Rank Regression Inference via Empirical Likelihood

Ellen E. Bishop

Joint Statistical Meetings

Minneapolis, MN August 7-11, 2005

RTI INTERNATIONAL

RTI International is a trade name of Research Triangle Institute

2951 Flowers Road ■ Atlanta, GA ■ USA 30341
Phone 770-234-5019 Fax 770-234-5030 e-mail ebishop@rti.org

Background

Rank regression is an alternative method for statistical analysis when the assumptions of parametric methods are not sufficiently met. Here the dispersion function is used to estimate beta. Confidence intervals are developed using an empirical likelihood (EL) ratio method.

EL Motivations

- No assumptions are required about the distribution of the residuals
- Does not require constant variance
- Does not require variance estimation, which is required for the normal approximation method
- The confidence intervals are asymmetric
- The confidence intervals reflect the data set

Dispersion Function

Estimate β by minimizing a measure of dispersion of the residuals.

$D(Y - X\beta) = \sum_{i=1}^{N} a(R(y_i - x_i\beta)) * (y_i - x_i\beta)$ where a(·) is a score function,

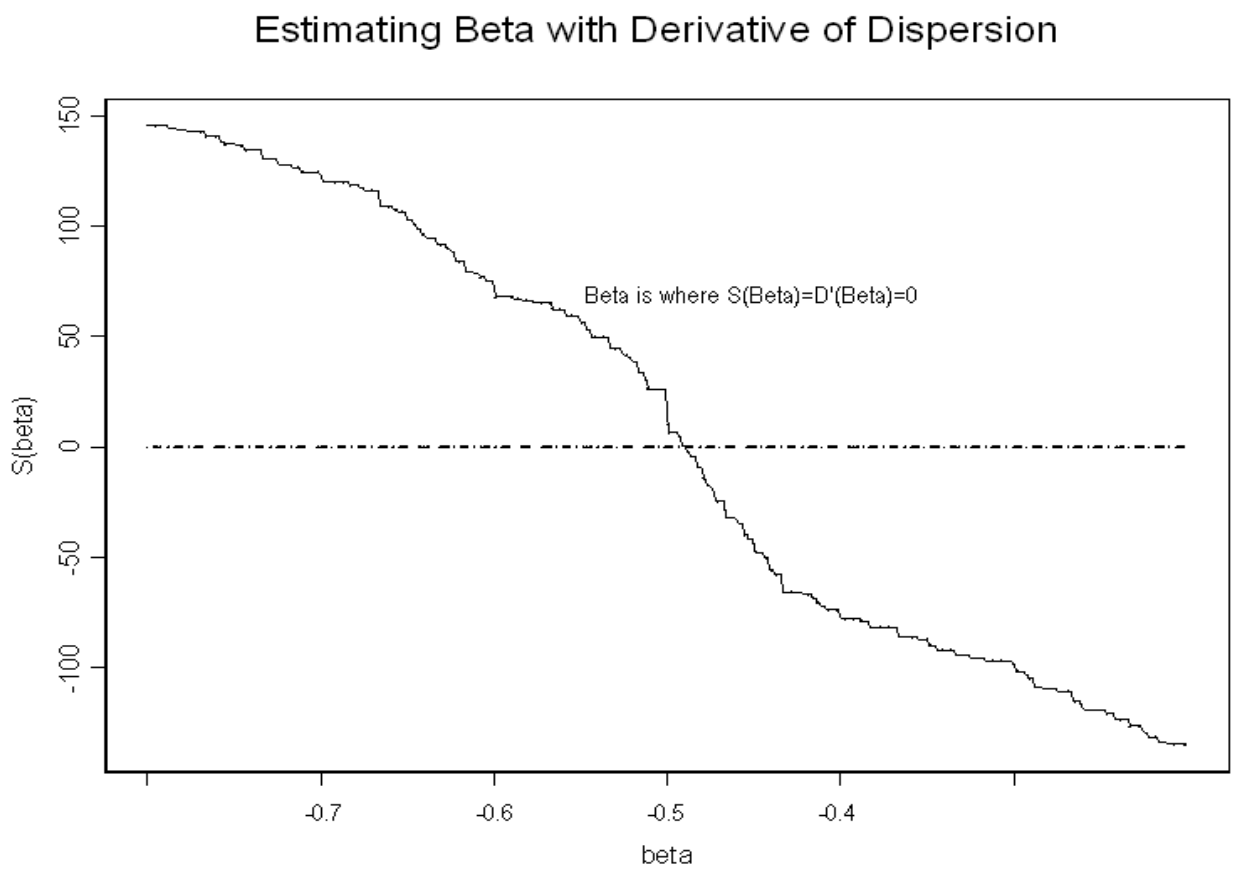
$a(i) = \sqrt{12}\left[\frac{i}{(N+1)} - \frac{1}{2}\right]$ and $R(y_i - x_i\beta)$ is the rank of the errors.

