Asymptotically Optimal Prediction

David J. Olive*

Southern Illinois University

June 8, 2011

Abstract

This paper presents asymptotically optimal prediction intervals and prediction regions. The prediction intervals are for a future response Y_f given a vector $\boldsymbol{x}_f$ of predictors when the regression model has the form $Y_i = m(\boldsymbol{x}_i) + e_i$ where m is a function of $\boldsymbol{x}_i$ and the errors e_i are iid. The prediction regions are for a future vector of measurements $\boldsymbol{x}_f$ from an elliptically contoured distribution. The techniques perform well for moderate sample sizes as well as asymptotically.

KEY WORDS: nonlinear regression; prediction intervals; prediction regions; regression; generalized additive models.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu.

1 INTRODUCTION

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $p \times 1$ vector of predictors $\boldsymbol{x}$. An important regression model is

$$Y_i = m(\boldsymbol{x}_i) + e_i \tag{1}$$

for $i = 1, ..., n$ where m is a function of $\boldsymbol{x}_i$ and the errors e_i are continuous and iid. Many of the most important regression models have this form, including the multiple linear regression model and many time series, nonlinear, nonparametric and semiparametric models. If $\hat{m}$ is an estimator of m, then the ith residual is $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$.

Olive (2007) showed how to form asymptotically optimal prediction intervals for such models when the errors are iid from a continuous unimodal distribution. A problem with these intervals is that for many regression models and estimators, large n is needed for the intervals to perform well. Prediction intervals derived for multiple linear regression using least squares (OLS) did perform well. This paper derives asymptotically optimal prediction intervals that perform well for many models for moderate n.

A large sample $100(1-\alpha)\%$ prediction interval (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \xrightarrow{P} 1 - \alpha$ as the sample size $n \to \infty$. Following Olive (2007), let ξ_α be the α percentile of the error e, i.e., $P(e \leq \xi_\alpha) = \alpha$. Let $\hat{\xi}_\alpha$ be the sample α percentile of the residuals. Consider predicting a future observation Y_f given a vector of predictors $\boldsymbol{x}_f$ where $(Y_f, \boldsymbol{x}_f)$ comes from the same population as the past data $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Let $1 - \alpha_2 - \alpha_1 = 1 - \alpha$ with $0 < \alpha < 1$ and $\alpha_1 < 1 - \alpha_2$ where $0 < \alpha_i < 1$. Then $P[Y_f \in (m(\boldsymbol{x}_f) + \xi_{\alpha_1}, m(\boldsymbol{x}_f) + \xi_{1-\alpha_2})] = 1 - \alpha$.

Assume that $\hat{m}$ is consistent: $\hat{m}(\boldsymbol{x}) \xrightarrow{P} m(\boldsymbol{x})$ as $n \to \infty$. Then $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) \xrightarrow{P}$

$Y_i - m(\boldsymbol{x}_i) = e_i$ and, under "mild" regularity conditions, $\hat{\xi}_\alpha \xrightarrow{P} \xi_\alpha$. If $a_n \xrightarrow{P} 1$ and $b_n \xrightarrow{P} 1$, then

$$(\hat{L}_n, \hat{U}_n) = (\hat{m}(\boldsymbol{x}_f) + a_n \hat{\xi}_{\alpha_1}, \hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{1-\alpha_2}) \tag{2}$$

is a large sample $100(1-\alpha)\%$ PI for Y_f.

According to regression folklore, the percentiles of the residuals are consistent estimators, $\hat{\xi}_\alpha \xrightarrow{P} \xi_\alpha$, under "mild" regularity conditions, and this consistency is the basis for using QQ plots. The folklore is true for linear models: sufficient conditions are $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ and the $\boldsymbol{x}_i$ are bounded in probability. See Olive and Hawkins (2003), Welsh (1986) and Rousseeuw and Leroy (1987, p. 128).

As an example, consider the multiple linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown iid zero mean errors e_i with variance σ^2. Let the "leverage" $h_f = \boldsymbol{x}_f^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_f$ and use the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}_{OLS}$ to find $\hat{Y}_f = \boldsymbol{x}_f^T \hat{\boldsymbol{\beta}}_{OLS}$. Let $\hat{\xi}_\alpha$ be the sample quantile of the residuals. Following Olive (2007), let

$$a_n = b_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n}{n-p}} \sqrt{(1+h_f)}. \tag{3}$$

Then a large sample semiparametric $100(1-\alpha)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n \hat{\xi}_{\alpha/2}, \hat{Y}_f + a_n \hat{\xi}_{1-\alpha/2}). \tag{4}$$

A PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. The PI (4) is asymptotically optimal on a large class of unimodal continuous symmetric error distributions. For more general distributions, an asymptotically optimal PI can be created by applying the shorth(c) estimator to the residuals where $c = \lceil n(1-\alpha) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. See Grübel (1988). That is, let $r_{(1)}, ..., r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\alpha_1}, \tilde{\xi}_{1-\alpha_2})$ correspond to the interval with the smallest length. Following Olive (2007), a 100 $(1-\alpha)$% PI for Y_f is

$$(\hat{Y}_f + a_n \tilde{\xi}_{\alpha_1}, \hat{Y}_f + a_n \tilde{\xi}_{1-\alpha_2}) \tag{5}$$

where a_n is given by (3). This prediction interval performs well for moderate n for multiple linear regression and least squares.

