Ridge regression technique to determine the environmental influences on tef (*Eragrostis tef*) grain yield

Legesse Kassa Debusho
Department of Statistics, University of Pretoria, Information Technology Building, Pretoria, 0002, Republic of South Africa
Email: Legesse.debusho@up.ac.za

Accepted 21 April 2008

Tef grain yield is dependent on a number of component characters such as plant height and panicle length. These characters and consequently yield are governed by a large number of factors including environmental factors. The objective of this paper was to determine the environmental influences on tef grain yield. The effects of eleven environmental variables on tef yield were studied using least squares and ridge regression analyses. The results revealed that the least squares estimates of regression coefficients for seven environmental variables did not give a correct indication of the influence of the variables on tef yield but that the estimates from ridge regression were stable. In addition, the ridge regression model has a lower mean square error. Tef grain yield is positively correlated to rainfall, average monthly minimum temperatures, soils dominated by silt with adequate quantities of available nitrogen while it is negatively influenced by average monthly maximum temperatures and clay soils. The results obtained from this study are in accordance with previous literature about the effects of environmental factors on yield of tef cultivars.

Keywords: Grain yield, multicollinearity, ridge regression, tef, variance inflation factor

Introduction

Tef, *Eragrostis tef* (Zucc.) Trotter, is an important cereal crop in Ethiopia and Eritrea. Outside these countries there is a growing interest in using tef. For example, small scale commercial production of tef has begun in a few areas of the wheat belts of the USA, Canada, and Australia. Tef has been introduced to South Africa and is cultivated as a forage crop and in recent years it has been cultivated as a cereal crop in northern Kenya (National Research Council, 1996; Ketema, 1997). Several endemic and non-endemic species of *Eragrostis* are found in Ethiopia. Existing genetic diversity for tef indicates that it originated and was domesticated in Ethiopia (Ketema, 1997), confirming a historic report by Vavilov (1951) that identified Ethiopia as the centre of origin and diversity of tef. Tef ranks first among all the cereals, pulses and oil crops in area coverage and is cultivated on about 22.7% of the total area cultivated and it constitutes about 18.7% (second to maize) of the gross yearly grain production of cereals in Ethiopia (CSA, 2002). An extensive review of literature on tef is given by Kebebw *et al.* (2003) and Kassa *et al.* (2006).

Firstly, the plants' potential for growth and development is determined by their genetic composition and secondly, on the environment which in the broadest sense includes soil and climatic factors. Tef grain yield is dependent on a number of component characters such as plant height, panicle length, etc. (Kassa *et al.*, 2001). These characters and consequent yield are governed by a large number of factors including environmental influences (Katiyar *et al.*, 1979).

As far as the effect of environmental factors on the growth of tef was concerned, only broad generalizations were made which indicated that the crop is adapted to a great diversity of climatic conditions and soil types (Ebba, 1969; Assefa, 1978). Murphy (1968) studied different soils in the accessible parts of Ethiopia and in some instances, their relationships with cereals.

The impact of the environmental factors on grain yield of tef can be assessed using multiple regression analysis. Usually, these independent (explanatory) variables are highly correlated with each other (Kassa *et al.*, 2003), which is called multicollinearity. Multicollinearity concerns the situation where, because of a strong relationship among the explanatory variables, it becomes difficult to estimate their separate effects on the dependent variable, i.e. tef grain yield. In the presence of multicollinearity, the least squares estimation method may result in regression coefficients with much larger/smaller value or different sign than expected (Neter, *et al.*, 1996). The method of ridge regression developed by Hoerl and Kennard (1970) resolves the problem of multicollinearity in the data by introducing biased estimates of the regression coefficients but with a higher precision than least squares estimates. Thus, the objective of this study was to investigate the effect of several environmental factors on tef grain yield using ridge regression analysis.

