TEACHER'S CORNER

In this department *The American Statistician* publishes articles, reviews, and notes of interest to teachers of the first mathematical statistics course and of applied statistics courses. The department includes the Accent on Teaching Materials section; suitable contents for the section are described under the section heading. Articles and notes for the department, but not intended specifically for the section, should be useful to a substantial number of teachers of the indicated types of courses or should have the potential for fundamentally affecting the way in which a course is taught.

Teaching About Approximate Confidence Regions Based on Maximum Likelihood Estimation

William Q. MEEKER and Luis A. ESCOBAR

Maximum likelihood (ML) provides a powerful and extremely general method for making inferences over a wide range of data/model combinations. The likelihood function and likelihood ratios have clear intuitive meanings that make it easy for students to grasp the important concepts. Modern computing technology has made it possible to use these methods over a wide range of practical applications. However, many mathematical statistics textbooks, particularly those at the Senior/Masters level, do not give this important topic coverage commensurate with its place in the world of modern applications. Similarly, in nonlinear estimation problems, standard practice (as reflected by procedures available in the popular commercial statistical packages) has been slow to recognize the advantages of likelihood-based confidence regions/intervals over the commonly use "normal-theory" regions/intervals based on the asymptotic distribution of the "Wald statistic." In this note we outline our approach for presenting, to students, confidence regions/intervals based on ML estimation.

KEY WORDS: Asymptotic approximation; Confidence interval; Large sample approximation; Profile likelihood.

1. INTRODUCTION

Because of its versatility and favorable large sample asymptotic properties, the method of maximum likelihood (ML) is probably the most widely used method of estimation for parametric statistical models. Applications extend to important areas like time series, survival analysis, categorical data analysis, variance components, spatial data analysis, errors in variables, and so on. Also, linear and nonlinear least squares estimators are equivalent to ML estimators based on an assumed normal distribution for the residual term. The standard textbooks in these areas (e.g., Agresti 1990; Bates and Watts 1988; Box and Jenkins 1976; Cox and Oakes 1984; Cressie 1991; Fuller 1987; Lawless 1982; Nelson 1990; Searle, Casella, and McCulloch 1992; Seber and Wild 1989) usually discuss ML estimation and, in some cases, related methods of computing confidence regions/intervals for the parameters or functions of the parameters.

Most textbooks on the theory of mathematical statistics (e.g., Bain and Engelhardt 1987; Casella and Berger 1990; Cox and Hinkley 1974; Rao 1973; Stuart and Ord 1991) describe ML estimation and go on to describe related confidence regions/intervals. For some models (e.g., linear regression with normally distributed residuals and no censoring) there are useful results based on exact distribution theory. In general, however, we have to rely on asymptotic theory.

Many mathematical statistics textbooks do not give a sense of the wide range of areas where ML methods are used. Also, because students at this level are only exposed to first-order asymptotic results, they leave their theory courses with the incorrect impression that there is little difference between the asymptotically equivalent Wald and likelihood-based methods of setting confidence regions/intervals. It is clear, however, that the likelihood-based methods have important advantages. We suggest an approach for teaching this material that we feel is more interesting, more useful, and more in line with today's computational capabilities.

2. OVERVIEW OF STANDARD ASYMPTOTIC SAMPLING DISTRIBUTION THEORY

2.1 Likelihood Ratio and Wald Statistics

Assume that we have a model with a vector $\theta = (\theta_1, \theta_2)$ of unknown parameters, partitioned in order to obtain a confidence region for θ_1 with θ_2 being nuisance parameters, and let $L(\theta)$ denote the corresponding likelihood. We also let $k_1 = \text{length}(\theta_1)$ and let $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2)$ denote the corresponding ML estimators. Also, define the profile likelihood for θ_1 as

$$R(\theta_1) = \max_{\theta_2} \left[\frac{L(\theta_1, \theta_2)}{L(\widehat{\theta})} \right].$$

William Q. Meeker is Professor of Statistics, Department of Statistics, Iowa State University, Ames, IA 50011. Luis A. Escobar is Professor of Statistics, Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803. The authors thank two anonymous referees and an associate editor who made a number of comments and suggestions that helped improve this article.

