Statistica & Applicazioni - Vol. IX, n. 1, 2011, pp. 35-46

MEDIAN RANKED SET SAMPLING FOR POLYNOMIAL REGRESSION

Mohd T. Alodat*
Gottfried Jetschke**

SUMMARY

The purpose of this article is to study the ploynomial linear regression model under the median ranked set sampling (MRSS) scheme introduced by Muttlak (1997). If the response variable can be more easily ranked than quantified, then we use the MRSS to collect data by ranking on the response variable. We obtain estimators and confidence intervals for the polynomial regression parameters under MRSS when the errors have a symmetric distribution. We also show that the least square estimators, under MRSS, are more efficient than their SRS counterparts and give illustrating examples.

Keywords: *Median Ranked Set Sample, Polynomial Regression.*

1. INTRODUCTION

In many environmental, ecological and agricultural studies, the measurements of a variable of interest are often expensive (Särndal, Swensson and Wretman, 1992; Tillé, 2006). In such situations, it is better to use another sampling technique instead of using the classical sampling schemes such as the Simple Random Sampling (SRS). The first one who introduced a sampling scheme suiting to such situations was McIntyre (1952) who called his new technique the Ranked Set Sampling (RSS). He was aiming to estimate the mean pasture yields. As a result, he indicated, without providing any theoretical proof, that the RSS is more efficient than SRS for estimating population mean. The RSS can be described as follows.

i. Select m independent simple random samples each of size m from the population of interest and rank each set, without quantifying its units, using a visual inspection or any cheap method.
ii. Then the first ranked unit from the first set, the second ranked unit from the second set,..., the largest ranked unit from the last set are chosen for actual quantification.

As an illustration example of steps (i)-(ii), assume that we are interested in estimating the average yield of a population which consists of 500 olive trees. Collecting the yield of an olive tree is time-consuming. Instead of selecting a SRS of size m, we may select m simples random sample each of size m and rank their yields using a cheap

* Department of Statistics - Yarmouk University Irbid - JORDAN
(e-mail: ✉ alodatmts@yahoo.com).
** Institut of Ecology - University of Jena - JENA, GERMANY (e-mail: gottfried.vetschke@uni-jena.de).

method (e.g. farmer experience) without collecting their yields. From the first set we collect only the one which has (according to farmer experience) the lowest yield. From the second set we collect only the one which has the second lowest yield. We continue in this way until the yield of the tree of largest yield from the m^{th} set is collected.

It can be noticed that RSS produces a set of independent measurements $Y_1, ..., Y_m$ such that Y_i is a random observation from the i^{th} population order statistic induced by the population of interest. This procedure can be repeated r times to get a sample of size mr. McIntyre (1952) has pointed out that the sample mean of an RSS of size mr is more efficient than the mean of a SRS of the same size as an estimator to the population mean. The RSS works better than the SRS unless the visual ranking is made with errors. However, the literatures of RSS show in several cases that the RSS and its modifications can work better than the SRS even if portion of the selected samples are ranked incorrectly Chen, Bai and Sinha (2004).

The McIntyre's RSS has been studied and modified by several authors. As an excellent review for such studies and modifications we refer to Chen *et al.* (2004). One of such modifications is called the median ranked set sampling (MRSS) which was introduced by Muttlak (1997). This technique consists of selecting m simple random samples each of size m and quantifying the median of each set. For symmetric populations, he showed that the MRSS is more efficient than RSS for estimating the population mean. Muttlak (2001) studied several estimators for the parameters of the simple linear regression model under different ranked set sampling schemes.

Assuming that the X values are known constants, Muttlak (1995) used the RSS to estimate the parameters of the simple linear regression model. He showed that there is no improvement in the accuracy of parameters relative to their SRS counterparts. Assuming that the predictor variable X is random, then a slightly better improvement on the efficiency can be achieved (Al-Saleh and Samawi, 2002; Samawi and Abu-dayyeh, 2002).

In this paper, we propose to use the MRSS to study the polynomial regression model. As a result we show that the MRSS leads to improve the efficiency of parameters estimators in a comparison with their SRS counterparts.

The rest of the paper is organized as follows. In Section 2, we present the polynomial regression under MRSS. While in Section 3, we derive the least squares estimates of the regression parameters and we show that they are more efficient than their SRS counterparts. In Sections 4 and 5, we obtain tests, confidence and prediction intervals. In Section 6, we conduct a simulation study for confidence intervals lengths. In Section 7, we apply our result to an ecological study.