A problem with prediction intervals is choosing a_n and b_n so that the intervals have short length and coverage close to or higher than the nominal coverage for a wide variety of regression models when n is moderate. Section 2 shows how to modify (4) and (5) to achieve these goals while Section 3 covers prediction regions for a future vector of measurements $\boldsymbol{x}_f$.

2 Asymptotically Optimal Prediction Intervals

The technique used to produce asymptotically optimal PIs that perform well for moderate samples is simple. Find $\hat{Y}_f$ and the residuals from the regression model. For a wide range of regression models, extrapolation occurs if $h_f > 2p/n$: if $\boldsymbol{x}_f$ is too far from the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, then the model may not hold and prediction can be arbitrarily bad. This result suggests replacing (3) by

$$a_n = b_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n+2p}{n-p}}. \tag{6}$$

Let $p_n = \min(1-\alpha+0.05, 1-\alpha+p/n)$ for $\alpha > 0.1$ and

$$p_n = \min(1-\alpha/2, 1-\alpha+10\alpha p/n), \quad \text{otherwise.} \tag{7}$$

Set $q_n = p_n$. If $1-\alpha < 0.999$ and $q_n < 1-\alpha+0.001$, set $q_n = 1-\alpha$.

Let $\alpha_n = 1 - q_n$. Then

$$(\hat{L}_n, \hat{U}_n) = (\hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{\alpha_n/2}, \hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{1-\alpha_n/2}) \tag{8}$$

is a large sample $100(1-\alpha)\%$ PI for Y_f that is similar to (2) and (4).

Let $c = \lceil n q_n \rceil$. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\alpha_1}, \tilde{\xi}_{1-\alpha_2})$ correspond to the interval with the smallest length. Then the asymptotically optimal $100\ (1-\alpha)\%$ large sample PI for Y_f is

$$(\hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{\alpha_1}, \hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{1-\alpha_2}), \tag{9}$$

and is similar to (5).

To see that the PI (9) is asymptotically optimal, assume that the sample percentiles of the residuals converge to the population percentiles of the iid unimodal errors: $\hat{\xi}_\alpha \xrightarrow{P} \xi_\alpha$. Also assume that the population shorth $(\xi_{\alpha_1}, \xi_{1-\alpha_2})$ is unique and has length L. Since $b_n \to 1$, $\hat{m}(\boldsymbol{x}_f) \xrightarrow{P} m(\boldsymbol{x}_f)$, and $q_n = 1-\alpha$ for large enough n, it is enough to show that the shorth of the residuals converges to the population shorth of the e_i: $(\tilde{\xi}_{\alpha_1}, \tilde{\xi}_{1-\alpha_2}) \xrightarrow{P} (\xi_{\alpha_1}, \xi_{1-\alpha_2})$. Let L_n be the length of $(\tilde{\xi}_{\alpha_1}, \tilde{\xi}_{1-\alpha_2})$. Let $0 < \delta < 1$ and $0 < \epsilon < L$ be arbitrary. Assume n is large enough so that $q_n = 1-\alpha$. Then $P(L_n > L + \epsilon) \to 0$ since $(\hat{\xi}_{\alpha_1}, \hat{\xi}_{1-\alpha_2})$ covers $100\ (1-\alpha)\%$ of the data and $L_n = \tilde{\xi}_{1-\alpha_2} - \tilde{\xi}_{\alpha_1} \le \hat{\xi}_{1-\alpha_2} - \hat{\xi}_{\alpha_1} \xrightarrow{P} L$ as $n \to \infty$ since the sample percentiles are consistent and the shorth is the smallest interval covering $100\ (1-\alpha)\%$ of the data. If $P(L_n < L - \epsilon) > \delta$ eventually, then the shorth is an interval covering 100 $(1-\alpha)$% of the cases that is shorter than the population shorth with positive probability δ. Hence at least one of $\hat{\xi}_{1-\alpha_2}$ or $\hat{\xi}_{\alpha_1}$ would not converge, a contradiction. Since ϵ and δ were arbitrary, $L_n \xrightarrow{P} L$. If $P(\tilde{\xi}_{\alpha_1} < \xi_{\alpha_1} - \epsilon) > \delta$ eventually, then $P(\tilde{\xi}_{1-\alpha_2} < \xi_{1-\alpha_2} - \epsilon/2) > \delta$ eventually since $L_n = \tilde{\xi}_{1-\alpha_2} - \tilde{\xi}_{\alpha_1} \xrightarrow{P} L = \xi_{1-\alpha_2} - \xi_{\alpha_1}$. But such an interval (of length going to L in probability with left endpoint less than $\xi_{\alpha_1} - \epsilon$ and right endpoint less than $\xi_{1-\alpha_2} - \epsilon/2$) contains more than $100(1-\alpha)$% of the cases with probability going to one since the population shorth is unique shortest interval covering $100(1-\alpha)$% of the mass. Hence there is an interval covering $100(1-\alpha)$% of the cases that is shorter than the shorth, with probability going to one, a contradiction. The case $P(\tilde{\xi}_{\alpha_1} > \xi_{\alpha_1} + \epsilon) > \delta$ can be handled similarly. Since ϵ and δ were arbitrary, $\tilde{\xi}_{\alpha_1} \xrightarrow{P} \xi_{\alpha_1}$. The proof that $\tilde{\xi}_{1-\alpha_2} \xrightarrow{P} \xi_{1-\alpha_2}$ is similar.