Under regularity conditions (e.g., Lehmann 1983, p. 429) the ML estimators are unique, and we have, if the true value of the parameter vector is $\theta_1 = \theta_{10}$, the following important and well-known asymptotic distributional results. As $n \to \infty$, both the "likelihood ratio subset statistic"

$$-2\log[R(\theta_{10})]$$

and the "Wald subset statistic"

$$[\hat{\theta}_1 - \theta_{10}]'[\hat{\Sigma}_{\hat{\theta}_1}]^{-1}[\hat{\theta}_1 - \theta_{10}]$$

follow a chi-square distribution with k_1 degrees of freedom. Here we take $\hat{\Sigma}_{\hat{\theta}_1}$ to be the estimate of the variance–covariance matrix of $\hat{\theta}_1$ obtained by taking the first k_1 rows and columns of the inverse of the observed information matrix for θ.

In situations where the number of model parameters increases with sample size, ML estimators may not even be consistent (the standard regularity conditions assume that the number of parameters is fixed). As discussed by Kalbfleisch and Sprott (1970), this "incidental parameters" problem is an indication that asymptotic approximations may not be adequate when the number of parameters is large relative to the number of observations.

2.2 Confidence Regions/Intervals Based on Asymptotic Sampling Distribution Theory

The asymptotic distributional results in Section 2.1 provide approximate tests of hypotheses, and these tests can be inverted to obtain (approximate) confidence regions/intervals (e.g., Lehmann 1986). These approximate confidence regions have the usual frequentist interpretation: The probability that a random confidence region/interval will cover the true value of θ_1 is, in large samples, approximately $1 - \alpha$.

The following asymptotic results suggest that the resulting regions/intervals are asymptotically equivalent:

- An approximate $100(1 - \alpha)\%$ likelihood-based confidence region for θ_1 is the set of all values of θ_{10} such that

$$-2\log[R(\theta_{10})] < \chi^2_{(1-\alpha;k_1)}.$$

- An approximate $100(1 - \alpha)\%$ normal-theory (Wald) confidence region for θ_1 is the set of all the θ_{10}'s in the ellipsoid

$$[\hat{\theta}_1 - \theta_{10}]'[\hat{\Sigma}_{\hat{\theta}_1}]^{-1}[\hat{\theta}_1 - \theta_{10}] \le \chi^2_{(1-\alpha;k_1)}, \qquad (1)$$

where $\chi^2_{(1-\alpha;k_1)}$ is the $1 - \alpha$ quantile of the chi-square distribution with k_1 degrees of freedom. Cox and Hinkley (1974, p. 321) provided an explicit proof that Wald and likelihood-ratio confidence regions are asymptotically equivalent. In the Appendix we show that the Wald confidence region (interval) can be interpreted as a confidence region (interval) based using a quadratic approximation to the log likelihood. Similar ideas were presented for estimation problems with a single parameter in Sprott (1973) and for estimation problems with nuisance parameters in Sprott (1980). In certain special cases (particularly when the log likelihood function is quadratic in the unknown parameters), the Wald and the likelihood ratio statistics are equivalent and there is exact distribution theory.

A third alternative, the so-called score statistic, also has the same asymptotic distribution, and can also be viewed as a quadratic approximation to the log likelihood (e.g., Sprott 1980).

2.3 Using Transformations to Improve the Wald Approximation

The accuracy of the Wald approximation depends on parameterization. Sprott (1973, 1975) suggested that the normality of the relative likelihood (or quadratic shape of the log-likelihood) be used as a criterion to judge the adequacy of the large sample approximation. Cook and Weisberg (1990), in the context of nonlinear regression, gave examples and a method of plotting the profile likelihood that allows assessment of the Wald approximation. Anscombe (1964) and Sprott (1973, 1975), for example, showed how reparameterization can, to some degree, be used to improve the asymptotic approximation. The basic idea is to find a parameterization that will, as much as possible, make the log-likelihood approximately quadratic. Of course, finding a good parameterization may be nearly as difficult as using the likelihood ratio method (but once determined for a class of problems, an appropriate reparameterization could save computational time for that class of problems). Sprott (1973) also indicated and gave an example of a situation where finding such a transformation will *not* be possible (also see Example 3 in Section 4.2).

Likelihood-based confidence regions/intervals are invariant to such transformations, and generally do as well or better than the best transformation. This is closely related to the "parameter effects" ideas described, for example, in Bates and Watts (1988).

3. CHOOSING BETWEEN LIKELIHOOD AND WALD APPROACHES

Particularly when the focus is on theory, students are often left with the mistaken impression that the Wald and likelihood approaches provide equally accurate approximations (e.g., Rao 1973, p. 418). In fact, most commercial statistical computer packages use the inferior Wald approach (SAS JMP's nonlinear regression procedure is a notable exception) with nonlinear estimation. The reason for this is some combination of (a) lack of knowledge among statisticians about the advantages of likelihood-based methods, and (b) more complicated computations are required for the likelihood-based approach.

