

blmr: Broken Line Model Regression

Marc Adams May 20, 2013

Introduction

This draft introduces the theory and use of the R package 'blmr' for exact confidence regions in broken line regression. The examples demonstrate the value of exact inference.

Theory

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

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

$$y_i = \alpha + \beta(x_i - \theta)_+ + e_i \quad (2)$$

$$y_i = \beta(x_i - \theta)_+ + e_i \quad (3)$$

with $x_1 \leq x_2 \leq \dots \leq x_n$ and $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma})$, where $\theta, \alpha, \beta', \beta, \sigma$ are unknown but $\mathbf{\Sigma}$ is known. Notation $a_- = \min(a, 0)$ and $a_+ = \max(a, 0)$. Model (2) and its horizontal reflection (-2) are threshold models. Model (3) applies for a known threshold level, or for multivariate regression as shown in Example 3.

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

Conditional inference incorporates the uncertainty of unknown parameters to determine the exact probability distribution of a test statistic. Student's t , for example, is the distribution of a sample mean conditional on a sufficient statistic for the unknown 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 regions for the changepoint θ in models (1) and (2), and for the two-parameter changepoint (θ, α) in model (2). Knowles et al. (1991) also developed a formula for rapid evaluation, which 'blmr' implements.

'blmr' augments this theory: Their method derives an exact significance test for (θ, α) in model (1). The theory extends naturally to the case σ known. And these exact significance tests degenerate to simpler forms for a postulate changepoint value outside of $[x_1, x_n]$ (Knowles and Siegmund 1989).

Approximate-F (AF) is another inference method that is common in broken line 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. Simulations

Coverage frequencies of .95-confidence intervals on 100 random models, 1 m. iterations

		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, the coverage frequency of the 0.95-confidence interval is 95.2% by CLR but 90.7% by AF for a first-line slope -1, second-line slope +0.5, changepoint at $x = 3$, 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

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

$$\alpha = 0 \quad \beta' = -1 \quad \beta = 2 - 2.5 \cdot U \quad \sigma = 0.1 + 2 \cdot U \quad \Sigma = I,$$

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 $y_i = \alpha + \beta'(x_i - \theta)_- + \beta(x_i - \theta)_+ + \sigma \cdot N(0,1)$ and counted how often $SL(\theta) > .05$. Coverage frequencies should be accurate to $\pm 0.05\%$.

2. Drinking and driving surveys

Yearly drinking-and-driving surveys fit a broken-line trend, when they are adjusted by a seasonal index based on monthly surveys for a similar question. The annual surveys by TIRF (1998-2007) asked respondents if in the past 30 days they had driven within two hours after a drink, while the monthly surveys by CAMH (1999-2002) asked if in the past 30 days they had driven within one hour after two drinks. The 'blmr' help page lists the log-odds data and the covariance matrix.

This exploratory analysis makes the strong assumption that the two surveys follow the same seasonal pattern. If the seasonal adjustment were valid, however, the 95% confidence intervals for the changepoint would be

CLR	AF
[2001.29, 2002.88]	[1998.92, 2002.82]

The wide difference here is due to plateaus in the significance levels. 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 (1). The inference mathematics assume that any line slope is possible, extending to an instantaneous drop near December 1998 in this example.

3. Multivariate regression

'blmr' can estimate a changepoint in multivariate regression. Canonical reduction transforms a multivariate regression problem to the form of model (3) for a likelihood-ratio significance test, as Siegmund and Zhang (1994) described. See Hoffman and Kunze (1971, ch.6) and Lehmann (2005, sec. 7.1).

An error-correction model fit US Income and Expenditures data. Construct an orthogonal matrix with first row 1 and second row 'e' to annihilate parameters B_0 , B_1 . In R, the commands are

```
> do this  
> do that
```

This procedure works because the likelihood ratio statistic uses the optimal values for the unknown parameters. The canonical model lets these optimal values reduce their correspondent errors to zero always. Thus they have no effect on inferences, and analysis can omit them.

This procedure tests the hypothesis of a change in one specific parameter, assuming continuity in its coefficient times value. A modified analysis would need to test for an arbitrary change in the regression model.

Conclusion

If a broken line with Normal errors does represent the relationship between a design variable and responses, this package 'blmr' solves the inference step for the changepoint. Fitting a broken line can reveal the plausible region for a change, but practical cause-effect relations usually have smooth transitions. Any statistical analysis should examine the fit of the model and the error distribution with graphs and significance tests, interpret results and report possible alternatives.

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, Ontario: Author. Available at http://www.camh.net/pdf/eb021_ddmonthly.pdf.

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

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.

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, pp. 292-316.

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