The above results show that the shorth of the residuals behaves well when the sample percentiles are consistent. We conjecture that for the uniform distribution, where the populuation shorth is not unique, the PI (9) is asymptotically optimal under mild regularity conditions, but the endpoints of the PI fail to converge to population percentiles.

For asymptotic optimality, can not have extrapolation. Also, even if the coverage converges to the nominal coverage, the length of the PI need not be asymptotically shortest unless the highest $1-\alpha$ density region of the probability density function of the iid errors is an interval. The highest density region is an interval for unimodal distributions, but need not be an interval for multimodal distributions for fixed α. Also see Cai, Tian, Solomon and Wei (2008).

Notice that the technique computes a PI for coverage $q_n \geq 1-\alpha$ which converges to the nominal coverage $1-\alpha$ as $n \to \infty$. Suppose $n \leq 20p$. Then the nominal 95% PI uses

$q_n = 0.975$ while the nominal 50% PI uses $q_n = 0.55$. Prediction distributions depend both on the error distribution and on the variability of the estimator $\hat{m}$. This variability is typically unknown but converges to 0 as $n \to \infty$. Also, residuals tend to underestimate the errors for small n. For small n, ignoring estimator variability and using $q_n = 1 - \alpha$ resulted in undercoverage as high as $\min(0.05, \alpha/2)$. Letting the "coverage" q_n decrease to the nominal coverage $1 - \alpha$ inflates the length of the PI for small n, compensating for the unknown variability of $\hat{m}$.

The geometry of the "asymptotically optimal prediction region" is simple. The region is the area between two parallel lines with unit slope. Consider a plot of $m(\boldsymbol{x}_i)$ versus Y_i on the vertical axis. The identity line with zero intercept and unit slope is $E(Y_i) = m(\boldsymbol{x}_i)$. Let (L_i, U_i) be the asymptotically optimal 95% prediction interval containing $m(\boldsymbol{x}_i)$. For example, if the errors are iid $N(0, \sigma^2)$, then $Y_i | m(\boldsymbol{x}_i) \sim N(m(\boldsymbol{x}_i), \sigma^2)$, and $(L_i, U_i) = (m(\boldsymbol{x}_i) - 1.96\sigma, m(\boldsymbol{x}_i) + 1.96\sigma)$. Then the upper line has unit slope and passes through $(m(\boldsymbol{x}_i), U_i)$ while the lower line has unit slope and passes through $(m(\boldsymbol{x}_i), L_i)$.

A response plot of $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$ versus Y_i has identity line $\hat{E}(Y_i) = \hat{m}(\boldsymbol{x}_i)$. The region corresponding to pointwise prediction intervals is between two lines with unit slope passing through the points $(\hat{m}(\boldsymbol{x}_i), \hat{U}_i)$ and $(\hat{m}(\boldsymbol{x}_i), \hat{L}_i)$, respectively, where $(\hat{L}_i, \hat{U}_i)$ is the asymptotically optimal prediction interval (9) for Y_f if $\boldsymbol{x}_f = \boldsymbol{x}_i$. Olive (2007) suggested a similar plot for PIs (4) and (5), but the region was not between two parallel lines since the length of PIs (4) and (5) depends on h_f.

For the multiple linear regression model, expect the points in the response plot to scatter in an evenly populated band for $n > 5p$. Other regression models, such as generalized additive models, may need a much larger sample size n.