Material and methods

Data sources

Data were collected from the Debre Zeit Agricultural Research Centre in Ethiopia, located at $8^{o}44'$ N, $38^{o}58'$ E. Field trials were conducted at four sites (Debre Zeit black soil, Debre Zeit light soil, Akaki and Koka) and during three main growing seasons using a randomized complete block design with four replications. Twelve genotypes in this multilocation trial were sown on 4 m^2 field plots at a spacing of 2 m between blocks and 1.5 m between plots within blocks. The data collected were grain yield and eleven environmental variables, namely altitude in meters (x_1), rainfall (mm) (x_2), average monthly minimum temperature (^{o}C) (x_3), average monthly maximum temperature (^{o}C) (x_4), nitrogen content (%) (x_5), phosphorus (mg kg^{-1}) (x_6), pH (x_7), sand content (%) (x_8), organic matter (%) (x_9), silt (%) (x_{10}) and clay (%) (x_{11}). Climatic data from the National Meteorological Services Agency of Ethiopia were compared with data from the Centre.

Statistical analysis

Ridge regression procedure

In matrix notation the multiple linear regression model is given as:

$$y = X\beta + e$$

where y is an $n \times 1$ vector of grain yields, X is an $n \times p$ matrix of explanatory variables which, in this case, comprises climate and soil variables, β is a $p \times 1$ vector of unknown regression coefficients and e is the $n \times 1$ vector of experimental errors with mean 0 and variance σ^2. Ridge regression is a method of obtaining biased estimates of regression coefficients. In ridge regression a non-negative constant k is added to the diagonal of the correlation matrix, $X'X$ where X' is the transpose of X, among the regression variables before inverting the matrix for least squares estimation. Therefore the ridge estimates of the regression coefficients, β (Neter *et al*., 1996) is given as:

$$\hat{\beta}^* = (X^T X + kI)^{-1} X^T y$$

where I is the identity matrix. Note that, when $k = 0$, the ridge estimator is identical to the least squares estimator. Hoerl and Kennard (1970) justify the use of the ridge regression by providing the existence of a value of k for which the mean square error for ridge regression coefficient estimates is less than that for the ordinary least squares regression estimates. In cases like that, variance is a decreasing function of k while bias in the estimates is an increasing function of k. Thus, as k increases, the mean square error of a coefficient decreases to a minimum and then increases.

The choice of an appropriate value of k is based upon two factors: regression coefficients must be stable and variance inflation factor must be small. An aid to choosing values of k which lead to stable regression coefficients is the ridge trace. The ridge trace is a simultaneous plot of the values of the estimated ridge standardized regression coefficients for different values of k, usually between 0 and 1. It serves to display the complex interrelationships that may exist between the explanatory variables and the effect of these interrelationships on the estimation of the regression coefficients.

The variance inflation factor (*VIF*) is a method of detecting multicollinearity and given by (Neter *et al*., 1996):

$$(VIF)_i = \frac{1}{1 - R^2}; \quad i = 1,2,..., p$$

where p is the number of explanatory variables in the model and R^2 is the coefficient of determination when x_1 is regressed on the $(p - 1)$ X variables in the model.

Stability in the estimates should not be used solely as a criterion for selecting a value of since a great degree of bias can be introduced if k increases. Thus, the ridge trace should be used in conjunction with *VIF* (Hoerl & Kennard, 1970). The ridge estimates $\hat{\beta}^*$ may fluctuate widely as k is changed slightly from 0 and some may even change their numerical signs. Gradually, however, these wide fluctuations cease and the magnitudes of the estimates tend to move slowly toward zero as k is increased further. At the same time, the values of $(VIF)_i$ tend to fall rapidly as k is changed from 0 and gradually the $(VIF)_i$ values also tend to change only moderately as k is increased further. One therefore examines the ridge trace and the *VIF* values and chooses the smallest value of k where it is deemed that the regression coefficients first become stable in the ridge trace and the *VIF* values have become sufficiently small.

Once the presence of multicollinearity in the explanatory variables was established, ridge regression analysis was used to analyse 18 different values of k. All analyses were done with the SAS/STAT statistical package (SAS Institute, 2004).

Results and discussion

The correlation matrix among the environmental (explanatory) variables is presented in Table 1. The correlation coefficients suggest that most environmental variables, for instance x_1 and x_4, x_3 and x_{10}, x_7 and x_{11}, are correlated, which present the problem of multicollinearity.