3.1 Advantages of Likelihood-Based Methods

Especially more recently, a number of authors have recognized and reported the advantages of likelihood-based inference (e.g., Beale 1960; Cox and Hinkley 1974, pp. 342–343; Cox and Oakes 1984, p. 36; Kalbfleisch and Prentice 1980, p. 48; Lawless 1982, p. 525). Numerous simulation studies have shown clear advantages for likelihood-based intervals for a number of specific models (e.g., Donaldson and Schnabel 1987; Meeker 1987; Ostrouchov and Meeker 1988; Vander Wiel and Meeker 1990). On the theoretical side, in the context of nonlinear regression, Bates and Watts (1988) made a strong case on the basis of parameter-effects curvature–transformation. Also see Cook and Weisberg (1990).

3.2 Computational Issues

Likelihood-based regions/intervals are more difficult to compute than the corresponding Wald regions/intervals. In a problem with nuisance parameters, finding a likelihood-based region/interval involves finding the roots of a function where each function evaluation requires a constrained maximum likelihood estimation. Venzon and Moolgavkar (1988) and Cook and Weisberg (1990) described useful algorithms for the process. With improvements in computing technology, this is less of a problem today than in the past, and the direction for the future is clear; lack of appropriate easy-to-use software is the main problem.

4. EDUCATIONAL ISSUES

4.1 Outline of Lecture Material

When teaching this material, we suggest, instead of the traditional approach, something similar to the following outline for the presentation of these important results. Emphasis given to the different points will depend on whether the presentation is for a theory or for a methods course.

- Present asymptotic distributional results and applications to hypothesis testing for both the Wald and the likelihood ratio methods.
- Show how to invert the hypothesis tests to construct confidence regions/intervals for both methods. Indicate computational differences, noting that, in general, the inversion of the likelihood ratio test must be done numerically.
- Use the simple proof in the Appendix to show that the region/interval given by the Wald approach is based on a quadratic approximation to the log-likelihood.
- Show, by example, that the quadratic approximation for the log-likelihood could give inaccurate and, in extreme cases, nonsensical intervals (e.g., negative lower limits for scale parameters or limits for probabilities that are outside of the range of [0–1]) in situations where the likelihood is far from quadratic.
- As discussed in Section 2.3, show that the accuracy of the Wald approach depends on parameterization, and describe and illustrate how transformations of parameters can, to some degree, be used to improve the adequacy of the Wald approach.
- Describe the close relationship between likelihood-based confidence intervals and Bayesian highest posterior density (HPD) intervals (e.g., Casella and Berger 1990, p. 424; Severini 1991).

An important point of the discussion, made frequently by others in the past (e.g., Box and Jenkins 1976, pp. 224–226), but not sufficiently in standard textbooks, is that one should, in unfamiliar nonlinear estimation problems, examine a plot of the likelihood and profile likelihoods rather than just computing summary test statistics.

4.2 Examples

Examples like the following are useful for illustrating the points in Section 4.1.

Example 1. This example addresses inferences on lognormal distribution parameters and quantiles. It provides illustrative analytical formulas for the approximate likelihood-based methods and has an exact solution (based on the noncentral t distribution) with which to compare. Also, the example is similar to practical problems where exact results are not available and where asymptotic approximations are commonly used: problems with censored data and nonnormal distributions, like the Weibull and Gamma.

The lognormal density is

$$f_T(t;\mu,\sigma) = \frac{C}{\sqrt{\sigma^2 t}}\exp\left\{-\frac{1}{2\sigma^2}[\log(t)-\mu]^2\right\} = \frac{C}{\exp(y)}\exp\left\{-\frac{1}{2}\left[\frac{(y-\mu)^2}{\sigma^2}+\log(\sigma^2)\right]\right\}$$

where $C = 1/\sqrt{2\pi}$ and $y = \log(t)$. The likelihood for n independent observations from a lognormal distribution is $L(\mu,\sigma) = \prod_{i=1}^{n} f_T(t_i;\mu,\sigma)$. Setting the first partial derivatives of $\log[L(\mu,\sigma)]$ with respect to μ and σ equal to 0 and solving, one gets the ML estimates, $\hat{\mu} = \bar{y}$ and $\hat{\sigma} = \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2/n}$ where $\bar{y}$ is the sample mean of the y_i's.