Example 1. Chambers and Hastie (1993, pp. 251, 516) examine an environmental study that measured the four variables Y = ozone concentration, solar radiation, temperature, and wind speed for $n = 111$ consecutive days. Figure 1 shows the response plot with the pointwise large sample 95% PI bands for the generalized additive model. Here $\hat{m}(\boldsymbol{x})$ = estimated additive predictor (EAP). Note that the plotted points scatter about the identity line in a roughly evenly populated band, and that 3 of the 111 PIs (9) corresponding to the observed data do not contain Y.

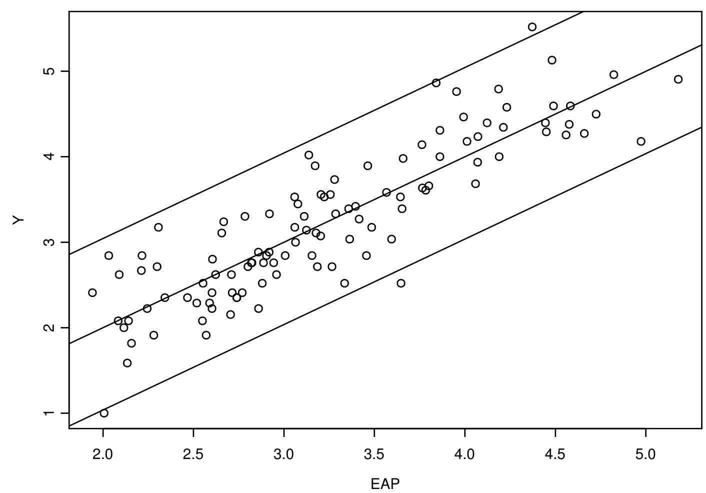

Figure 1: Pointwise Prediction Interval Bands for Ozone Data

A small simulation study compares the PI lengths and coverages for sample sizes $n = 50, 100$ and 1000 for PIs (8) and (9). Values for PI (8) were denoted by scov and slen while values for PI (9) were denoted by ocov and olen. The five error distributions in the simulation were 1) N(0,1), 2) t_3, 3) exponential(1) -1, 4) uniform$(-1, 1)$ and 5) $0.9N(0,1) + 0.1N(0,100)$. The value $n = \infty$ gives the asymptotic coverages and lengths and does not depend on the model. So these values are same for multiple linear and nonlinear regression as well as generalized additive models.

Software for the simulations is described in the conclusions section. The multiple linear regression model with $E(Y_i) = 1 + x_{i1} + \cdots + x_{i7}$ was used. The vectors $(x_1, ..., x_7)^T$ were iid $N_7(\mathbf{0}, \boldsymbol{I}_7)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. For nonlinear regression $Y_i = m(\boldsymbol{x}_i) + e_i$, $E(Y_i) = m(\boldsymbol{x}_i) = \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \beta_4 x_{i2}^2 + \beta_5 x_{i3} + \beta_6 x_{i3}^2$. For the generalized additive model, the additive predictor $m(\boldsymbol{x}_i) = \alpha + \sum_{j=1}^{3} S_j(x_{ij})$. Both the nonlinear regression and generalized additive model had the same mean function $m(\boldsymbol{x}_i) = x_{i1} + x_{i1}^2$. Thus $\boldsymbol{\beta} = (1, 1, 0, 0, 0, 0)^T$, $\alpha = 0$, $S_1(x_{i1}) = x_{i1} + x_{i1}^2$, $S_2(x_{i2}) = 0$ and $S_3(x_{i3}) = 0$. For these two models, the vectors $(x_1, x_2, x_3)^T$ were iid $N_3(\mathbf{0}, \boldsymbol{I}_3)$.

The Olive (2007) PIs (4) and (5) are tailored for multiple linear regression but are liberal (too short) for moderate n for many other techniques. The new PIs (8) and (9) are meant to have coverage near or higher than the nominal coverage for moderate n and for a wide variety of techniques and are longer than PIs (4) and (5). For multiple linear regression, the new PIs (8) and (9) were conservative (too long with roughly 98% coverage for the 95% PI and 70% or 60% coverage for the 50% PI) for $n = 50$ and 100 compared to (4) and (5) for least squares, least absolute deviations L_1 and an M-estimator using the *Splus* functions `l1fit` and `rreg`. See MathSoft (1999, pp. 293-295.)

The PIs (8) and (9) for nonlinear regression and generalized additive models appear to have coverage near the nominal values in the simulations. For $n = 50$ and 100, the PIs for nonlinear regression were usually roughly 10% longer than those for generalized additive models.