The ridge trace of the standardized regression coefficients with various values of k are given in Figure 1. The regression coefficients seem to be stable at k = 0.4 - 0.5, but are very

Table 1 Correlation matrix among the environmental variables[a]

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
Altitude (x_1)	1.000										
Rainfall (x_2)	-0.508	1.000									
Min (x_3)	-0.986	0.496	1.000								
Max (x_4)	-0.987	0.503	0.984	1.000							
N (x_5)	-0.289	0.114	0.138	0.174	1.000						
P (x_6)	0.442	-0.256	-0.572	-0.522	0.636	1.000					
pH (x_7)	-0.227	0.136	0.305	0.247	-0.245	-0.735	1.000				
Sand (x_8)	0.250	-0.161	-0.395	-0.342	0.769	0.978	-0.714	1.000			
Organic (x_9)	-0.382	0.187	0.339	0.314	0.529	-0.154	0.693	-0.042	1.000		
Silt (x_{10})	-0.963	0.497	0.983	0.967	0.116	-0.646	0.464	-0.474	0.463	1.000	
Clay (x_{11})	0.556	-0.277	-0.515	-0.492	-0.530	-0.240	-0.678	0.099	-0.981	-0.623	1.000

[a] x_3: average monthly minimum temperature, x_4: average monthly maximum temperature, x_5: nitrogen, x_6: phosphorus, x_7: pH, x_8: sand, x_9: organic matter, x_{10}: silt and x_{11}: clay.

unstable at small *k*-values. However, just looking at the ridge trace, it is difficult to choose the optimum value of *k*.

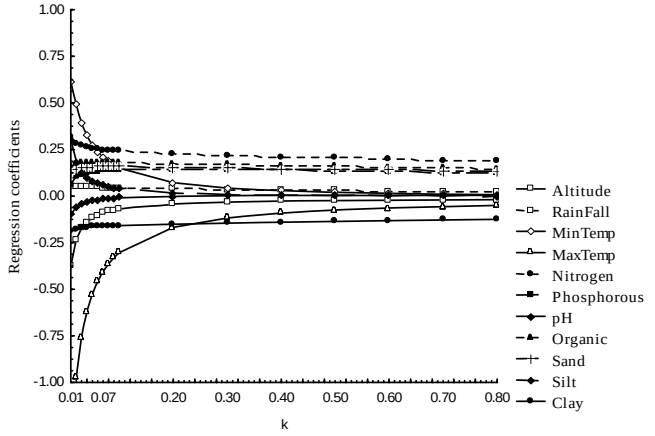

Figure 1 Ridge trace of estimated standardized regression coefficients.

Table 2 Variance inflation factor for regression coefficients for biasing constants $k = 0.0$, $k = 0.40$ and $k = 0.45$

Explanatory variables	$k = 0.0$ $(VIF)_i$	$k = 0.40$ $(VIF)_i$	$k = 0.45$ $(VIF)_i$
Altitude	662.636	8.006	7.257
Rainfall	1.405	1.201	1.197
Average monthly minimum temperature	314.939	7.800	7.113
Average monthly maximum temperature	119.271	7.331	6.729
Nitrogen	647.391	5.271	4.792
Phosphorus	712.324	7.013	6.349
pH	570.165	4.263	3.908
Sand	645.732	5.834	5.299
Organic	673.097	5.308	4.833
Silt	736.056	10.225	9.217
Clay	694.087	6.357	5.767

Neter *et al*. (1996) stated that if the variance inflation factor exceeds 10 then it is an indication that the associated coefficients are poorly estimated because of multicollinearity. In this study, the variance inflation factors at $k = 0.45$ are less than 10 for all explanatory variables (Table 2). Thus, the optimum value of the biased constant *k* is chosen at 0.45 in combination with variance inflation factor.

The values of the standardized regression coefficients at $k = 0.0$ and $k = 0.45$ are illustrated in Table 3. The ridge regression coefficients $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_6$, $\hat{\beta}_7$, $\hat{\beta}_8$, and $\hat{\beta}_9$ have changed sign compared with least squares estimates. In some cases, the ridge regression procedure changes the non-significant least squares regression coefficients, for instance with *p*-value 0.806, to a significant estimated coefficients (Draper & Smith, 1998). Further, the decrease in the mean square error is about 14.93 % at $k = 0.45$ compared with $k = 0.0$. Therefore, the ridge regression model appears better than the least squares model.

The general results obtained from ridge regression analysis using $k = 0.45$ are in accordance with literature on the effect of environmental factors on yield of tef cultivars (Assefa, 1978; Isak, 1982). For instance, it is expected that rainfall and tef grain yield are positively related but because of multicollinearity the least squares estimate (k = 0.0), $\hat{\beta}_2$ has negative value, which is opposite to what is expected. The Central Highlands of Ethiopia where most of the tef is grown get an average annual rainfall of 950-1000 mm, but in extreme cases can be as high as 2500 mm (Ketema, 1993).