Expressing the likelihood as

$$L(\mu,\sigma) = \frac{C^n}{\exp(n\hat{\mu})}\exp\left\{-\frac{n}{2}\left[\frac{\hat{\sigma}^2+(\hat{\mu}-\mu)^2}{\sigma^2}+\log(\sigma^2)\right]\right\}$$

shows that $(\hat{\mu},\hat{\sigma})$ are sufficient statistics for μ and σ. Then the relative likelihood takes the simple form

$$R(\mu,\sigma) = \frac{L(\mu,\sigma)}{L(\hat{\mu},\hat{\sigma})} = \exp\left\{-\frac{n}{2}\left[\frac{\hat{\sigma}^2-\sigma^2+(\hat{\mu}-\mu)^2}{\sigma^2}+\log\left(\frac{\sigma^2}{\hat{\sigma}^2}\right)\right]\right\}. \quad (2)$$

For fixed σ, the value of μ that maximizes $R(\mu,\sigma)$ is $\tilde{\mu} = \hat{\mu}$. Substituting this for μ in (2) gives a simple expression for $R(\sigma)$. For fixed μ, the value of σ that maximizes $R(\mu,\sigma)$ is $\tilde{\sigma} = \sqrt{\hat{\sigma}^2+(\hat{\mu}-\mu)^2}$. Substituting this for σ in (2) gives a simple expression for $R(\mu)$.

We now consider the lognormal P quantile $T_P = \exp(\mu + Z_P\sigma)$ where Z_P is the standard normal P quantile. To do this, we reparameterize in terms of T_P and σ by substituting $\log(T_P) - Z_P\sigma$ for μ on the right-hand side of (2), giving the relative likelihood

$$R(T_P,\sigma) = \exp\left\{-\frac{n}{2}\left[\frac{\hat{\sigma}^2-\sigma^2+(v-\sigma Z_P)^2}{\sigma^2}+\log\left(\frac{\sigma^2}{\hat{\sigma}^2}\right)\right]\right\} \quad (3)$$

where we use $v = \log(T_P) - \hat{\mu}$ to simplify the presentation.

To obtain $R(T_P)$, maximize $R(T_P,\sigma)$ with respect to σ for fixed values of T_P. Setting $\partial\log[R(T_P,\sigma)]/\partial\sigma = 0$ and simplifying gives $\sigma^2 + vZ_P\sigma - (\hat{\sigma}^2+v^2) = 0$. One of the roots of this quadratic is always positive and the other is always negative. It can be shown that the positive root

$$\tilde{\sigma} = \frac{1}{2}\left[-vZ_P + \sqrt{v^2Z_P^2+4\hat{\sigma}^2+4v^2}\right]$$

is a maximum. Then substituting $\tilde{\sigma}$ for σ in (3) gives $R(T_P)$, the profile likelihood for T_P.

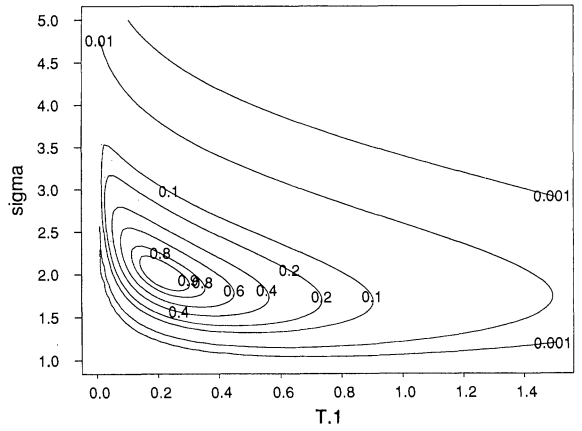

Figure 1. Relative Likelihood $R(T_{.1}, \sigma)$ for the Lognormal Distribution With $n = 10$, $\bar{y} = 1$, and $\hat{\sigma} = 2$.

We now specialize to the .1 quantile and use for data $n = 10$, and the sufficient statistics $\hat{\mu} = 1$ and $\hat{\sigma} = 2$. Figure 1 is a contour plot of $R(T_P, \sigma)$. Because $\chi^2_{(1-\alpha;2)} = -2\log(\alpha)$, the α contour of a two-dimensional relative likelihood corresponds to an approximate $100(1 - \alpha)\%$ joint confidence region for T_P, σ.