The PIs for the generalized additive model were computed using the R function `gam`. See Hastie and Tibshirani (1990) and Wood (2006). The PI (8) is not asymptotically

Table 1: PIs for Generalized Additive Models

error type	n	95% slen	PI olen	95% scov	PI ocov	50% slen	PI olen	50% scov	PI ocov
1	50	5.126	4.998	0.959	0.950	1.862	1.674	0.596	0.520
1	100	4.691	4.515	0.968	0.957	1.662	1.528	0.570	0.516
1	1000	3.994	3.944	0.954	0.949	1.379	1.351	0.514	0.505
1	∞	3.920	3.920	0.95	0.950	1.349	1.349	0.50	0.50
2	50	9.444	8.630	0.951	0.943	2.385	2.153	0.576	0.512
2	100	8.245	7.596	0.962	0.954	2.042	1.878	0.577	0.532
2	1000	6.523	6.388	0.950	0.946	1.584	1.553	0.499	0.489
2	∞	6.365	6.365	0.950	0.950	1.530	1.530	0.50	0.50
3	50	5.186	4.823	0.958	0.948	1.573	1.275	0.611	0.526
3	100	4.677	4.156	0.967	0.955	1.382	1.063	0.603	0.533
3	1000	3.771	3.227	0.954	0.952	1.112	0.774	0.509	0.512
3	∞	3.664	2.996	0.950	0.950	1.099	0.693	0.50	0.50
4	50	2.634	2.598	0.961	0.958	1.237	1.087	0.593	0.506
4	100	2.318	2.272	0.972	0.968	1.155	1.028	0.561	0.480
4	1000	1.936	1.926	0.959	0.954	1.014	0.969	0.499	0.486
4	∞	1.900	1.900	0.950	0.950	1.00	1.00	0.50	0.50
5	50	19.689	17.747	0.944	0.935	2.976	2.693	0.608	0.548
5	100	18.754	16.230	0.955	0.946	2.352	2.164	0.580	0.534
5	1000	13.855	12.930	0.946	0.943	1.602	1.569	0.510	0.504
5	∞	13.490	13.490	0.950	0.950	1.507	1.507	0.50	0.50

optimal with error type 3.

The simulation used 5000 runs and gave the proportion $\hat{p}$ of runs where Y_f fell within the nominal $100(1-\alpha)\%$ PI. The count $m\hat{p}$ has a binomial$(m = 5000, p = 1 - \delta_n)$ distribution where $1 - \delta_n$ converges to the asymptotic coverage $(1 - \delta)$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/5000} = 0.0031$ and 0.0071 for $p = 0.05$ and 0.5, respectively. Hence an observed coverage $\hat{p} \in (.941, .959)$ for 95% and $\hat{p} \in (.479, .521)$ for 50% PIs suggests that there is no reason to doubt that the PI has the nominal coverage.

Table 1 shows that for $n = 1000$, the coverages and lengths are near the asymptotic $n = \infty$ values. For the 95% PI (9), the coverages were in or near (.94,.96) while the 50% PI (9) was sometimes slightly conservative. The coverage for the 50% PI (8) was near 60% for $n = 50$. PI (9) is recommended since its asymptotic optimality does not depend on the symmetry of the error distribution.

3 Prediction Regions

Asymptotically optimal prediction regions use ideas similar to those in the previous section. Some notation is needed. Let the ith case $\boldsymbol{x}_i$ be a $p \times 1$ random vector, and suppose the n cases are collected in an $n \times p$ matrix $\boldsymbol{X}$ with rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$.

The classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ of multivariate location and dispersion is the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})(\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})^\mathrm{T}. \tag{10}$$

Some important joint distributions for $\boldsymbol{x}$ are completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$.

An important model is the elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with probability density function $f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})]$ where $k_p > 0$ is some constant and g is some known function. The multivariate normal (MVN) $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution is a special case.

Let the $p \times 1$ column vector $T(\boldsymbol{X})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{X})$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X})) = (\boldsymbol{x}_i - T(\boldsymbol{X}))^T \boldsymbol{C}^{-1}(\boldsymbol{X})(\boldsymbol{x}_i - T(\boldsymbol{X})) \tag{11}$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{X})$ is $D_i(T(\boldsymbol{X}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. Following Johnson (1987, pp. 107-108), the population squared Mahalanobis distance

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{12}$$

and for elliptically contoured distributions, U has probability density function (pdf)

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{13}$$

The volume of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1}(\boldsymbol{z} - \overline{\boldsymbol{x}}) \le h^2\} \text{ is equal to } \frac{2\pi^{\mathrm{p}/2}}{\mathrm{p}\Gamma(\mathrm{p}/2)} \mathrm{h}^{\mathrm{p}} \sqrt{\det(\boldsymbol{S})}, \tag{14}$$

see Johnson and Wichern (1988, pp. 103-104).