2. POLYNOMIAL REGRESSION UNDER MRSS

Consider the following polynomial linear regression model

$$Y = \beta_0 + \beta_1 x + ... + \beta_k x^k + \varepsilon, \tag{1}$$

where Y is the response variable, X is the predictor variable, ß's are the regression

parameters and e is a random error independent of X. Also we assume that the random error e has a symmetric distribution with zero mean and finite variance. If $g(\varepsilon;\sigma)$ and $G(\varepsilon;\sigma)$ denote the probability density function (pdf) and the cumulative distribution function (cdf) of e, then $Var(\varepsilon) = \sigma^2 D$, where

$$D = \int_{-\infty}^{\infty} \nu^2 g(\nu;1)\, d\nu.$$

Model (1) has been considered in practice by several authors. Mashario, Kazato, Kanato, Ichirow, Aya N. Jyoken and Hiroyuki (2007) considered a quadratic regression model to estimate the total plant biomass and the plant water content in a whole area of their study. They regressed the plant biomass/plant water content on the height/volume of a tree. For more details about the applications of polynomial regression models see Yarranton (1971), Mead (1971), Palmer and Hussain (1997) and Chen and Wang (2004).

Let $x_1, ..., x_r$ be distinct values of the variable X chosen for regression design. Let m be a positive integer. For each value of the x_i's, we assume that the experiment has been repeated $n = 2m - 1$ times, and $Y_{ji}$and ε_{ji}, $i = 1, ..., r$, $j = 1, ..., n$ be the corresponding response values of Y and their errors, respectively. Applying the median ranked set sampling procedure on the response values and for each $j = 1, ..., r$, let $Z_j = Median\ \{Y_{j1}, ..., Y_{jn}\}$ denote the median of $Y_{j1}, ..., Y_{jn}$.

Also let $\delta_j = Median\ \{\varepsilon_{j1}, ..., \varepsilon_{jn}\}$. Hence we may relate the values of Z_j's to x_j's via the linear model

$$Z_j = \beta_0 + \beta_1 x_j + ... + \beta_k x_j^k + \delta_j, \tag{2}$$

where $j = 1, ..., r$. The random variables $\delta_1, ..., \delta_r$ are independent and identically distributed with pdf

$$f_\delta(\nu;\sigma) = \frac{(2m-1)!}{(m-1)!^2\sigma} G(\nu;\sigma)^{m-1} G(-\nu;\sigma)^{m-1} g(\nu;\sigma), \qquad -\infty < \nu < \infty.$$

Since the pdf $g(\nu;\sigma$ is symmetric function about zero, then so is $f_\delta(\nu,\sigma)$. Therefore, it is straightforward to show that $E(\delta_j) = 0$ and $Var(\delta_j) = \sigma^2 D_m$, where

$$D_m = \frac{(2m-1)!}{(m-1)!^2} \int_{-\infty}^{\infty} \nu^2 f_\delta(\nu,\sigma).$$

Using the vector notation, we may write the above model as

$$Z = W\ + \Delta,$$

where

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_r \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & x_r & x_r^2 & \cdots & x_r^k \end{pmatrix} \quad \mathbf{\Delta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_r \end{pmatrix}, \quad \mathbf{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

3. LEAST SQUARES ESTIMATOR

Under model (2), the least squares estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_m$, is given by

$$\hat{\boldsymbol{\beta}}_m = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Z}.$$

It is easy to check that $\hat{\boldsymbol{\beta}}_m$ is an unbiased estimator for $\boldsymbol{\beta}$ with covariance matrix given by

$$Cov(\hat{\boldsymbol{\beta}}_m) = \sigma^2 D_m(\mathbf{W}'\mathbf{W})^{-1} = D_m Cov(\hat{\boldsymbol{\beta}}_1).$$

The diagonal elements of $Cov(\hat{\boldsymbol{\beta}}_m)$ represent the variances of $\hat{\beta}_0, \hat{\beta}_1, ... + \hat{\beta}_k$ respectively. To show that $\hat{\boldsymbol{\beta}}_m$is more efficient $\hat{\boldsymbol{\beta}}_1$ it suffices to show that $D_m < 1$. It is well-known in order statistics theory that this problem is not an easy task. Hence we show that $D_m < 1$ for some distributions and different values of m as presented in the Tables 1, 2, 3 and 4. It can be noticed from these tables that the conjecture $D_m < 1$ can be verified for large class of distributions.