Figure 2 compares plots of $R(T_{.1})$ as well as the corresponding Wald quadratic approximations $\exp[(-1/2)W_1]$ and $\exp[(-1/2)W_0]$ in the scale of $T_{.1}$ and $\log(T_{.1})$, respectively. These special cases of (1) are computed as

$$W_1(T_P) = \frac{[\hat{T}_P - T_P]^2}{\widehat{\text{var}}[\hat{T}_P]}$$

$$W_0(T_P) = \frac{[\log(\hat{T}_P) - \log(T_P)]^2}{\widehat{\text{var}}[\log(\hat{T}_P)]}$$

where $\widehat{\text{var}}[\log(\hat{T}_P)] = (\hat{\sigma}^2/n)[1 + Z_P^2/2]$ is obtained by evaluating the second derivatives of the log-likelihood at the ML estimates and, using the delta method, $\widehat{\text{var}}[\hat{T}_P] = \hat{T}_P^2 \widehat{\text{var}}[\log(\hat{T}_P)]$. Corresponding numerical values of the confidence interval endpoints can be read from the graph, but are also shown in Table 1. These show that the likelihood-based intervals provide the best approximation to the exact-theory interval (obtained by using Table A.12 in Hahn and Meeker 1991). The quadratic approximation

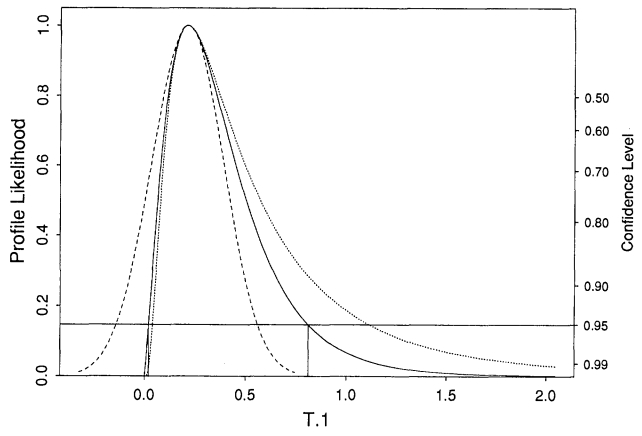

Figure 2. Profile Likelihood $R(T_{.1})$ (solid curve) for the Lognormal Distribution With $n = 10$, $\bar{y} = 1$, and $\hat{\sigma} = 2$ with Corresponding Quadratic Approximations in $T_{.1}$ (dashed curve) and $\log(T_{.1})$ (dotted curve).

Table 1. Comparison of Confidence Intervals For Example 1

Confidence Interval Method	Endpoints	
	Lower	Upper
Exact	.014	.793
Likelihood	.020	.812
Wald in $\log(T_{.1})$	.039	1.116
Wald in $T_{.1}$	−.141	.560

in $T_{.1}$ results in a nonsensical negative lower confidence interval endpoint. The commonly used quadratic approximation in $\log(T_{.1})$ corrects this problem but still does relatively poorly for the upper endpoint of the confidence interval.

Example 2. Figure 3, computed from the model used in Arnold, Beaver, Groeneveld, and Meeker (1993), provides another, quite different example where the quadratic approximation to the log-likelihood profile is also inadequate. In their model and this particular set of data, the matrix of second partial derivatives of the log-likelihood with respect to the model parameters, evaluated at the maximum likelihood estimates of the parameters, is nearly singular. This gives an indication that one cannot obtain a Wald-based confidence interval for λ. The log-likelihood profile, however, gives a more accurate picture of the information that the data provide about λ.

Example 3. Figure 4, computed from the model used in Meeker (1987), provides an example where the quadratic approximation provided by the Wald statistic would be seriously inadequate. Nor is there a transformation that will make the log-likelihood approximately quadratic. Because the log-likelihood profile for "proportion defective" p flattens out at a high level, it is quite plausible that the data could have come from a model in which the proportion defective was as high as 1. The Wald-based confidence interval for the proportion defective is, however, [.0034, .0200] which, in this case, would be seriously misleading.