β can be estimated by finding the value that will solve $D'(\beta) = 0$.

α is estimated as the median of $y_i - x_i\hat{\beta}$

$Note: \alpha$ does not effect the rank

Estimating Beta with Dispersion Function
Approximate Beta where D'(Beta)=S(Beta) = 0



The Empirical Likelihood Ratio is Defined As:

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i,$$ where $L(\cdot)$ is the empirical likelihood function and F_n is its maximum, the empirical distribution function.

To estimate a parameter such as mean μ, under mild regularity conditions:

$$R(\mu) = \sup\left\{ \prod_{i=1}^{n} np_i \,\middle|\, p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i x_i = \mu \right\}$$

Also, $-2\log R(\mu) \sim \chi_1^2$

Now Consider the Dispersion Function and Estimating β

Let $\sum_{i\text{-}1}^{n} w_i = D'(\beta)$. It has been shown that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \xrightarrow{D} N(0,\sigma^2) \text{ and } \frac{\sum_{i=1}^{n} w_i^2}{n} \xrightarrow{\Pr} \sigma^2 \text{ as } n \to \infty$$

$$R(\beta_o) = \sup\left\{ \prod_{i=1}^{n} np_i \,\middle|\, \sum_{i=1}^{n} p_i = 1, p_i \geq 0, \sum_{i=1}^{n} p_i w_i = 0 \right\}, \text{ where}$$

w_i is derived with some algebra as:

$$w_i = \sqrt{12}(x_i - \bar{x})\left[\frac{R(y_i - x_i\beta)}{n+1} - \frac{1}{2}\right]$$

Using A Lagrange Multiplier and Knowledge About the Likelihood Distribution:

Using Lagrange multipliers $\prod_{i=1}^{n} np_i$ is maximized when

$p_i = \frac{1}{n}(1+\lambda w_i)^{-1}$ where λ satisfies the equation

$f(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{w_i}{1+\lambda w_i} = 0.$ This ultimately equates to

$$-2\log R(\beta_0) = 2\sum_{i=1}^{n}\log(1+\lambda w_i) \sim \chi^2_{1,.05}$$

A confidence interval is $\mathrm{R} = \left\{\beta : -2\log\mathrm{R}(\beta) \le \chi^2_{1,\alpha}\right\}$

Solve using a bi-section converging algorithm

Simulation Method

True β = -0.5, $\alpha = 100$

Three sample sizes 20, 50, 200

Generate random number for 6 error distributions, ε

Simulate $y = \alpha + \beta x + \varepsilon$ for 2000 sets

Estimate $\hat{\beta}$, using the dispersion function

Create 90%, 95% and 99% confidence intervals

EL Simulation Results

95% Confidence Interval for Beta when true Beta=-0.5 and sample size=20

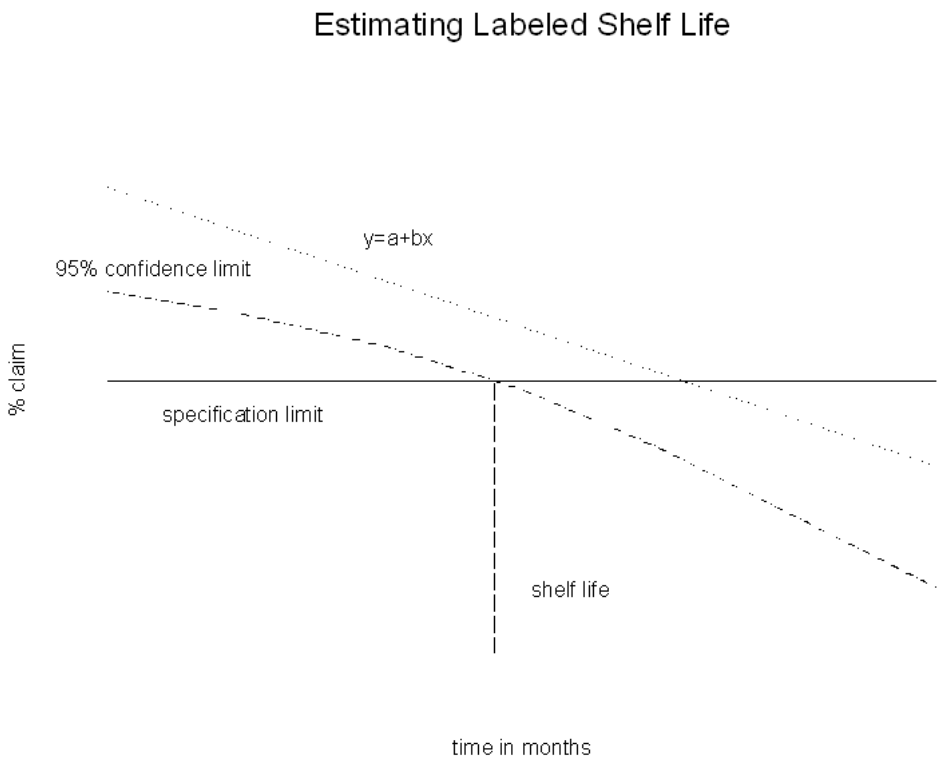
Error	Method	Coverage	Lower	Upper	Length
Normal(0,1)	NA	0.952	-0.960	0.248	1.208
	EL	0.947	-1.002	0.079	1.081
Weibull(2,3)	NA	0.967	-0.973	0.216	1.189
	EL	0.944	-1.075	0.134	1.209
Unif(-2,2)	NA	0.999	-1.125	0.522	1.647
	EL	0.962	-1.156	0.227	1.383

