`lm.br`: An R Package for Broken Line Regression

Marc Adams

Abstract

The R package `lm.br` delivers exact tests and exact confidence regions for a changepoint in linear or multiple linear regression. This package implements the likelihood theory of conditional inference. Examples demonstrate its use and show some properties of the broken-line models.

1 Theory

A broken-line model consists of two straight lines joined at a changepoint. Three variants are

$$y_i = \alpha + \beta(x_i - \theta)_- + \beta'(x_i - \theta)_+ + e_i \tag{1}$$

$$y_i = \alpha + \beta(x_i - \theta)_- + e_i \tag{2}$$

$$y_i = \beta(x_i - \theta)_- + e_i \tag{3}$$

denoting $a_- = min(a, 0)$ and $a_+ = max(a, 0)$, where $e \sim N(0, \sigma^2\Sigma)$. Parameters θ, α, β, β', σ are unknown but Σ is known. Model (2) is a threshold model, while model (3) would apply for a known threshold level. Inference about a parameter uses the assumed model and resulting distribution of a test statistic.

A test statistic D assigns a numeric value to a postulate parameter value, p_0, depending on the model and the observations. $D(p_0)$ is itself a random variable because it is a function of the random observations. A significance level is the probability that D could be worse than the observed value, $SL(p_0) = Pr[D(p_0) > D(p_0)_{obs}]$, based on the model. The set of postulate values such that $SL > \alpha$ is a $100(1-\alpha)\%$ confidence region for the true parameter value.

Conditional inference incorporates sufficient statistics that account for the other, unknown parameters. This refinement determines the exact distribution of a test statistic, even for small data sets. Student's t, for example, is the distribution of a sample mean conditional on a sufficient statistic for the variance. See Kalbfleisch (1985, ch. 15), Cox and Hinkley (1974, ex. 5.1, 5.5).

The likelihood-ratio is an optimal test statistic. Knowles and Siegmund (1989) examined an exact significance test, using likelihood-ratio, for the null hypothesis of a single line versus a broken line. Knowles, Siegmund, and Zhang (1991) derived the conditional likelihood-ratio (CLR) significance tests for the nonlinear parameter in semilinear regression. Siegmund and Zhang (1994) applied these tests to get exact confidence intervals for the changepoint θ in models (1) and (2), and exact confidence regions for the two-parameter changepoint (θ, α) in model (2). Knowles et al. (1991) also developed a formula to evaluate these tests rapidly, which `lm.br` implements.

`lm.br` extends this theory. Their method derives the conditional likelihood-ratio test for (θ, α) in model (1). The theory adapts to the case σ known, which is useful for the Normal approximation of a binary random variable (Cox and Snell, 1989, eq. 2.28). And these tests simplify for a postulate changepoint value outside the range of x-values.

For comparisons, Approximate-F (AF) is another inference method that is common in nonlinear regression. The AF method estimates the distribution of a likelihood-ratio statistic by its asymptotic χ^2 distribution with partial conditioning on a sufficient statistic for the variance. See Draper and Smith (1998, sec. 24.6). Simulations and examples cover all of the above theory.

2 Simulation Tests

Coverage frequencies of the 95% confidence interval on 100 arbitrary models

		CLR	**AF**
10 observations,	$x_1 - 1 < \theta < x_{10} + 1$	**95.0 – 95.2**	**90.0 – 97.5**
30 observations,	$x_{10} < \theta < x_{20}$	**95.0 – 95.2**	**90.8 – 95.0**
100 observations,	$x_{10} < \theta < x_{20}$	**95.0 – 95.2**	**91.3 – 95.0**

To give one specific example, coverage frequency is 95.2% by CLR but 90.7% by AF for a first-line slope -1, second-line slope +0.5, changepoint $\theta = 3$, and 10 observations at x = (1.0, 1.1, 1.3, 1.7, 2.4, 3.9, 5.7, 7.6, 8.4, 8.6) with $\sigma = 1$.

