Tutorial for the R package chngpt

Youyi Fong

August 8, 2018

1 Introduction

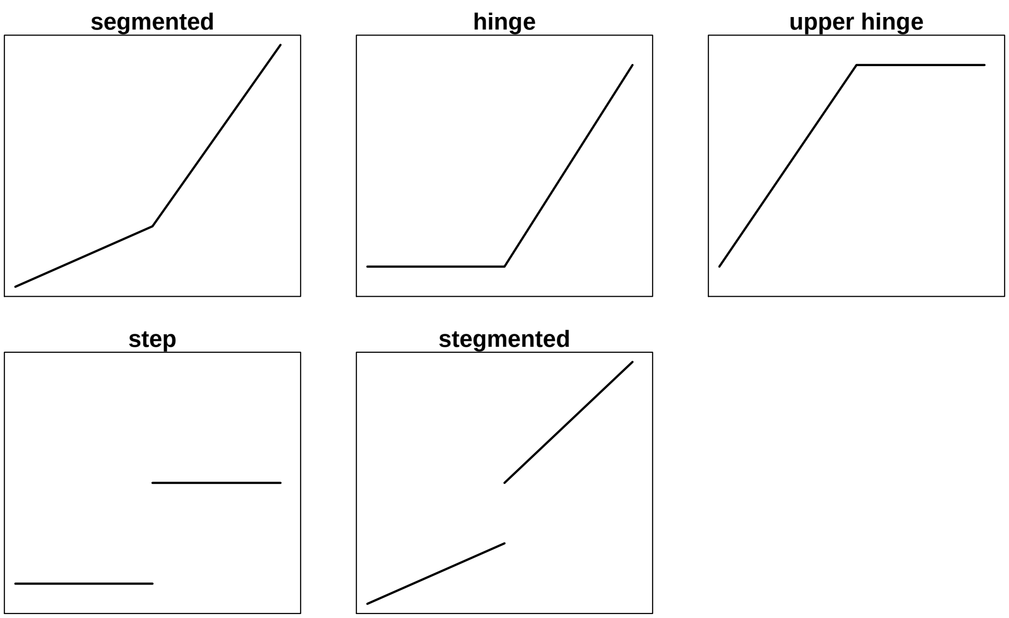

Figure 1.1: Types of threshold effects. Hinge, upper hinge and segmented are called continuous threshold models, while step and stegmented are called discontinuous threshold models.

Let e be the threshold parameter, x be the predictor with threshold effect, and z be additional predictors. Let $I(x > e) = 1$ if $x > e$ and 0 otherwise; $(x - e)_+ = x - e$ if $x \geq e$ and 0 otherwise; and $(x - e)_- = x - e$ if $x < e$ and 0 otherwise. The threshold effects shown in Figure 1.1 can be written as:

$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x - e)_+ + \gamma x \qquad \text{(segmented)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x - e)_+ \qquad \text{(hinge)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x - e)_- \qquad \text{(upper hinge)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 I (x > e) \qquad \text{(step)}$$
$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x - e)_+ + \gamma x + \beta_2 I (x > e) . \qquad \text{(stegmented)}$$

2 Examples

The examples below are organized by type of threshold effects and regression models. Before we get into specific examples, here are some notes that are of general interest:

- The fitted model has a component named best.fit, which is the glm or coxph fit at the estimated threshold parameter. This could be useful to know if one would like to extract information from model.

- If the models contain interaction between a thresholded covariate and a non-thresholded covariate, be careful. The package provides some support, but it is incomplete.

2.1 Segmented and hinge linear regression

For continuous threshold linear regression, we have developed a grid search method for estimation that is super fast (Fong, 2018). Together with the observation that bootstrap confidence intervals have better coverage than robust analytical confidence intervals (Fong et al., 2017b) for continuous threshold linear models, we recommend setting `est.method="fastgrid"` and `var.type="bootstrap"` in the call to chngptm.

