R package ‘lm.br’

Marc Adams September 16, 2013

Introduction

‘lm.br’ delivers exact tests and exact confidence regions for a changepoint in linear or multivariate linear regression. This package implements the likelihood theory of conditional inference. Examples demonstrate its use and note properties of the broken line models.

A broken-line model consists of two straight lines joined at a changepoint. Formally, the broken-line models are

$$y_i \;=\; \alpha \;+\; \beta(x_i-\theta)_- \;+\; \beta'(x_i-\theta)_+ \;+\; e_i \qquad \text{(LL)}$$

$$y_i \;=\; \alpha \;+\; \beta(x_i-\theta)_- \;+\; e_i \qquad \text{(LT)}$$

$$y_i \;=\; \beta(x_i-\theta)_- \;+\; e_i \qquad \text{(LT0)}$$

$\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{\Sigma})$, where θ, α, β, β', σ are unknown but $\mathbf{\Sigma}$ is known. Notation $a_- = \min(a,0)$ and $a_+ = \max(a,0)$. Model LT and its horizontal reflection TL are threshold models. Model LT0 would apply for a known threshold level.

The likelihood-ratio is a test statistic. A test statistic ‘*D*’ assigns a numeric value to a postulate parameter value, p_0 , based 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 $SL > 1 - \alpha$ is a $100\alpha\%$ confidence region.

Conditional inference incorporates sufficient statistics to account for the other, unknown parameters. This refinement determines the exact distribution of a test statistic, even for small samples. The familiar 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).

Knowles, Siegmund and Zhang (1991) derived the conditional likelihood-ratio (CLR) significance tests for the non-linear parameter in semilinear regression. Siegmund and Zhang (1994) applied these tests to get exact confidence intervals for the changepoint θ in models LL and LT, and exact confidence regions for the two-parameter changepoint (θ, α) in model LT. Knowles et al. (1991) also developed a formula for rapid evaluation, which ‘lm.br’ implements.

‘lm.br’ augments this theory. Their method derives an exact significance test for (θ,α) in model LL. The theory adapts for the case σ known. And these exact significance tests simplify for a postulate changepoint value outside $[\, x_{min}, x_{max} \,]$ (Knowles & Siegmund 1989).

Approximate-F (AF) is another inference method that is common for nonlinear regression, but it is not exact. 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, chap. 24).

Examples

1. Simulation Tests

Table – Coverage frequencies of the 95% confidence interval for 100 random 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 $\sigma = 1$.

The formulae that generated the random models are

$$n = 10 \qquad x_1 = 1, \quad x_i = x_{i-1} + 2 \cdot U \quad \text{for } i > 1 \qquad \theta = x_1 - 1 + (x_n - x_1 + 2) \cdot U$$

$$\alpha = 0 \qquad \beta = -1 \qquad \beta' = 2 - 2.5 \cdot U \qquad \sigma = 0.1 + 2 \cdot U \qquad \mathbf{\Sigma} = \mathbf{I},$$

or $n = 30$ or $n = 100$ and $\theta = x_{10} + (x_{20} - x_{10}) \cdot U$, where $U \sim \text{Uniform}(0,1)$. For each model, the program output one million sets of random $y_i = \alpha + \beta\,(x_i - \theta)_{-} + \beta'(x_i - \theta)_{+} + \sigma \cdot \text{N}(0,1)$ and counted how often $\text{SL}(\theta) > .05$. Coverage frequencies should be accurate to ±0.05% .

2. Drinking and Driving

Drinking and driving might have followed a broken line trend. Yearly surveys were adjusted by a seasonal index based on monthly surveys for a similar question. The annual surveys asked respondents if in the past 30 days they had driven within two hours after a drink, while the monthly surveys asked if in the past 30 days they had driven within one hour after two drinks. The figure shows the survey results without and with seasonal adjustment, and the exact 95% confidence region for the changepoint if the adjustment were valid.

