039305656 0.3644561  -0.04564143  0.1242527

Threshold:
   est (lower upper)
   220   -680    240
```

2.2 Continuous two-phase linear regression with random intercepts

The following code fits the linear mixed model:

$$Y = a + \alpha^T z + \gamma x + \beta (x - e)_+ + \epsilon$$
$$a \sim N(0, \sigma_a)$$
$$\epsilon \sim N(0, \sigma_\epsilon)$$

Variance estimates are being developed.

```
dat=sim.twophase.ran.inte(threshold.type="segmented", n=50, seed=1)
fit = chngptm (formula.1=y~z+(1|id), formula.2=~x, family="gaussian", dat,
 type="segmented", est.method="grid", var.type="none")
summary(fit)
plot(fit, which=1, plot.individual.line=T, lcol="gray", lwd=.5)
```

```
No variance estimate available.

(Intercept)           z           x (x-chngpt)+      chngpt
  2.7154145   0.3514853   1.7894006   2.5695986   5.1571429
```

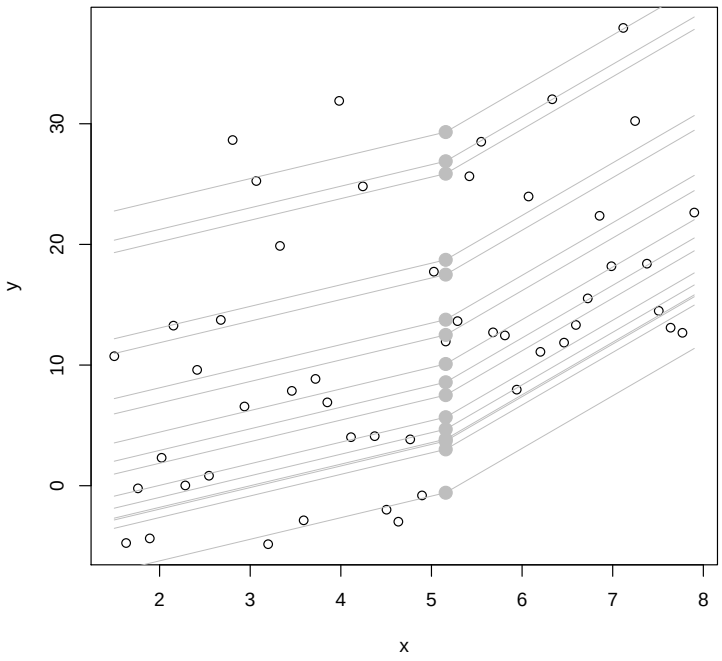

Figure 2.2: Each line correponds to one id.

2.3 Continuous three-phase linear regression

The following code fits a three-phase linear regression model. The default estimation method is *fastgrid* and the default variance type is *bootstrap*.

$$\eta = \beta_1 + \beta_2 z + \beta_3 x + \beta_4 (x - e)_+ + \beta_5 z x + \beta_6 z (x - e)_+$$

```
fit=chngptm (formula.1=pressure~1, formula.2=~temperature, pressure, type="M111",
 family="gaussian", ci.bootstrap.size=20)
summary(fit)
```

```
Change point model threshold.type:  M111

Coefficients:
                               est Std. Error*       (lower        upper)     p.value*
(Intercept)            -3310.976868 1006.428099 -5283.575941 -1338.3777950 1.002481e-03
temperature               11.417859    3.015001     5.508457    17.3272614 1.524670e-04
(temperature-chngpt1)-    -3.862734    1.612508    -7.023249    -0.7022192 1.659850e-02
(temperature-chngpt2)-    -7.425005    1.700649   -10.758278    -4.0917324 1.265528e-05

Threshold:
         est Std. Error (lower upper) p.value
chngpt.1 240   30.61224    180    300      NA
chngpt.2 320   31.12245    259    381      NA
```

2.4 Continuous two-phase logistic regression

For continuous two-phase logistic regression, a fast grid search method for estimation is not yet available. In addition, we have observed that bootstrap confidence intervals have similar coverage as robust analytical confidence intervals (Fong et al., 2017b). Thus, we recommend either `var.type="bootstrap"` or `var.type="robust"` in the call to chngptm. Note that when it is set to *robust*, an auxiliary fit needs to be supplied, which is generally a smooth parametric model with enough but not too many degrees of freedom.
To estimate a hinge logistic regression model, we call

```
library(splines)
 fit=chngptm(formula.1=y~birth, formula.2=~NAb_SF162LS, dat.mtct,
 type="hinge", family="binomial",
 est.method="smoothapprox", var.type="robust",
 aux.fit=glm(y~birth + ns(NAb_SF162LS,3), dat.mtct, family="binomial"))
summary(fit)
```