TABLE 1. - *The values of D_m for different values of m for N*(0,1)*, Laplace*(0,1)*, Student* t(5) *and Logistic*(0,1)

n	*N(0,1)*	*Laplace(0,1)*	*Student t(5)*	*Logistic(0,1)*
2	0.4487	0.6389	0.5830	1.1290
3	0.2868	0.3512	0.3502	0.7899
4	0.2104	0.3256	0.2498	0.5676
5	0.1661	0.1751	0.1941	0.4426
6	0.1372	0.1383	0.1587	0.3626
7	0.1168	0.1138	0.1341	0.3071

TABLE 2. - *The values of D_m for different values of m for mixtures of N*(0,1)*, Laplace*(0,1)

n	*pN(0,1)+qt₅* *p*			*pN(0,1)+qLaplace(0,1)* *p*			*pN(0,1)+qLogist(0,1)* *p*		
	0.2	**0.5**	**0.8**	**0.2**	**0.5**	**0.8**	**0.2**	**0.5**	**0.8**
2	1.1411	0.9347	0.7498	0.5926	0.6084	0.6260	0.6941	0.8898	1.1192
3	0.6815	0.5386	0.4188	0.3331	0.3494	0.3502	0.4124	0.5285	0.6746
4	0.4812	0.3714	0.2834	0.2281	0.2421	0.2380	0.2925	0.3741	0.4806
5	0.3703	0.2808	0.2116	0.1719	0.1841	0.1786	0.2265	0.2890	0.3726
6	0.3001	0.2246	0.1677	0.1373	0.1481	0.1421	0.1847	0.2354	0.3041
7	0.2519	0.1865	0.1383	0.1139	0.1236	0.1176	0.1559	0.1985	0.2568

An estimator of σ^2 is

$$\hat{\sigma}_m^2 = \frac{(\mathbf{Z}-\hat{\mathbf{Z}})'(\mathbf{Z}-\hat{\mathbf{Z}})}{(r-k-1)D_m} = \frac{\mathbf{Z}(\mathbf{I}-\mathbf{P})\mathbf{Z}}{(r-k-1)D_m},$$

where $\hat{\mathbf{Z}} = \mathbf{W}\hat{\boldsymbol{\beta}}_m$ and $\mathbf{P} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}$. Since

$$\begin{aligned} E\big((\mathbf{Z}-\hat{\mathbf{Z}})'(\mathbf{Z}-\hat{\mathbf{Z}})\big) &= \sigma^2 D_m Tr(\mathbf{I}-\mathbf{P}) + \mathbf{W}(\mathbf{I}-\mathbf{P})\mathbf{W}\boldsymbol{\beta} \\ &= (r-k-1)D_m\sigma^2 \end{aligned}$$

where $Tr(.)$ denotes the trace of matrix, then $\hat{\sigma}_m^2$ is an unbiased estimator for σ^2. If $m = 1$, then the above model reduces to the simple random sampling model. The efficiency of any coefficient estimator with respect to its SRS counterpart is $\frac{1}{D_m}$. Since $D_m < 1$, then, in general, the median ranked set sampling estimator is more efficient than its SRS counterpart in general.

TABLE 3. - *The values of D_m for different values of m for mixtures of N(0,1), Laplace(0,1)*

m	*pLap(0,1) +qLogis(0,1)* p			*pLap(0,1)+qt5* p			*pLogis(0,1)+ qt5* p		
	0.2	**0.5**	**0.8**	**0.2**	**0.5**	**0.8**	**0.2**	**0.5**	**0.8**
2	0.5514	0.5083	0.4707	0.5890	0.5252	0.4747	1.0644	0.7778	0.5571
3	0.3353	0.3153	0.2975	0.3331	0.3113	0.2951	0.6425	0.4692	0.3461
4	0.2408	0.2284	0.2173	0.2281	0.2195	0.2134	0.4588	0.3359	0.2513
5	0.1877	0.1789	0.1710	0.1719	0.1687	0.1668	0.3566	0.2618	0.1973
6	0.1538	0.1471	0.1409	0.1373	0.1367	0.1368	0.2916	0.2144	0.1624
7	0.1303	0.1249	0.1199	0.1139	0.1146	0.1158	0.2466	0.1816	0.1379

The efficiency values of $\mathbf{e}_j'\hat{\boldsymbol{\beta}}_m$ with respect to $\mathbf{e}_j'\hat{\boldsymbol{\beta}}_1$, where $\mathbf{e}_j$ is the j^{th} unit vector in the standard basis of the $(k+1)$-dimensional Euclidean space, are reported in Table 4.