A program created the arbitrary models using $U \sim Uniform(0, 1)$ with

$$n = 10 \quad x_1 = 1 \quad x_i = x_{i-1} + 2U \text{ for } i > 1 \quad \theta = x_1 - 1 + (x_n - x_1 + 2)U$$
$$\alpha = 0 \quad \beta = -1 \quad \beta' = 2 - 2.5U \quad \sigma = 0.1 + 2U$$

or n= 30 or n= 100 and $\theta = x_{10} + (x_{20} - x_{10})U$. For each model, the program

generated one million sets of random $y_i = \alpha + \beta(x_i - \theta)_- + \beta'(x_i - \theta)_+ + \sigma N(0,1)$ and counted how often $SL(\theta) > .05$. These coverage frequencies should be accurate to $\pm 0.05\%$.

3 Examples

3.1 Broken Line Regression

A broken-line model could fit drinking-and-driving survey results. Yearly surveys, taken in different months over the years, were adjusted by a seasonal index based on monthly surveys for a similar question (TIRF, 1998–2007; CAMH, 2003). The annual surveys asked respondents if in the past 30 days they had driven within two hours after one drink, while the monthly surveys asked if in the past 30 days they had driven within one hour after two drinks. Figure 1 shows the survey results without and with seasonal adjustment, and the exact 90% confidence region for a changepoint if the adjustment were valid.

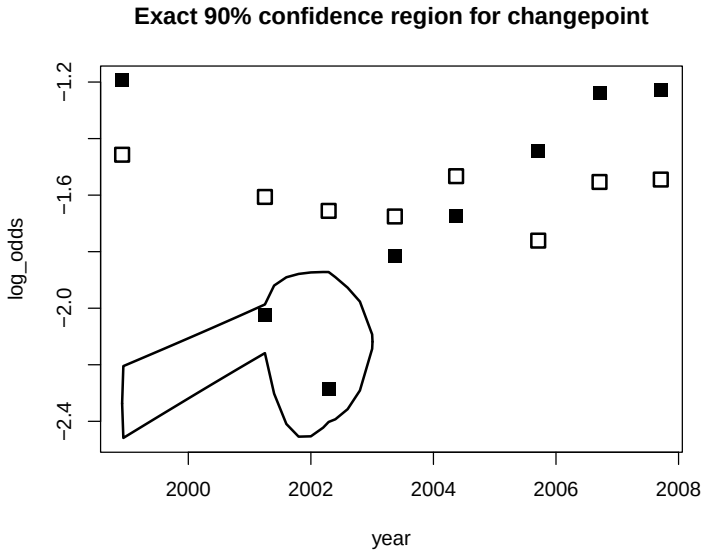

Figure 1: Drinking-and-driving surveys log-odds (blank squares) and log-odds with seasonal adjustment (solid squares) versus year, and the exact 90% confidence region for a changepoint (θ, α) by CLR.

Data input, and commands to get confidence intervals for the changepoint, are

```
> log_odds <-  c( -1.194, -2.023, -2.285, -1.815, -1.673,
+   -1.444, -1.237, -1.228 )
```

```
> year <-  c( 1998.92, 2001.25, 2002.29, 2003.37, 2004.37,
+   2005.71, 2006.71, 2007.71 )
> VarCov <-  matrix(  c(0.0361,  0,  0,  0,  0,  0,  0,  0,
+   0,  0.0218,  0.0129,  0,  0,  0,  0,  0,
+   0,  0.0129,  0.0319,  0,  0,  0,  0,  0,
+   0,  0,  0,  0.0451,  0.0389,  0,  0,  0,
+   0,  0,  0,  0.0389,  0.0445,  0,  0,  0,
+   0,  0,  0,  0,  0,  0.0672,  0.0607,  0.0607,
+   0,  0,  0,  0,  0,  0.0607,  0.0664,  0.0607,
+   0,  0,  0,  0,  0,  0.0607,  0.0607,  0.0662), nrow=8, ncol=8)
> dd <- lm.br( log_odds ~ year, w= VarCov, inv= T, var.known= T )
> dd$ci( )

95-percent confidence interval for changepoint 'theta' by CLR
    [ 2001.29, 2002.88 ]

> dd$ci( method = "AF" )

95-percent confidence interval for changepoint 'theta' by AF
    [ 1998.92, 2002.82 ]
```

The wide difference between the CLR and AF confidence intervals above is due to plateaus in the significance level on end-intervals. Both the CLR and AF methods give a constant significance level for all postulate values θ_0 on $[x_1, x_2]$, on $[x_{n-1}, x_n]$, or outside of $[x_1, x_n]$, in a model (1) with $x_1 < x_2 < ... < x_n$. (Coverage probability on these intervals is still exactly 95% by CLR, as the simulation tests show.) The inference assumes that any line slope is possible, extending to an instantaneous drop near December 1998 in this example.