More EL Simulation Results

95% Confidence Interval for Beta when true Beta=-0.5 and sample size=20

Error	Method	Coverage	Lower	Upper	Length
Exp(1)	NA	0.939	-0.806	-0.130	0.676
	EL	0.916	-0.805	-0.126	0.679
LogN(0,1)	NA	0.967	-0.969	0.133	1.102
	EL	0.949	-1.020	0.243	1.263
Logis(1,1)	NA	1.000	-1.312	0.863	2.175
	EL	0.954	-1.467	0.545	2.012

Applied Problem

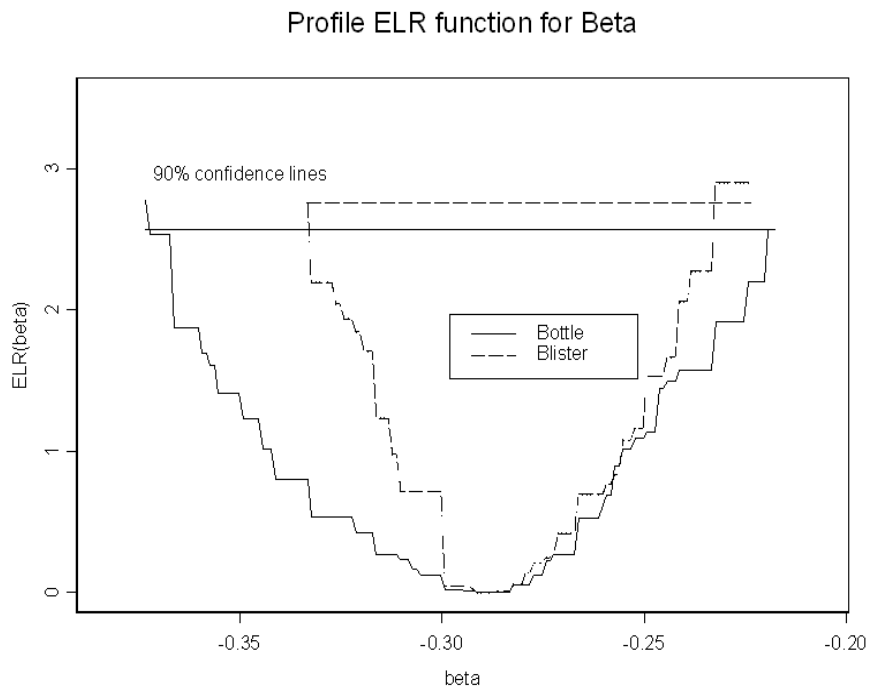


Applied Problem

- Five batches of a drug product stored in one of two package types, bottle or blister.
- Measurements of the drug's potency are taken at six time points, 0,3,6,9,12,and 18 months.
- The measurement is recorded as a % of claim.

Applied Problem

Shelf Life for drug stored in bottle or blister package
90% Confidence Interval for Beta
Bottle [-0.37,-0.22] Blister [-0.33, -0.23]



Shelf-life Using Adjusted Specification Beta estimated using Least Squares vs. Dispersion

Package	Result	Least Square	Dispersion
Bottle	Average Beta	-0.326	-0.284
	Shelf-life (months)	22	22
Blister	Average Beta	-0.254	-0.296
	Shelf-life (months)	21	20

Conclusion

- Empirical Likelihood Ratio Confidence Interval
  - accurate coverage probabilities
  - does not require variance estimation
  - weak assumptions about a distribution
  - asymmetric intervals
  - range respecting, reflect the data set

Acknowledgements

References are detailed in the accompanying paper, Rank Regression Inference via Empirical Likelihood

Dr. Zhao, my advisor, for his time, support and knowledge.

RTI for sponsoring my attendance at JSM.