TABLE 4. - *Efficiency values of $\hat{\boldsymbol{\beta}}_m$ with respect to $\hat{\boldsymbol{\beta}}_1$*

M	N(0, 1)	Laplace(0,1)	Student *t* (5)	Logistic(0,1)
2	2.2287	1.5652	1.7153	0.7753
3	3.4868	2.8474	2.8555	1.2660
4	4.7529	3.0712	4.0032	1.7618
5	6.0205	5.7110	5.1520	2.2594
6	7.2886	7.2307	6.3012	2.7579
7	8.5616	8.7874	7.4571	3.2563

4. TESTING AND CONFIDENCE INTERVALS

In this section, we rely on the pivotal method (Shao, 2003) to derive confidence intervals and tests for the parameters of interest. To derive a confidence interval for a linear combination, say $\mathbf{a}'\hat{\boldsymbol{\beta}}$ of $\hat{\boldsymbol{\beta}}$, we note that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_m &= (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}(\mathbf{W}\hat{\boldsymbol{\beta}} + \Delta) \\ &= \hat{\boldsymbol{\beta}} + (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\Delta.\end{aligned}$$

Hence, for any vector $\mathbf{a}$, $\mathbf{a}'\hat{\boldsymbol{\beta}}_m - \mathbf{a}'\boldsymbol{\beta} = \mathbf{a}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\Delta$. Then the random quantity

$$\frac{\mathbf{a}'\hat{\boldsymbol{\beta}}_m - \mathbf{a}'\boldsymbol{\beta}}{\sigma} = \mathbf{a}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{e},$$

where $\mathbf{e}' = \left(\frac{\delta_{\mathbf{1}}}{\sigma}, ..., \frac{\delta_{\mathbf{r}}}{\sigma}\right)$, has a distribution free of the parameters, since δ_j's are independent and distributed as $f_\delta(\nu, 1)$. Also the quantity

$$\sqrt{\frac{Z'(I-P)Z}{\sigma^2(r-k-1)D_m}}$$

has a distribution which is free of the parameters. Hence, under MRSS, we define the following counterpart of the t-test in SRS case:

$$T^* = \sqrt{D_m(r-k-1)}\,\frac{\mathbf{a}'\hat{\boldsymbol{\beta}}_m - \mathbf{a}'\boldsymbol{\beta}}{\sqrt{\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}}} \tag{3}$$

which has the same distribution as that of the random variable

$$H = \sqrt{D_m(r-k-1)}\,\frac{\mathbf{a}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{e}}{\sqrt{\mathbf{V}'(\mathbf{I}-\mathbf{P})\mathbf{V}}} \tag{4}$$

where $\mathbf{V} = \mathbf{Z}/\sigma$. Then we reject the hypothesis H_0: $\mathbf{a}'\boldsymbol{\beta} = 0$ against $H_1 : \mathbf{a}'\boldsymbol{\beta} \neq 0$ at level α if $|T^*| > t^*_{\alpha/2}$. Similar tests can be obtained for one-sided hypotheses. If we are interested in testing the hypothesis

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0,$$

then the counterpart of the usual SRS F-test is

$$F^* = \frac{R^{*2}/(k-1)}{(1-R^{*2})/(r-k-1)}$$

where

$$R^{*2} = 1 - \frac{\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}}{\mathbf{Z}'(\mathbf{I}-\mathbf{J})\mathbf{Z}},$$

$\mathbf{1}' = (1, ..., 1)$ is an $(1 \times r)$ vector of ones and $\mathbf{J} = \mathbf{1}\mathbf{1}'/r$. We reject H_0 if $F^* > f^*_\alpha$, where f^*_α is the 100α quantile of the distribution of F^* which is free of the parame-

ters under H_0. Based on the statistic in (3), we may have the following 100(1-α)% confidence interval for $\mathbf{a}'\boldsymbol{\beta}$:

$$\mathbf{a}'\hat{\boldsymbol{\beta}}_m \pm t^*_\alpha \sqrt{\frac{\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}}{(r-k-1)D_m}}.$$