This example illustrates a further advantage of the profile likelihood approach: the theory can be generalized to handle situations in which the parameter vector is near

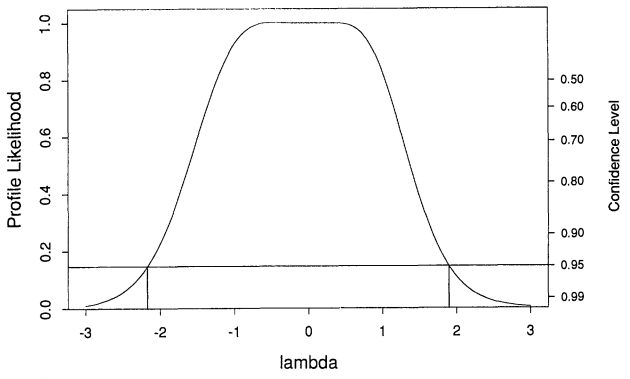

Figure 3. Profile Likelihood $R(\lambda)$ With Approximate 95% Likelihood-Based Confidence Interval for λ, Based on Model and Data From Arnold, Beaver, Groeneveld, and Meeker (1993).

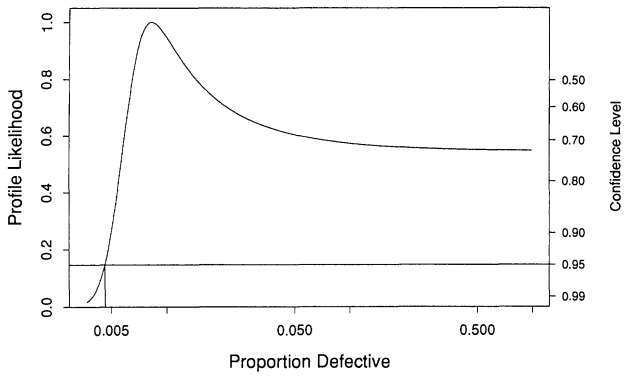

Figure 4. Profile Likelihood R(p) With Approximate 95% Likelihood-Based Confidence Interval for Proportion Defective p, Based on Model and Data From Meeker (1987).

or on the boundary of the parameter space (see Chernoff 1954 and Feder 1968).

5. DISCUSSION

We have attempted to build a case for the use of likelihood-based confidence intervals and regions and for giving these methods a more prominent place in the statistics curricula. We would, however, like to reiterate an important point made by Casella and Berger (1990, p. 416), among others. In general, there is no guarantee that likelihood-based methods are optimum, although they will seldom be too bad. Also, for some problems, even today, the computational effort will be substantial and, for specific problems, there may be other, simpler methods that one can or should use.

APPENDIX: WALD CONFIDENCE REGIONS APPROXIMATE PROFILE LIKELIHOOD CONFIDENCE REGIONS

In this Appendix we show that the Wald confidence region can be viewed as an approximation to the likelihood-based region. The approximation is based on quadratic approximation to the log-likelihood profile. We again use the partition $\theta = (\theta_1, \theta_2)$ in order to obtain a confidence region for θ_1 and θ_2 denotes the nuisance parameters. Minus 2 times the log likelihood profile for θ_1 is

$$\Lambda(\theta_1) = -2[\mathcal{L}(\theta_1, \widehat{\widehat{\theta}}_2) - \mathcal{L}(\widehat{\theta}_1, \widehat{\theta}_2)]$$

where $\mathcal{L}(\cdot)$ is the log-likelihood, $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2)$ is the MLE estimator of the parameters $\theta = (\theta_1, \theta_2)$, and $\widehat{\widehat{\theta}}_2 = \widehat{\widehat{\theta}}_2(\theta_1)$ is the MLE of θ_2 with θ_1 fixed.

Result. The confidence region for θ_1 obtained by inverting the Wald test is equivalent to the confidence region obtained by using a quadratic approximation to the log-likelihood profile.

Proof. In what follows, we assume the following mild regularity conditions: (a) the MLE of θ exist and it is unique, and (b) the observed information matrix of θ evaluated at $\widehat{\theta}$ is positive definite.

Expanding $\Lambda(\theta_1)$ in a Taylor series about $\widehat{\theta}_1$ gives the desired quadratic approximation for the log-likelihood profile:

$$Q(\theta_1) = \Lambda(\widehat{\theta}_1) + (\theta_1 - \widehat{\theta}_1)' \frac{\partial \Lambda(\theta_1)}{\partial \theta_1} + \left(\frac{1}{2}\right)(\theta_1 - \widehat{\theta}_1)' \frac{\partial^2 \Lambda(\theta_1)}{\partial \theta_1 \partial \theta_1'} (\theta_1 - \widehat{\theta}_1)$$