```
Change point model type:  hinge

Coefficients:
                            OR     p.value    (lower    upper)
(Intercept)           0.7026523 0.341429662 0.3388366 1.4571044
birthVaginal          1.2397649 0.523159883 0.6393632 2.4039809
(NAb_SF162LS-chngpt)+ 0.6712371 0.001332547 0.5270730 0.8548327

Threshold:
   26.3%   (lower   upper)
7.373374 5.472271 8.186464
```

The *chngptm* function supports the use of cbind in the formula, as the *glm* function does. For example,

```
dat.2=sim.chngpt("thresholded", "step", n=200, seed=1, beta=1, alpha=-1,
  x.distr="norm", e.=4, family="binomial")
dat.2$success=rbinom(nrow(dat.2), 10, 1/(1 + exp(-dat.2$eta)))
dat.2$failure=10-dat.2$success
fit.2a=chngptm(formula.1=cbind(success,failure)~z, formula.2=~x,
  family="binomial", dat.2, type="step")
```

2.5 Continuous two-phase Poisson regression

Only grid search method and bootstrap confidence intervals are supported, so getting the model fit with confidence intervals could take some time.

```
counts <- c(18,17,15,20,10,20,25,13,12)
outcome <- as.integer(gl(3,1,9))
treatment <- gl(3,3)
print(d.AD <- data.frame(treatment, outcome, counts))
fit.4=chngptm(formula.1=counts ~treatment, formula.2=~outcome, data=d.AD,
  family="poisson", type="segmented", var.type="bootstrap")
summary(fit.4)
```

2.6 Discontinuous two-phase GLM

Confidence interval for discontinuous threshold regression models can be constructed by m-out-of-n bootstrap.

```
fit=chngptm(formula.1=mpg~hp, formula.2=~drat, mtcars, type="step",
 family="gaussian", var.type="bootstrap", ci.bootstrap.size=100, m.out.of.n=20)
summary(fit)
```

```
Change point model threshold.type:  step

Coefficients:
                    est Std. Error*      (lower      upper)     p.value*
(Intercept) 27.29298302  2.89102342 21.62657712 32.95938892 3.706663e-21
hp          -0.05692654  0.01644498 -0.08915870 -0.02469439 5.369001e-04
drat>chngpt  5.24824935  2.72504835 -0.09284542 10.58934411 5.411325e-02

Threshold:
       est Std. Error     (lower     upper)    p.value
 3.9200000  0.4693878  3.0000000  4.8400000         NA
```

2.7 Two-phase Cox regression

The *chngpt* package also provides some support for estimation of threshold Cox regression models. What is missing, though, is confidence intervals for parameter estimates and hypothesis testing methods. See the help page on *chngpt* for an example.

2.8 Models with interaction terms

In the following example we fit a model with an interaction term.

$$\eta = \beta_1 + \beta_2 z + \beta_3 x + \beta_4 (x - e)_+ + \beta_5 z x + \beta_6 z (x - e)_+$$

```
fit=chngptm(formula.1=mpg ~hp, formula.2=~hp*drat, mtcars, type="segmented",
 family="gaussian", var.type="bootstrap", ci.bootstrap.size=100)
summary(fit)
```

```
Change point model threshold.type:  segmented

Coefficients:
                         est Std. Error*       (lower      upper)  p.value*
(Intercept)        71.0423961 107.7931740 -140.2322250 282.3170173 0.5098559
hp                 -0.5714405   0.7521618   -2.0456777   0.9027967 0.4474155
drat              -14.3708279  35.7034558  -84.3496013  55.6079456 0.6873122
(drat-chngpt)+     21.6073593  73.6732299 -122.7921714 166.0068899 0.7693032
hp:drat             0.1658607   0.2482010   -0.3206132   0.6523346 0.5039730
hp:(drat-chngpt)+  -0.1970979   0.5108437   -1.1983515   0.8041557 0.6996239

Threshold:
       est Std. Error     (lower     upper)    p.value
 3.2300000  0.4489796  2.3500000  4.1100000         NA
```

In the following example we fit a model with two interaction terms

$$\eta = \beta_1 + \beta_2 z_1 + \beta_3 z_2 + \beta_4 I(x > e) + \beta_5 z_1 I(x > e) + \beta_6 z_2 I(x > e)$$

```
fit=chngptm(formula.1=mpg~hp+wt, formula.2=~hp*drat+wt*drat, mtcars, type="step",
 family="gaussian", var.type="bootstrap", ci.bootstrap.size=100)
summary(fit)
```