5. PREDICTION INTERVAL FOR NEW VALUE OF X

Assume that we are interested in finding the predicted value of Y_0, the value of Y at $X = x_0$. To do so, let $\mathbf{x}'_0 = (1, x_0, ..., x_0^k)$ and $\hat{Y}_0 = \mathbf{x}'_0\hat{\boldsymbol{\beta}}_m$. Hence,

$$\begin{aligned}\hat{Y}_0 - Y_0 &= x'_0(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) - e_0 \\ &= \mathbf{x}'_0(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\Delta - e_0\end{aligned}$$

Dividing the last equation by σ, we reach to the following quantity

$$\frac{\hat{Y}_0 - Y_0}{\sigma} = \mathbf{x}'_0(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{e} - U \tag{4}$$

which has a distribution free of parameters. Also the quantity

$$\frac{\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}}{\sigma^2(r-k-1)D_m} = \frac{\mathbf{V}'(\mathbf{I}-\mathbf{P})\mathbf{V}}{D_m(r-k-1)} \tag{5}$$

has a distribution free of parameters. Hence the quantity

$$\frac{\hat{Y}_0 - Y}{\sqrt{\dfrac{\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}}{(r-k-1)D_m}}} = \frac{\mathbf{x}'_0(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{e} - U}{\sqrt{\dfrac{\mathbf{V}'(\mathbf{I}-\mathbf{P})\mathbf{V}}{D_m(r-k-1)}}} = Q, \text{ say}$$

where $U = e_0/\sigma$ has a distribution free of parameters. Hence for a $100(1-\alpha)\%$ prediction interval for Y_0 the value of Y at $X = x_0$ is

$$\hat{Y}_0 \pm q_{\frac{\alpha}{2}} \sqrt{\frac{\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}}{(r-k-1)D_m}}$$

where $q_{\alpha/2}$ is the $100(1-\alpha/2)$ quantile of the distribution of the random variable Q. Since the theoretical distribution of Q has a complicated form, we will rely on simulations to find the quantiles of Q. For polynomial regression without intercept, the above theory remains valid but we need to replace $(r-k-1)$ by $(r-k)$ and to remove from the matrices $\mathbf{W}$ and $\mathbf{x}'_0$ the first column.

6. SIMULATION STUDY

In this section, we conduct a simulation study to compare the expected length of the $100(1-\alpha)\%$ confidence interval for the slope obtained via the MRSS to its

SRS counterpart. We consider the following values of x: -1, -2.5, 3.3, -2.1, 2.6, -2.9, 1.2, 1.3, -1.5, 2.7, 3, 4, -6, 9, 8, -2, 1, 3, 2, 1, 2 ,3, 1, 2, -2. When $r = 5$, the first 5 of x values are chosen for the regression design, similarly for $r = 10$, 15 and 25. Also we do our simulation for $k = 5$, $m = 2$, 3. The quantiles of H are approximated by the sample quantiles of a large sample of 5000 observations obtained from the exact distribution of H using simulation. We present the simulation results in Tables 5 and 6. It can be noticed from these tables that the expected length of the confidence interval for the slope β_2 has the following properties:

i. *The expected length of the confidence interval is decreasing as a function of r*

ii. The expected length of the confidence interval is decreasing as a function of m

iii. The expected length for the MRSS interval is less than that for SRSS interval, hence MRSS gives shorter confidence interval for the slope.

iv. In general, and under MRSS, the interval derived from polynomial model with normal errors, i.e., e_i's are iid normal, is shorter than the intervals derived under the other symmetric distribution in this simulation work.