where the derivatives are evaluated at $\theta_1 = \widehat{\theta}_1$ and $\theta_2 = \widehat{\widehat{\theta}}_2(\theta_1)$. Because $\Lambda(\theta_1)$ and $\mathcal{L}(\theta_1, \widehat{\widehat{\theta}}_2)$ are maximized at $\theta_1 = \widehat{\theta}_1$, we have that $\Lambda(\widehat{\theta}_1) = 0$ and $\frac{\partial \Lambda(\theta_1)}{\partial \theta_1}\big|_{\theta_1 = \widehat{\theta}_1} = 0$. Thus the quadratic approximation simplifies to

$$Q(\theta_1) = \left(\frac{1}{2}\right)(\theta_1 - \widehat{\theta}_1)' \frac{\partial^2 \Lambda(\theta_1)}{\partial \theta_1 \partial \theta_1'} (\theta_1 - \widehat{\theta}_1).$$

The Wald subset statistic has the form

$$W(\theta_1) = (\theta_1 - \widehat{\theta}_1)' [I_{11} - I_{12} I_{22}^{-1} I_{21}] (\theta_1 - \widehat{\theta}_1)$$

where

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} = - \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_1'} & \dfrac{\partial^2 \mathcal{L}(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2'} \\ \dfrac{\partial^2 \mathcal{L}(\theta_1, \theta_2)}{\partial \theta_2 \partial \theta_1'} & \dfrac{\partial^2 \mathcal{L}(\theta_1, \theta_2)}{\partial \theta_2 \partial \theta_2'} \end{bmatrix}$$

is the observed information matrix [the derivatives here are evaluated at $(\theta_1, \theta_2) = (\widehat{\theta}_1, \widehat{\theta}_2)$].

Using Theorem 2.2 (p. 39) in Seber and Wild (1989), it follows that

$$\left(\frac{1}{2}\right) \frac{\partial^2 \Lambda(\theta_1)}{\partial \theta_1 \partial \theta_1'} = I_{11} - I_{12} I_{22}^{-1} I_{21}.$$

Thus $W(\theta_1) = Q(\theta_1)$ and $\exp[-(1/2)\, W(\theta_1)] \doteq R(\theta_1)$.

[*Received June 1993. Revised June 1994.*]

REFERENCES

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley.

Anscombe, F. J. (1964), "Normal Likelihood Functions," *Annals of the Institute of Statistical Mathematics*, 16, 1–19.

Arnold, B. C., Beaver, R. J., Groeneveld, R. A., and Meeker, W. Q. (1993), "The Nontruncated Marginal of a Truncated Bivariate Normal Distribution," *Psychometrika*, 58, 471–488.

Bain, L. J., and Engelhardt, M. (1987), *Introduction to Probability and Mathematical Statistics*, Boston: PWS Publishers.

Bates, D. M., and Watts, D. G. (1988), *Nonlinear Regression Analysis and Its Applications*, New York: John Wiley.

Beale, E. M. L. (1960), "Confidence Regions in Nonlinear Regression" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 22, 41–88.

Box, G. E. P., and Jenkins, G. W. (1976), *Time Series Analysis: Forecasting and Control* (revised ed.), San Francisco: Holden-Day.

Casella, G., and Berger, R. L. (1990), *Statistical Inference*, Belmont, CA: Wadsworth.

Chernoff, H. (1954), "On the Distribution of the Likelihood Ratio," *Annals of Mathematical Statistics*, 25, 573–578.

Cook, R. D., and Weisberg, S. (1990), "Confidence Curves in Nonlinear Regression," *Journal of the American Statistical Association*, 85, 544–551.

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

Cox, D. R., and Oakes, D. (1984), *Analysis of Survival Data*, New York: Methuen.

Cressie, N. A. C. (1991), *Spatial Statistics*, New York: John Wiley.

Donaldson, J. R., and Schnabel, R. B. (1987), "Computational Experience with Confidence Regions and Confidence Intervals for Nonlinear Least Squares," *Technometrics*, 29, 67–82.

