Statistical Methods for the Blood Beryllium Lymphocyte Proliferation Test*

Edward L. Frome† Matthew H. Smith‡ L. Gayle Littlefield‡

Richard L. Neubert§ Shirley P. Colyer‡

Abstract

The blood beryllium lymphocyte proliferation test (BeLPT) is a modification of the standard lymphocyte proliferation test that is used to identify persons who may have chronic beryllium disease. A major problem in the interpretation of BeLPT test results is outlying data values among the replicate well counts ($\approx 7\%$). A log-linear regression model is used to describe the expected well counts for each set of Be exposure conditions, and the variance of the well counts is proportional to the square of the expected count. Two outlier resistant regression methods are used to estimate stimulation indices (SIs) and the coefficient of variation. The first approach uses least absolute values (LAV) on the log of the well counts as a method for estimation; the second approach uses a resistant regression version of maximum quasi-likelihood estimation. **A major advantage of these resistant methods is that they make it unnecessary to identify and delete outliers**. These two new methods for the statistical analysis of the BeLPT data and the current outlier rejection method are applied to 173 BeLPT assays. **We strongly recommend the LAV method for routine analysis of the BeLPT.**

Outliers are also important when trying to identify individuals with beryllium hypersensitivity, since these individuals typically have large positive SF values. A new method for identifying large SIs using combined data from the not exposed group and the beryllium workers is proposed. The log(SI)s are described with a Gaussian distribution with location and scale parameters estimated using resistant methods. This approach is applied to the test data and results are compared with those obtained from the current method.

*Published in **Environmental Health Perspectives 104, Supplement 5 (1996), pp 957-968. Research was supported by the Offices of Occupational Medicine and Epidemiology and Health Surveillance, Environment, Safety and Health, U. S. Department of Energy under contract DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp., and DA-AC05-76OR00033 with Oak Ridge Associated Universities.**

†Mathematical Sciences Section, Computer Science and Mathematics Division, Oak Ridge National Laboratory, P. O. Box 2008, Oak Ridge, Tennessee 37831-6367

‡Medical Sciences Division, Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

§Consultant, Oak Ridge,Tennessee 37830

approaches for estimating the SIs and the coefficient of variation, ϕ.

Second Method Based on LAV Regression on Log(y)

The first approach based on the regression model is to take the log of the counts since this is the variance-stabilizing transformation and leads to a linear model in $z_{jk} = \log(y_{jk})$, i.e.

$$E(z_{jk}) = X_j\beta - \phi^2/2 \text{ and } Var(z_{jk}) \simeq \phi^2.$$

In the absence of outliers ordinary least squares on the transformed data would yield consistent estimates for the log(SI) parameters (5). The effect of outliers is minimized by using least absolute values (or some other robust method) on the z_{jk}. LAV regression — also known as L_1 norm, least absolute deviations (LAD) and minimum sum of absolute errors (MSAE) — is well known to be resistant to outliers and is an important particular case of a general class of robust methods known as M-estimators (6, 7). In general, LAV regression requires special computational resources to calculate parameter estimates (8). In this situation, however, it is only necessary to find the median of the log of the well counts for each set of design conditions (say $\tilde{z}_j$) and then subtract the control median for each harvest day from the beryllium-stimulate medians. Frome *et. al* present details in Appendices A and C of report a for an Oak Ridge National Laboratory (9). A resistant estimate of the coefficient of variation can then be obtained as

$$\tilde{\phi}_L = C \times median\{|z_{jk} - \tilde{z}_j|\},$$

where $C = 1.48 \times \sqrt{n/(n-p)}$, $n = 56$, and $p = 10$ (when the assay is complete). The value of C is chosen to make the estimate consistent for the standard deviation for a Gaussian error model and for consistency with the usual least squares results in which the estimated variance is multiplied by the correction factor $n/(n-p)$ (10) and S-PLUS function `mad` (11) which computes the median absolute deviation (MAD) estimate of the standard deviation. Alternative approaches to estimating ϕ have been discussed in the context of LAV regression (12, 7) and there is no consensus as to the best approach. In addition to the fact that ϕ is of direct interest, it is also needed to obtain an estimate of the parameter covariance matrix

$$\omega^2 (X'X)^{-1},$$

where $\omega^2 = [2f(0)]^{-2}$ is the asymptotic variance of the sample median (13). Following the approach of McGill it et. al (14) we assume that the underlying error distribution is Gaussian in the center and use $\tilde{\omega} = \sqrt{\pi/2}\tilde{\phi}_L$ to obtain an estimate of the standard deviation of the log of the stimulation indices. The appropriate diagonal term from $(X'X)^{-1}$ is $4/12$, and consequently the estimated standard deviation of log(SI) is $1.25\tilde{\phi}_L(0.58) = 0.72\tilde{\phi}_L$. The results of applying this approach to the data in Table 1 are shown in Table 4.

Table 4: Results of LAV estimation for log(y) of data in Table 1, $\tilde{\phi}_L = 0.367$

Experimental Conditions		z_{jk}				$\tilde{z}_j$	$(z_{jk} - \tilde{z}_j)/\tilde{\phi}_L$				$\tilde{\beta}$	$\exp(\tilde{\beta})$
Day 5	Controls	6.872	7.067	6.719	6.759	7.182	-0.8	-0.3	-1.3	-1.2		
	Controls	7.296	8.887	6.929	6.883	7.182	0.3	4.6	-0.7	-0.8		
	Controls	7.313	7.455	7.422	7.597	7.182	0.4	0.7	0.7	1.1		
	Be1	6.957	6.560	7.268	6.532	6.758	0.5	-0.5	1.4	-0.6	-0.423	0.655
	Be10	7.347	7.290	7.415	7.741	7.381	-0.1	-0.2	0.1	1.0	0.199	1.221
	Be100	8.181	8.662	8.297	8.562	8.429	-0.7	0.6	-0.4	0.4	1.248	3.483
Day 7	Controls	9.127	8.567	8.239	8.559	8.139	2.7	1.2	0.3	1.1		
	Controls	7.745	7.953	7.557	8.040	8.139	-1.1	-0.5	-1.6	-0.3		
	Controls	7.807	8.278	8.035	8.793	8.139	-0.9	0.4	-0.3	1.8		
	Be1	6.571	7.034	8.713	7.000	7.017	-1.2	0.0	4.6	0.0	-1.122	0.326
	Be10	6.667	6.741	7.922	6.480	6.704	-0.1	0.1	3.3	-0.6	-1.436	0.238
	Be100	8.706	9.030	8.832	9.254	8.931	-0.6	0.3	-0.3	0.9	0.792	2.207
Day 5	Pha	11.320	10.877	10.872	10.829	10.874	1.2	0.0	0.0	-0.1	4.792	120.50
	Candida	10.477	9.982	9.978	10.003	9.992	1.3	0.0	0.0	0.0	3.910	49.880

Third Method Based on Quasi-Likelihood Estimation

In the second regression model approach, the analysis is done on the original scale and estimation is based on an iterative weighted least squares (IWLS) algorithm. The use of IWLS for generalized linear (15) and nonlinear (16) regression functions leads to maximum likelihood estimates when the dependent variable is in the regular exponential family. McCullagh (17) extended this result to quasi-likelihood (QL) estimation, which requires specification of the mean and variance function. Extension of IWLS to resistant/robust regression has been described by Green (18) and Pregibon (19), and the computational approach described by Chambers and Hastie