TABLE 5. - *Expected lengths of the interval for β_2 when errors have N(0,1), Laplace(0,1)*

N(0,1)	*r*				
Scheme	5	10	15	20	25
SRS $m=1$	4.0617	2.4510	0.5411	0.4160	0.3764
MRSS $m=2$	2.7550	1.7220	0.3568	0.2709	0.2397
3	2.1413	1.3896	0.3012	0.2235	0.2005
4	1.9102	1.1942	0.2469	0.1903	0.1677
5	1.6726	1.0935	0.2282	0.1694	0.1462
6	1.5826	0.9831	0.2072	0.1527	0.1327
7	1.4312	0.9002	0.1761	0.1439	0.1277
Laplace(0,1)	*r*				
Scheme	5	10	15	20	25
SRS $m=1$	5.9597	3.6417	0.6915	0.5647	0.5180
MRSS $m=2$	3.3681	2.0206	0.4028	0.3456	0.2796
3	2.4464	1.5348	0.3113	0.2361	0.2129
4	2.0254	1.2310	0.2514	0.2026	0.1790
5	1.8028	1.0769	0.2186	0.1754	0.1484
6	1.5335	0.9799	0.1968	0.1501	0.1379
7	1.3269	0.8740	0.1682	0.1391	0.1189

TABLE 6. - *Expected lengths of the interval for* β_2 *when errors have* t_5, *Logistic*(0,1)

t_5	r				
Scheme	5	10	15	20	25
SRS $m=1$	5.4436	3.2575	0.6564	0.5309	0.4648
MRSS $m=2$	3.1634	1.8236	0.4118	0.3155	0.2746
3	2.4094	1.5492	0.3209	0.2551	0.2136
4	2.0691	1.3585	0.2548	0.2079	0.1798
5	1.8201	1.1710	0.2335	0.1884	0.1621
6	1.6485	1.0364	0.2106	0.1664	0.1454
7	1.4920	0.9621	0.1962	0.1527	0.1349
Logistic(0,1)	r				
Scheme	5	10	15	20	25
SRS $m=1$	6.9484	4.8724	1.0011	0.7687	0.6678
MRSS $m=2$	4.6131	2.9630	0.6089	0.4577	0.4136
3	3.6466	2.2685	0.4679	0.3814	0.3278
4	2.6886	1.9695	0.3868	0.3197	0.2749
5	2.4840	1.6922	0.3494	0.2803	0.2584
6	2.3769	1.5992	0.3047	0.2483	0.2228
7	2.1420	1.4828	0.2889	0.2202	0.2033

7. APPLICATION

In this section, we apply our theoretical results to real grassland data, obtained in a biodiversity research project in the federal state of Thuringia, East Germany. The data were collected from 78 montane grassland sites variables such as soil nutrient, soil water and pH, as well as biotic response variables, especially species richness, species abundances, percent cover and biomass (for more details see Kahmen, Perner, Audorff and Weisser, 2005). The observer might be especially interested in predicting the Shannon-Wiener entropy of a given site as a measure of diversity of that site in terms of species richness in the site. The Shannon-Wiener entropy of a community consisting of s species is defined by the equation $-\sum_{i=1}^{s} p_i \ln p_i$, where p_i denotes the relative abundance or amount of species i in a given plot. Species richness is defined as the total number of different species in a given plot and is much faster to observe, because it does not need any estimation of abundances. Figure 1 shows the scatter plot of Shannon-Wiener entropy vs. species richness for the 78 sites. The figure suggests a polynomial regression of order 2. Since Shannon-Wiener

entropy takes zero value when the number of species is zero, we consider the following second order polynomial regression

$$\text{Shannon Entropy} = \text{SpeciesRichness} + \beta_2 \times \text{SpeciesRichness}^2 + \varepsilon. \quad (6)$$

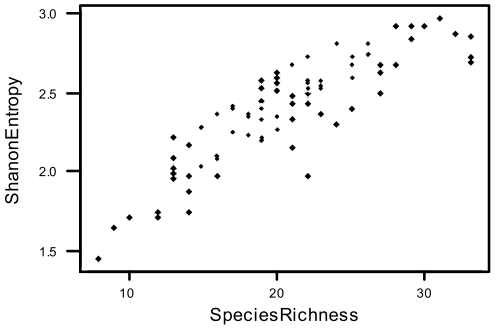

FIGURE 1. - *Shannon-Wiener entropy vs. Species richness for the 78 sites*

In this experiment, collecting cover or abundance data to calculate the Shannon-Wiener entropy is considered to be time consuming. Hence, it is reasonable to employ the MRSS technique to increase the efficiency of the study. For our purpose, we consider only the two variables:
Y = Shannon Entropy as response variable and X = Species richness as a predictor variable. For this we took a SRS of size $r = 10$ from the population. Alternatively, the MRSS was applied with $m = 2$, i.e. when the set size is 3. These two approaches produced the measurements which are presented in Table 7.