Feder, P. I. (1968), "On the Distribution of the Log Likelihood Ratio Statistic When the True Parameter is Close to the Boundary of the Hypothesis Regions," *Annals of Mathematical Statistics*, 39, 2044–2055.
Fuller, W. A. (1987), *Measurement Error Models*, New York: John Wiley.
Hahn, G. J., and Meeker, W. Q. (1991), *Statistical Intervals: A Guide for Practitioners*, New York: John Wiley.
Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley.
Kalbfleisch, J. D., and Sprott, D. A. (1970), "Applications of Likelihood Methods to Models Involving Large Numbers of Parameters" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 32, 175–208.
Lawless, J. F. (1982), *Statistical Models and Methods for Life Time Data*, New York: John Wiley.
Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: John Wiley.
——— (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: John Wiley.
Meeker, W. Q. (1987), "Limited Failure Population Life Tests: Application to Integrated Circuit Reliability," *Technometrics*, 29, 51–65.
Nelson, W. (1990), *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses*, New York: John Wiley.
Ostrouchov, G., and Meeker, W. Q. (1988), "Accuracy of Approximate Confidence Bounds Computed from Interval Censored Weibull and Lognormal Data," *Journal of Statistical Computation and Simulation*, 29, 43–76.
Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley.
Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: John Wiley.
Seber, G. A. F., and Wild, C. J. (1989), *Nonlinear Regression*, New York: John Wiley.
Severini, T. A. (1991), "On the Relationship Between Bayesian and Non-Bayesian Interval Estimates," *Journal of the Royal Statistical Society*, Ser. B, 53, 611–618.
Sprott, D. A. (1973), "Normal Likelihoods and Their Relation to Large Sample Theory of Estimation," *Biometrika*, 60, 457–465.
——— (1975), "Application of Maximum Likelihood Methods to Finite Samples," *Sankhyā*, 37, 259–270.
——— (1980), "Maximum Likelihood in Small Samples: Estimation in the Presence of Nuisance Parameters," *Biometrika*, 67, 515–523.
Stuart, A., and Ord, K. (1991), *Kendall's Advanced Theory of Statistics, Volume 2: Classical Inference and Relationship* (5th ed.), Oxford, U.K.: Oxford University Press.
Vander Wiel, S. A., and Meeker, W. Q. (1990), "Accuracy of Approximate Confidence Bounds Using Censored Weibull Regression Data from Accelerated Life Tests," *IEEE Transactions on Reliability*, 39, 346–351.
Venzon, D. J., and Moolgavkar, S. H. (1988), "A Method of Computing Profile-Likelihood-Based Confidence Intervals," *Applied Statistics*, 37, 87–94.

Graphical Interpretation of Variance Inflation Factors

Robert A. STINE

A dynamic graphical display is proposed for uniting partial regression and partial residual plots. This animated display helps students understand multicollinearity and interpret the variance inflation factor. The variance inflation factor is presented as the square of the ratio of t-statistics associated with the partial regression and partial residual plots. Examples using two small data sets illustrate this approach.

KEY WORDS: Collinearity; Interactive plots; Regression diagnostics

1. INTRODUCTION

This article focuses on the connection between the variance inflation factor (VIF) and two diagnostic plots for least squares regression, partial regression plots, and partial residual plots (added-variable plots and component-plus-residual plots). To help students master regression diagnostics, I have found it useful to point out explicitly the connections among them. Introductions to regression diagnostics at the level of Chatterjee and Price (1991) or Fox (1991) offer the student a variety of numerical and graphical diagnostics for judging the adequacy of a regression model. There are diagnostics for specification error, outliers, multicollinearity, nonlinearity, heteroscedasticity, and other faults. Rather than present each diagnostic individually, I find it useful to describe the connections among them, much as one needs to do in presenting the various types of random variables in an introductory course.

The presentation offered here is relatively elementary. The level is appropriate for students who do not know linear algebra, and I have found it useful in more advanced courses as well. The presentation relies upon imbedding the three diagnostics in a single dynamic plot. At one extreme of a slider control, this plot is the partial residual plot, which shows none of the effects of collinearity. As the control moves to the other extreme, it becomes the partial regression plot, which conveys the effects of multicollinearity. The plot dynamically updates its coordinates to suggest the effects of intermediate levels of multicollinearity.

2. THE DIAGNOSTICS

The VIF measures how much multicollinearity has increased the variance of a slope estimate. Suppose that we write the full-rank regression model for n independent observations as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\text{var}(\epsilon_i) = \sigma^2$. In vector form, the model is $Y = X\beta + \epsilon$ where X is the $n \times (k+1)$ matrix with columns $X_0, X_1, \ldots, X_k$ and X_0 is a column vector of 1s. The name of this diagnostic arises from writing the variance of the least squares estimator $\widehat{\beta}_j$ $(j = 1, \ldots, k)$ as (e.g., Belsley 1991, sec. 2.3)

$$\text{var}(\widehat{\beta}_j) = \sigma^2 (X'X)^{-1}_{jj} = \frac{\sigma^2}{SS_j} \text{VIF}_j,$$

Robert A. Stine is Associate Professor, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6302.