3.2 Multiple Regression

`lm.br` can test for a changepoint in multiple linear regression. `lm.br` tests for a change in one coefficient of the regression model, assuming continuity. It does not test for an arbitrary structural change that might include changes of two or more coefficients or discontinuity.

Liu, Wu, and Zidek (1997) suggested a changepoint for the coefficient of car weight in a linear fit of miles-per-gallon against weight and horsepower, for 38 cars of 1978-79 models. One of R's included datasets is the ratings for 32 cars, 1973-74 models. Analysis of this 1973-74 dataset by the exact conditional likelihood-ratio inference also shows evidence for a changepoint:

```
> lm.br( mpg ~ wt + hp, data = mtcars )

Call:
lm.br(formula = mpg ~ wt + hp, type = "LL", data = mtcars)
```

```
Changepoint and coefficients:
       theta         alpha   wt < theta   wt > theta           hp
     2.62000     25.02750     -8.81519     -2.51738     -0.03003

Significance Level of H0:"no changepoint" vs H1:"one changepoint"
    SL= 0.0110841  for theta0 = 1.32  by method CLR

95-percent confidence interval for changepoint 'theta' by CLR
    [ 2.13813, 5.14625 ]
```

For multiple regression, `lm.br` applies an orthogonal transformation to a canonical model (Siegmund and Zhang, 1994). One way to see how this method works is formulaic. The composite likelihood-ratio statistic uses optimal values for unknown parameters. A canonical model lets these optimal coefficients of linear terms reduce their correspondent errors to zero always. Thus they have no effect on inference, so the algebra can omit them. This elimination reduces a multiple-predictor model to a single-predictor model. See Hoffman and Kunze (1971, ch. 6), Lehmann and Romano (2005, sec. 7.1).

4 Summary

If a broken line with Normal errors represents the relationship between a factor and responses, then `lm.br` solves the inference step for the changepoint. This package uses the technique of conditional inference to allow for the other, unknown terms in the model. Fitting a broken line can reveal the plausible interval for a changepoint, although practical cause-effect relations usually have a smooth transition. Any statistical analysis should examine the fit of the model and the error distribution with graphs and significance tests, interpret results, and consider adjustments to the model or alternate models.

References

CAMH. Monthly variation in self-reports of drinking and driving in Ontario. CAMH Population Studies eBulletin no.21, Centre for Addiction and Mental Health, Toronto, 2003. URL `www.camh.net/pdf/eb021_ddmonthly.pdf`.

D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.

D. R. Cox and E. J. Snell. *Analysis of Binary Data*. Chapman and Hall, London, 2nd edition, 1989.

N. R. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, 3rd edition, 1998.

K. Hoffman and R. Kunze. *Linear Algebra*. Prentice Hall, Englewood Cliffs, NJ, 2nd edition, 1971.

J. G. Kalbfleisch. *Probability and Statistical Inference*, volume 2. Springer, New York, 2nd edition, 1985.

M. Knowles and D. Siegmund. On Hotelling′s approach to testing for a nonlinear parameter in regression. *International Statistical Review*, 57(3):205–220, 1989.

M. Knowles, D. Siegmund, and H. P. Zhang. Confidence regions in semilinear regression. *Biometrika*, 78(1):15–31, 1991.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 3rd edition, 2005.

J. Liu, S. Wu, and J. V. Zidek. On segmented multivariate regression. *Statistica Sinica*, 7:497–525, 1997.

D. Siegmund and H. P. Zhang. Confidence regions in broken line regression. In E. Carlstein, H. Muller, and D. Siegmund, editors, *Change-point Problems*, volume 23 of *IMS Lecture Notes*, pages 292–316. Institute of Mathematical Statistics, Hayward, CA, 1994.

TIRF. Road safety monitor: Drinking and driving. Technical report, Traffic Injury Research Foundation, Ottawa, Ontario, 1998–2007. URL `www.tirf.ca`.