TABLE 7. - *SRS and MRSS(m = 2) of size r = 5*

X (Species Richness)	Y (Shannon Entropy)	
	SRS	MRSS
13	1.95	2.08
16	2.08	2.11
20	2.26	2.60
25	2.40	2.67
33	2.86	2.37

The Kolmogorov-Smirnov (KS) goodness of fit test was performed to test for normality of Y values. Since the sample size $r = 78 < 100$, then the exact KS test is used rather than the asymptotic one. The exact KS test, which is available by the R software, produced a P-value of 0.15. Hence, data could be treated as a sample from normal population. In order, to compare our results to some reference values, we fitted the model (6) to the complete 78 sites. Also we fitted, respectively, the model

(6) and the model (2) to SRS and MRSS to data presented in Table 7. The results are summarized in Table 8.

TABLE 8. - *Parameters estimates and their standard errors*

Parameter	SRS(r = 5)	MRS(r = 5)	78 sites
β_1	0.16874 (0.01388)	0.1950 (0.00622)	0.1850 (0.00378)
β_2	−0.00258 (0.00052)	−0.00343 (0.000233)	−0.00311 (0.000156)

Although, the standard errors of estimators are estimated from data, it can be noticed, from the results presented in Table 5, that the application results show that there is an improvement in accuracy which are consistent with our calculations in Table 4.

ACKNOWLEDGEMENT

We are indebted to Deutsche Forschungsgemeinschaft (DFG) for covering the stay of M.T. Alodat at Jena University, and Yarmouk University for partial financial support. Also we would like to thank the anonymous referee for the constructive comments which helped us to improve the paper.

REFERENCES

Chen Z., Bai Z., Sinha B.K. (2004). *Ranked set sampling: Theory and Applications*. Springer, New York.

Chen Z., Wang Y.G. (2004). Efficient regression analysis with ranked-set sampling. *Biometrics*, **60**, 997-1004.

Kahmen A., Perner J., Audorff V., Weisser W., Buchmann N. (2005). Effects of plant diversity, community composition and environmental parameters on productivity in montane European grasslands. *Oecologia*, **142**(4), 606-615.

Mashario H., Kazato O., Kanato M., Ichirow K., Aya N. Jyoken I., Hiroyuki H. (2007). Estimation of plant biomass and plant water content through dimensional masurements of plant volume in the Dund-Govi province of Mongolia. *Grassland Science*, **53**(4), 217-225.

Mead R. (1971). A note on the use and misuse of regression models in ecology. *Journal of Ecology*, **59**(1), 215-219.

Muttlak H.A. (1995). Parameters Estimation in Simple Linear Regression. Using Ranked Set Sampling. *The Biometrical Journal*, **37**, 799-810.

Muttlak H.A. (1997). Median ranked set sampling. *Journal of Applied Statistical Science*, **6**(4), 245-255.

Muttlak H.A. (2001). Regression estimators in extreme and median ranked set samples. *Journal of Applied Statistical Science*, **28**(8), 1003-1017.

Palmer M.W., Hussain M. (1997). The unimodal (species richness-biomass) relationship in microcommunities emerging from soil sees banks. *Proceeding of the Oklahoma Academy of Science*, **77**, 17-26.

Ryan L.M., Huang W., Thurston S.W., Kelsey K.T., Wiencke J.K., Christian D.C. (2004). On the use of biomarkers for environmental health research. Statistical Methods for Medical Research, **13**(3), 207-225.

Samawi H.M., Aby-dayyeh W. (2002). On regression analysis with random regressors using ranked samples. *International Journal of Information and Management Sciences*, **13**(3), 19-36.

Samawi H.M., Al-Saleh M.F. (2002). On regression analysis using bivariate ranked set samples. *Metron-International Journal of Statistics*, **LX**(3-4), 31-50.

Särndal C.E., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Shao J. (2003). *Mathematical Statistics.* Springer, New York.

Tillé Y. (2006). *Sampling Algorithms.* Springer, New York.

Yarranton G.A. (1969). Plant ecology: a unifying model. *Journal of Ecology*, **57**(1), 245-250.