To estimate a threshold linear regression model with a segmented-type change point for the relationship between *V3_BioV3B* and *NAb_score* in the *MTCT* dataset, we call

```
fit=chngptm (formula.1=V3_BioV3B~1, formula.2=~NAb_score, dat.mtct.2,
  type="segmented", family="gaussian",
  est.method="fastgrid", var.type="bootstrap", save.boot=TRUE)
```

- `formula.2` and `formula.1`: threshold variable and the rest of the model
- `type`: type of threshold model to fit
- `est.method` defaults to *fastgrid* and is recommended
- `var.type`: *bootstrap* method is recommended here
- `save.boot`: saves bootstrap samples for plotting bootstrap distributions

Calling `summary(fit)`, we get

```
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
```

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

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

and get

```
                    95%         95%
2.40795883 -0.06780353  4.88372120
```

Calling `plot(fit, which=1)` and `plot(fit, which=3)`, we get the two plots on the left-hand side of Figure 2.1. Changing `est.method` to `smoothapprox` in the model fit led us to the two plots on the right-hand side.

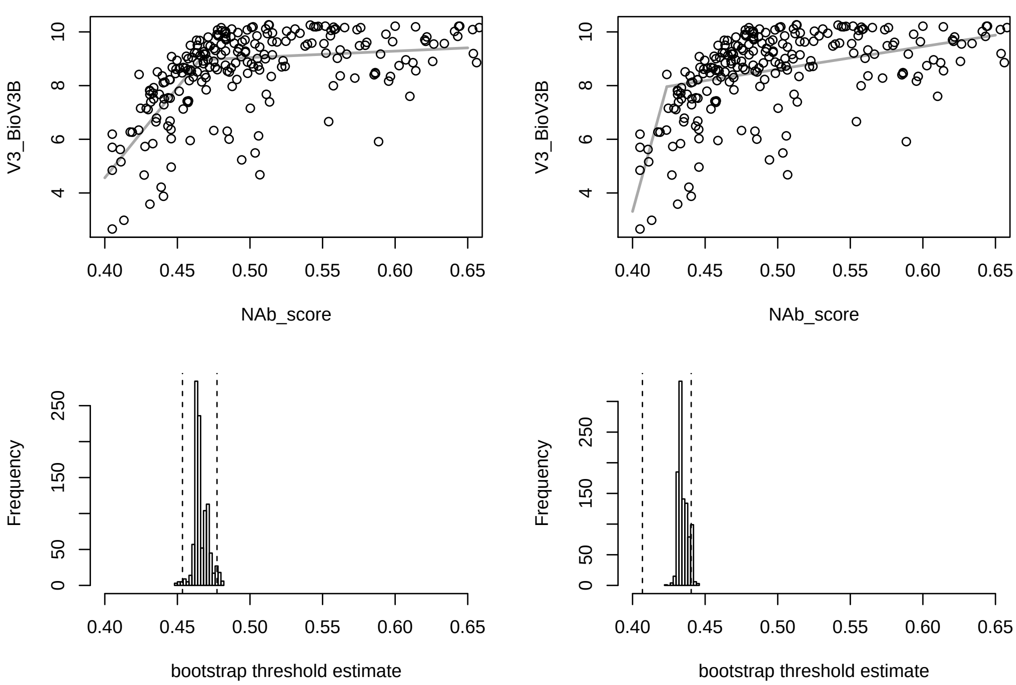

Figure 2.1: This is a replicate of Fong (2018) Figure 1. Left: results by fast grid search; right: results by smooth approximation search. Top: scatterplots with fitted models (gray lines); bottom: bootstrap distributions of the threshold estimate from 10^3 replicates. The dashed lines correspond to the 95% symmetric bootstrap confidence interval.

A second example To estimate a threshold linear regression model with a segmented-type change point in *Girth* for the *trees* dataset, we call

```
fit=chngptm(formula.1=Volume~1, formula.2=~Girth, data=trees,
type="segmented", family="gaussian",
 var.type="bootstrap", weights=NULL)
```

- `formula.2` and `formula.1`: threshold variable and the rest of the model
- `type`: type of threshold model to fit
- `var.type`: *bootstrap* method is recommended for confidence interval
- `weights` can be supplied

Calling `summary(fit)`, we get

```
Change point model type:  segmented

Coefficients:
                       est     p.value*     (lower     upper)
(Intercept)     -24.614440 1.985482e-04 -37.580354 -11.648527
Girth             3.993966 9.288973e-11   2.785558   5.202373
(Girth-chngpt)+   4.266618 8.261144e-04   1.765770   6.767467

Threshold:
   est (lower upper)
  16.0   12.9   19.1
```

Calling `plot(fit)`, we get Figure 2.2.