Note that if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\ \boldsymbol{\Sigma})$, then

$$D^2(T, \boldsymbol{C}) = (\boldsymbol{x}-T)^T \boldsymbol{C}^{-1}(\boldsymbol{x}-T) = (\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-T)^T [\boldsymbol{C}^{-1} - d^{-1}\boldsymbol{\Sigma}^{-1} + d^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-T)$$

$$= d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-1/2}).$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_p$.

Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\ \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \le h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \le h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \le h\} \tag{15}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)} \tag{16}$$

by (14). A future observation (random vector) $\boldsymbol{x}_f$ is in region (15) if $D_{\boldsymbol{x}_f} \le h$.

A large sample $(1-\alpha)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \xrightarrow{P} 1 - \alpha$. Let $p_n = \min(1 - \alpha + 0.05, 1 - \alpha + p/n)$ for $\alpha > 0.1$ and

$$p_n = \min(1 - \alpha/2, 1 - \alpha + 10\alpha p/n), \quad \text{otherwise.}$$

$$\text{Set} \ \ \mathrm{q_n} = \mathrm{p_n}. \ \ \text{If} \ \ \ 1 - \alpha < 0.999 \ \ \text{and} \ \ \mathrm{q_n} < 1 - \alpha + 0.001, \ \ \text{set} \ \ \mathrm{q_n} = 1 - \alpha. \tag{17}$$

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then (15) is a large sample $(1-\alpha)100\%$ prediction region if $h = D_{(up)}$ where $D_{(up)}$ is the q_nth sample quantile of the D_i. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid from an elliptically contoured distribution, then region (15) is asymptotically optimal in that its volume converges in probability to the volume of the minimum volume covering region $\{\boldsymbol{z} : (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \le u_{1-\alpha}\}$ where $P(U \le u_{1-\alpha}) = 1 - \alpha$ and U has pdf given by (13). The classical parametric multivariate normal large sample prediction region uses $D_{\boldsymbol{x}_f} \equiv MD_{\boldsymbol{x}_f} \le \sqrt{\chi^2_{p,1-\alpha}}$.

Notice that $p_n = 1-\alpha/2$ or $p_n = 1-\alpha+0.05$ for $n \leq 20p$ and $p_n \to 1-\alpha$ as $n \to \infty$. If $q_n = 1-\alpha$, then (15) is a large sample prediction region, but taking q_n given by (17) improves the finite sample performance of the region.

The Olive and Hawkins (2010) RMVN estimator $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ is an easily computed $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ under regularity conditions (E1) that include a large class of elliptically contoured distributions, and $c = 1$ for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. The RMVN estimator also give useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data even when certain types of outliers are present.

Three new prediction regions will be considered. The nonparametric region uses the classical estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = D_{(up)}$. The semiparametric region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(up)}$. The parametric MVN region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{p,q}$ where $P(W \leq \chi^2_{p,q}) = q$ if $W \sim \chi^2_p$. All three regions are asymptotically optimal for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributions with nonsingular $\boldsymbol{\Sigma}$. The first two regions are asymptotically optimal under (E1). For distributions with nonsingular covariance matrix $c_X\boldsymbol{\Sigma}$, the nonparametric region is a large sample $(1-\alpha)100\%$ prediction region, but regions with smaller volume may exist.

Example 2. Buxton (1920) gives five measurements on 87 men: *height*, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. The 5 outliers have *heights* that were recorded to be about 19mm and head lengths recorded as the heights. The DD plot of the classical Mahalanobis distances MD versus the RMVN distances RD can be used to visualize the prediction regions. Figure 2 shows the DD plot where points to the left of the vertical line are in the nonparametric large sample 90% prediction region. Points below the horizontal line are in the semiparametric region. The horizontal line