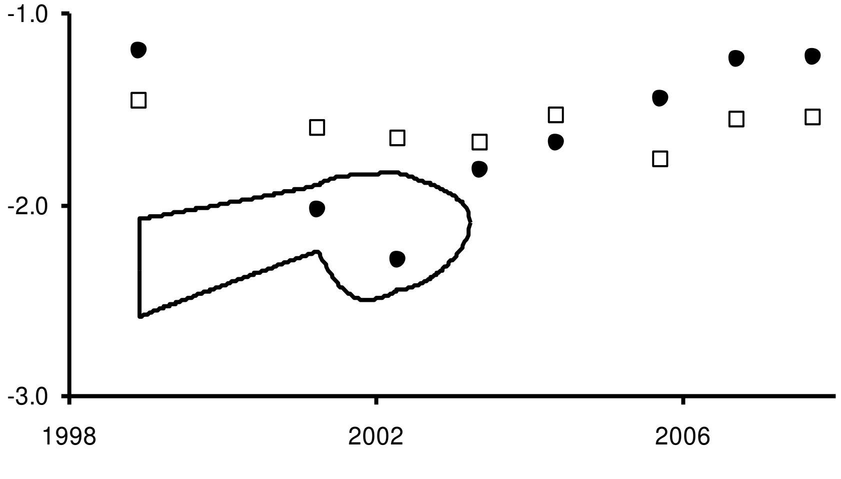

Figure – Drinking-and-driving surveys □ log-odds and ● log-odds with seasonal adjustment versus year, and the exact 95% confidence region for the changepoint (θ, α) .

In *R* the commands are

```
> library( lm.br )

> 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$cr( )

> dd$ci( )
    95% confidence interval for the changepoint 'theta'
                         [ 2001.29, 2002.88 ]      by method  CLR
> dd$ci( .95, 'af' )
                         [ 1998.92, 2002.82 ]      by method  AF
```

The wide difference in confidence intervals here is due to plateaus in the significance levels on end-intervals. Both the CLR and the AF methods give a constant significance level for all θ_0 on $(x_1, x_2]$, on $[x_{n-1}, x_n)$, and outside (x_1, x_n), in model LL. The inference assumes that any line slope is possible, extending to an instantaneous drop near Dec 1998 in this example.

3. Multivariate Regression

'lm.br' can test for a changepoint in multivariate regression. 'lm.br' tests for a coefficient change in the first term of the regression model, assuming continuity. It does not test for an arbitrary structural change that might involve multiple parameters or discontinuity.

Liu et al. (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, 1978–79 models. One of *R*'s included datasets is the ratings for 32 cars, 1973–74 models. Analysis of this dataset by conditional likelihood-ratio using 'lm.br' also shows some evidence for a changepoint:

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

       Broken-line type   LL

       Significance Level of H0:"no changepoint" vs H1:"one changepoint"
          SL= 0.0110841  for theta0 = 0.93  by method CLR

       95% confidence interval for the changepoint 'theta' by CLR
           [ 2.13813, 5.14625 ]

       Changepoint and coefficients
           theta     1-vector   wt < theta   wt > theta          hp
         2.62000     25.02750     -8.81519     -2.51738    -0.03003
```

'lm.br' applies canonical reduction for changepoint inferences in multivariate regression (Siegmund and Zhang 1994). One way to see how this theory works is formulaic. The composite likelihood-ratio statistic uses optimal values for unknown parameters. A canonical model lets these optimal coefficients for other terms reduce their correspondent errors to zero always. Thus they have no effect on inference, so the algebra can omit them. This elimination reduces the multivariate model to a univariate model. See Hoffman and Kunze (1971, ch.6) and Lehmann (2005, sec. 7.1).

Conclusion

If a broken line with Normal errors represents the relationship between a factor and responses, 'lm.br' solves the inference step for the changepoint. Fitting a broken line can reveal the plausible region for a changepoint, although practical cause-effect relations have a smooth transition usually. 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 alternative models.

References

Centre for Addiction and Mental Health (2003), "Monthly Variation in Self-Reports of Drinking and Driving in Ontario," *CAMH Population Studies eBulletin* [online], no. 21, Toronto: Author. Available online at http://www.camh.net/pdf/eb021_ddmonthly.pdf .

Draper, N.R. and Smith, H. (1998), *Applied Regression Analysis* (3rd ed.), New York: Wiley.

Hoffman, K. and Kunze, R. (1971), *Linear Algebra* (2nd ed.), Englewood Cliffs, NJ: Prentice Hall.

Kalbfleisch, J.G. (1985), *Probability and Statistical Inference* (Vol.2; 2nd ed.), New York: Springer.

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

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

Lehmann, E.L. and Romano, J.P. (2005), *Testing Statistical Hypotheses* (3rd ed.), New York: Springer.

Siegmund, D. and Zhang, H.P. (1994), "Confidence regions in broken line regression," in *Change-point Problems*, IMS Lecture Notes – Monograph Series, vol. 23, eds E. Carlstein, H. Muller and D. Siegmund, Hayward, CA: Institute of Mathematical Statistics, 292-316.

Traffic Injury Research Foundation (1998-2007), *Road Safety Monitor: Drinking and Driving*, Ottawa: Author. Available online at http://www.tirf.ca .

Liu, J., et al. (1997), "On Segmented Multivariate Regression," *Statistica Sinica*, 7, 497-525.