To test whether there is a change point (Fong et al., 2015), we call

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

When printed, we get

```
        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.

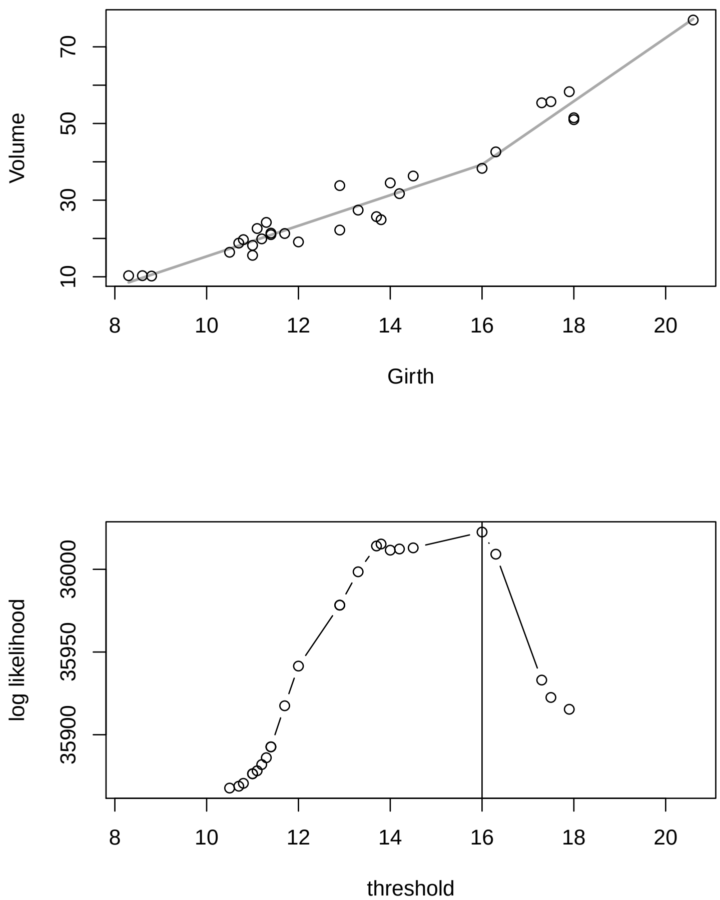

Figure 2.2: (a) Scatterplot of timber volume vs girth. The gray line shows the fitted segmented model. (b) Log likelihood of the submodel versus threshold parameter.

2.2 Segmented and hinge logistic regression

For continuous threshold 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 logistic regression model with a hinge-type change point in *NAb_SF162L* for the MTCT dataset, 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"))
```

- `formula.2` and `formula.1`: threshold variable and the rest of the model
- `type`: type of threshold model to fit
- `est.method`: *smoothapprox* is recommended
- `var.type`: *robust* is recommended for confidence interval
- `aux.fit`: required for *robust* variance estimation

Calling `summary(fit)`, we get

```
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
```

To test whether there is a change point (Fong et al., 2015), we call

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

When printed, we get

```
        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.

cbind 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")
```

2.3 Continuous threshold Poisson models

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)
```

2.4 Discontinuous threshold regression models

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

2.5 Threshold Cox models

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

3 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.

The *hinge* model is a special case of the *segmented* model with the slope before threshold fixed at 0. If the 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).

4 Implementation details

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, 2018) is a new method that is super fast and gives exact solutions; it is only available for continuous threshold linear regression now.

5 Statistical inference details

Robust confidence interval methods are described in Fong et al. (2017b). For linear regression, we recommend symmetric bootstrap confidence interval, as described in Fong (2018).

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

Acknowledgement

We are grateful for questions and comments from researchers around the world interested in using *chngpt*, which have led to great improvement to the package.

References

Fong, Y. (2018), "Fast Bootstrap Confidence Intervals for Continuous Threshold Linear Regression," *Journal of Computational and Graphical Statistics*, in press.

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.