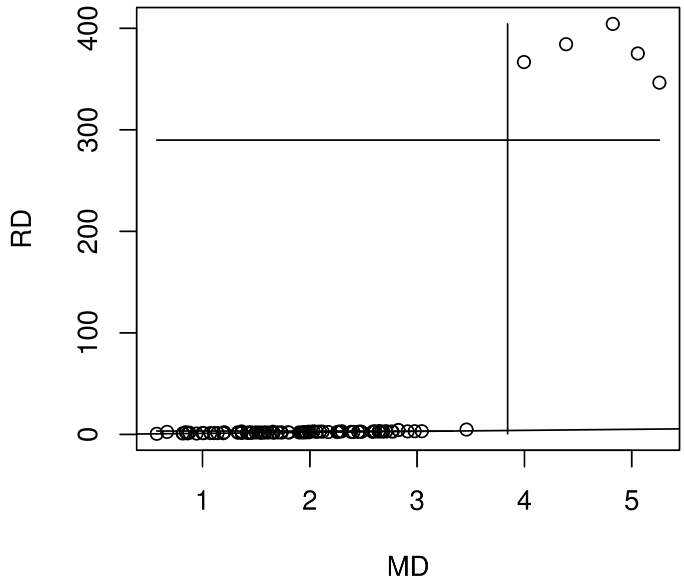

Figure 2: Prediction Regions for Buxton Data

at $RD = 3.33$ corresponding to the parametric MVN 90% region is obscured by the identity line. This region contains 78 of the cases. Since $n = 87$, the nonparametric and semiparametric regions used the 95th quantile. Since there were 5 outliers, this quantile was a linear combination of the largest clean distance and the smallest outlier distance. The semiparametric 90% region blows up unless the outlier proportion is small.

Figure 3 shows the DD plot and 3 prediction regions after the 5 outliers were removed. The classical and robust distances cluster about the identity line and the three regions are similar, with the parametric MVN region cutoff again at 3.33, slightly below the semiparametric region cutoff of 3.44.

Boente and Farall (2008) discuss a similar problem, but their theory uses the Stahel Donoho estimator which currently can only be computed for $p \leq 2$. Stahel Donoho algo-

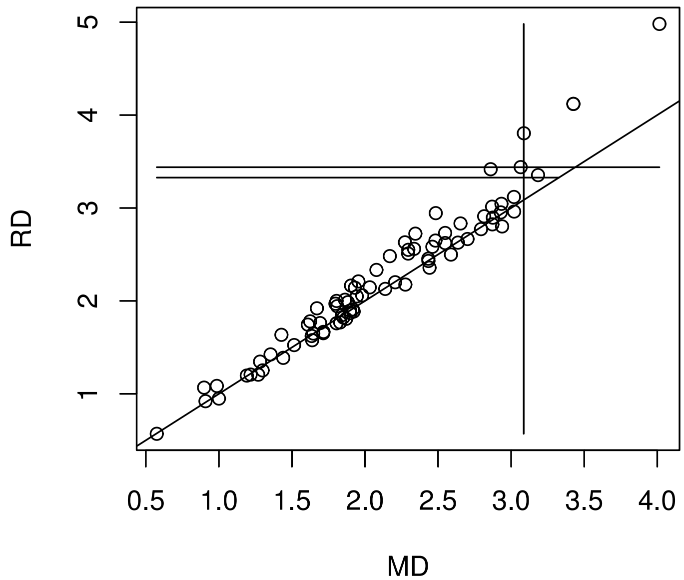

Figure 3: Prediction Regions for Buxton Data without Outliers

rithms use randomly chosen directions and are inconsistent if randomly chosen directions result in inconsistent estimators. Di Bucchianico, Einmahl and Mushkudiani (2001) use the minimum volume ellipsoid (MVE) estimator to cover m out of n cases to produce MVE prediction regions, but the technique can only be used on tiny data sets.

Simulations for the prediction regions used $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ where $\boldsymbol{A} = diag(\sqrt{1}, ..., \sqrt{p})$, $\boldsymbol{w} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{w} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{w} \sim MVT_p(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy(0,1). All simulations used 5000 runs and $\alpha = 0.1$.

For large n, the semiparametric and nonparametric regions are likely to have coverage near 0.90 because the coverage on the training sample is slightly larger than 0.9 and $\boldsymbol{x}_f$ comes from the same distribution as the $\boldsymbol{x}_i$. For $n = 10p$ and $2 \leq p \leq 40$, the

Table 2: Coverages for 90% Prediction Regions

$\boldsymbol{w}$ dist	n	p	ncov	scov	mcov	voln	volm
MVN	600	30	0.906	0.919	0.902	0.503	0.512
MVN	1500	30	0.899	0.899	0.900	1.014	1.027
LN	1000	10	0.903	0.906	0.567	0.659	0+
MVT(1)	1000	10	0.914	0.914	0.541	22634.3	0+

semiparametric region had coverage near 0.9. The ratio of the volumes

$$\frac{h_i^p \sqrt{det(\boldsymbol{C}_i)}}{h_2^p \sqrt{det(\boldsymbol{C}_2)}}$$

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric region, and $i = 3$ was the parametric MVN region. The volume ratio converges in probability to 1 for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data, and the ratio converges to 1 for $i = 1$ if (E1) holds. The parametric MVN region often had coverage much lower than 0.9 with a volume ratio near 0, recorded as 0+. The volume ratio tends to be tiny when the coverage is much less than the nominal value 0.9. For $10p \leq n \leq 20p$, the nonparametric region often had good coverage and volume ratio near 0.5.