```
Change point model threshold.type:  step

Coefficients:
                          est Std. Error*       (lower      upper)     p.value*
(Intercept)       30.83332346  4.06186261  22.87207274 38.79457417 3.176122e-14
hp                -0.02389962  0.02760935  -0.07801395  0.03021471 3.866903e-01
wt                -2.58756410  1.17757075  -4.89560276 -0.27952543 2.799370e-02
drat>chngpt       11.69827186 28.02745000 -43.23553015 66.63207386 6.763959e-01
hp:I(drat>chngpt) -0.00894615  0.20736123  -0.41537415  0.39748185 9.655877e-01
wt:I(drat>chngpt) -3.22148003 21.48073350 -45.32371769 38.88075762 8.807878e-01

Threshold:
       est Std. Error     (lower     upper)    p.value
 3.7000000  0.2806122  3.1500000  4.2500000         NA
```

3 Testing examples

Testing methods are described in Fong et al. (2015) and Fong et al. (2017a).

An example in linear regression:

```
test=chngpt.test(formula.null=Volume~1, formula.chngpt=~Girth, trees,
 type="segmented", family="gaussian")
test
```

```
        Maximum of Likelihood Ratio Statistics

data:  trees
Maximal statistic = 17.694, change point = 15.388, p-value = 0.00014
alternative hypothesis: two-sided
```

The first line gives the type of test carried out, and it is maximal likelihood ratio test here, which is the default. In addition, a plot function can be called on the test object to show the score or likelihood ratio statistic as a function of candidate change points.

An example in logistic regression:

```
test=chngpt.test(formula.null=y~birth, formula.chngpt=~NAb_SF162LS, dat.mtct,
 type="hinge", family="binomial", main.method="score")
test
```

```
        Maximum of Score Statistics

data:  dat.mtct
Maximal statistic = 3.3209, change point = 7.0347, p-value = 0.00284
alternative hypothesis: two-sided
```

The first line gives the type of test carried out, and it may be maximal likelihood ratio test. In addition, a plot function can be called on the test object to show the score or likelihood ratio statistic as a function of candidate change points.

4 Further considerations

4.1 Model choice

The choice of threshold effects is typically through a combination of domain knowledge and modeling. One modeling approach is to first examine the relationship using local polynomial regression. To choose among the segmented, hinge, and upper hinge models formally, we can use Wald tests. For example, if the question is framed as choosing between segmented and hinge models, we can fit a segmented model and then look at the slope before threshold in the summary function output. If the estimate is not significantly different from 0, then it is justifiable to fit a hinge model. We can also look at the slope after threshold, which is not displayed as part of the summary function output, but can be obtained by calling lincomb (see example in Section 2.1.1). If this estimate is not significantly different from 0, then it is justifiable to fit an upper hinge model. If the hinge or upper hinge model is reasonable, it is preferred over the segmented model because the model can be estimated with substantially higher precision (Fong et al., 2017b; Elder and Fong, 2019).

4.2 Estimation and inference methods

There are three types of search methods for finding the MLE (maximum likelihood estimator). Users generally do not need to worry about setting the argument, which is *est.method*, since the function chooses the most appropriate one by default. In the order of development, the three search methods are grid, smooth approximation, and fastgrid. The grid method is the most general and the slowest; it is recommended when other methods are not available. The smooth approximation method (Fong et al., 2017a) involves approximating the likelihood function with a differentiable function to allow gradient-based search; it is available for linear and logistic regression and mostly recommended for logistic regression only. Fastgrid (Fong, 2019; Elder and Fong, 2019) is a new method that is super fast and gives exact solutions; it is only available for certain threshold linear regression models.

Robust confidence interval methods are described in Fong et al. (2017b).

Hypothesis testing methods are described in Fong et al. (2015) and Fong et al. (2017a).

References

Elder, A. and Fong, Y. (2019), "Estimation and Inference for Upper Hinge Regression Models," *Environmental and Ecological Statistics*, 26, 287–302.

Fong, Y. (2019), "Fast Bootstrap Confidence Intervals for Continuous Threshold Linear Regression," *Journal of Computational and Graphical Statistics*, 28, 466–470.

Fong, Y., Di, C. and Permar, S. (2015), "Change point testing in logistic regression models with interaction term," *Statistics in medicine*, 34, 1483–1494.

Fong, Y., Huang, Y., Gilbert, P. and Permar, S. (2017a), "chngpt: threshold regression model estimation and inference," *BMC Bioinformatics*, 18, 454–460.

Fong, Y., Chong, D., Huang, Y. and Gilbert, P. (2017b), "Model-robust Inference for Continuous Threshold Regression Models," *Biometrics*, 73, 452–462.