Simulations and Table 2 suggest that for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data, the coverages (ncov, scov and mcov for prediction regions $i = 1$, 2, and 3, respectively) for the 3 regions are near 90% for $n = 20p$ and that the volume ratios voln and volm are near 1 for $n = 50p$. With fewer than 5000 runs, this result held for $2 \leq p \leq 80$. For the non–elliptically contoured LN data, the nonparametric region had voln well under 1, but the volume ratio blew up for $\boldsymbol{w} \sim MVT_p(1)$.

4 Conclusions

Formulas (7) and (17) were used to inflate the asymptotically optimal prediction intervals (8) and (9) and the asymptotically optimal nonparametric and semiparametric prediction regions so that the coverage was good for a wide variety of models when n is moderate.

The Olive (2007) PIs (4) and (5) are tailored for multiple linear regression but are too short for many other techniques for moderate n. PIs (8) and (9) are generally longer than PIs (4) and (5) and have coverage near or higher than the nominal value for many techniques even for moderate n, say $n > 10$(model degrees of freedom). PIs (8) and (9) are quite conservative for multiple linear regression for moderate n. These PIs are useful since the error distribution does not need to be known.

The new nonparametric and semiparametric prediction regions appear to have good coverage for $n > 10p$ and may be the first easily computed prediction regions that are effective when the underlying multivariate distribution is unknown. We conjecture that the semiparametric region is also a nonparametric large sample prediction region, but currently only have theory for elliptically contoured distributions. The plotted points in the DD plot follow the identity line with zero intercept and unit slope for multivariate normal data, but follow some other line through the origin if the data is from some other elliptically contoured distribution satisfying (E1). The DD plot is also useful for detecting outliers. See Olive and Hawkins (2010).

Simulations were done in *Splus* and *R*. See R Development Core Team (2008). The Buxton data and programs in the collection of functions *rpack.txt* are available at (www. math.siu.edu/olive/ol-bookp.htm). For multiple linear regression, `pisim` simulates PIs

(4) and (5) while the *Splus* function `pisim4` simulates PIs (8) and (9) using OLS, L_1 and M-estimators. The function `pisim3` was used to create Table 1 while `pisim5` uses `nls` to simulate PIs for nonlinear regression. Care is needed when using `pisim5` since for some versions of *R/Splus*, the `nls` function will fail to converge for some runs. Using nruns = 500 is less likely to cause an error than nruns=5000. The function *predsim* was used for Table 2. The function *ddplot4* was used to produce Figures 2 and 3. The function *covrmvn* computes the RMVN estimators.

Acknowledgements

This work was supported by the National Science Foundation under the grant DMS 0600933.

References

Boente, G., Farall, A., 2008. Robust multivariate tolerance regions: influence function and monte carlo study. Technometrics 50, 487-500.

Buxton, L.H.D., 1920. The anthropology of Cyprus. J R Anthropological Institute of Great Britain and Ireland 50, 183-235.

Cai, T., Tian, L., Solomon, S.D., Wei, L.J., 2008. Predicting future responses based on possibly misspecified working models. Biometrika 95, 75-92.

Chambers, J.M., Hastie, T.J., (eds.) 1993. Statistical Models in S, Chapman & Hall, New York, NY.

Di Bucchianico, A., Einmahl, J.H.J., Mushkudiani, N.A., 2001. Smallest nonparametric tolerance regions. Ann. Statist. 29, 1320-1343.

Grübel, R., 1988. The length of the shorth. Ann. Statist. 16, 619-628.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London, UK.

Johnson, M.E., 1987. Multivariate Statistical Simulation. Wiley, New York, NY.

Johnson, R.A., Wichern, D.W., 1988. Applied Multivariate Statistical Analysis. 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

MathSoft, 1999. S-Plus 2000 Guide to Statistics. Vol. 1, Data Analysis Products Division, MathSoft, Seattle, WA.

Olive, D.J., 2007. Prediction intervals for regression models. Computat. Statist. Data Analys. 51, 3115-3122.

Olive, D.J., and Hawkins, D.M. (2003), "Robust Regression with High Coverage," *Statistics and Probability Letters,* 63, 259-266.

Olive, D.J., Hawkins, D.M., 2010. Robust multivariate location and dispersion. Preprint, see (www.math.siu.edu/olive/pphbmld.pdf).

R Development Core Team, 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* Wiley, New York, NY.

Welsh, A.H. (1986), "Bahadur Representation for Robust Scale Estimators Based on Regression Residuals," *The Annals of Statistics*, 14, 1246-1251.

Wood, S.N., 2006. Generalized Additive Models: an Introduction with R. Chapman & Hall/CRC, Boca Rotan, FL.