In this study it was found that the average monthly maximum temperature affects tef yield negatively and average monthly minimum temperature affects tef grain yield positively. This is in contradiction with Wolde (1974) who found that high temperatures have no effect on the yield of tef. Assefa (1978) determined that optimum temperature for high grain yield is 15^{o}-21^{o}C, confirming the findings of this study.

Tef grain yield is positively correlated to soils dominated by silt, with adequate quantities of available nitrogen, while grain yield is negatively influenced by clay soils (Table 3). This is in accordance with the findings of Isak (1982).

Different sources have also given different views on the optimum altitudes for tef production. It can be grown from sea level up to 2800 m (Ketema, 1997). However, according to experiences gained so far from national yield trials conducted at different locations across the country, tef performs excellently at an altitude of 1800–2100 m. According to Ketema (1997) good growth can also be obtained at an altitude range of 1700–2200 m. This study, however, failed to prove the expected positive impact of altitude on tef yield. This may be attributed to the fact that only three different data values for altitude were used for analysis and unlike other environmental variables, altitude will not change with seasons.

Conclusion

This paper demonstrates the rationale of using ridge regression in agricultural research to deal with the multicollinearity problem of data. The instability of the least squares estimates can be seen in Figure 1. Further, it was observed that the least squares estimates of the regression coefficients for rainfall, average monthly minimum temperature, phosphorus, pH, sand content and organic matter do not give a true indication of the influence of these components on tef grain yield (Table 3). Therefore, when faced with the problem of multicollinearity in the data, ridge regression analysis is an important tool to identify a correct model. When used with caution and care, ridge regression analysis offers considerable value in obtaining reasonable and stable coefficients, as well as in attaining a higher precision, i.e. minimum mean square error (Draper & Smith, 1998). The results obtained from ridge regression analysis using $k = 0.45$ agree with the broad generalizations made by most authors about the effects of environmental factors on yield of tef cultivars. However, the study failed to show the positive relation that was expected between altitude of the site and tef yield. Further work needs to be done in order to confirm the findings by including more locations at different altitudes

.

Table 3 Least squares and ridge estimate of the standardized regression coefficients

Standardized regression coefficients[b]	Least square (k = 0.0)			Ridge (k = 0.45)		
	Estimate	Standard error	p-value	Estimate	Standard error	p-value
β_1	-4.401	3.846	0.255	-0.026	0.086	0.765
β_2	-0.029	0.065	0.659	0.031	0.059	0.602
β_3	-3.996	1.526	0.009	0.024	0.086	0.781
β_4	-4.858	0.858	<0.001	-0.083	0.085	0.332
β_5	1.035	4.120	0.806	0.208	0.081	0.012
β_6	-1.159	4.383	0.792	0.137	0.084	0.106
β_7	-1.268	3.917	0.747	0.003	0.079	0.966
β_8	-0.297	4.290	0.924	0.162	0.082	0.076
β_9	-0.401	4.199	0.945	0.135	0.081	0.052
β_{10}	3.787	4.201	0.367	0.004	0.088	0.968
β_{11}	-0.478	4.360	0.913	-0.138	0.083	0.100
Mean square error		0.965			0.821	

[b] β_1 : coefficient of x_1

References

ASSEFA, M., 1978. Floral morphogenesis, temperature effect on growth and development and variation in nutritional composition and distribution among cultivars in *Eragrostis tef* (Zucc.) Trotter. PhD Thesis, University of Wisconsin, Madison, Wisconsin.

CENTRAL STATISTICAL AUTHORITY (CSA), 2002. Ethiopian agricultural sample enumeration: Report on the preliminary results of area, production and yield of temporary crops. Part 2, Addis Ababa, Ethiopia.

DRAPER, N.R. & SMITH, H., 1998. Applied regression analysis, 3rd edn, John Wiley & Sons, Inc., New York.

EBBA, T., 1969. Tef (*Eragrostis tef*). The cultivation, usage, and some of the known diseases and insect pests. Part I. Agri. Expt. Sta. Bull. No. 60. Haile Selassie I University, College of Agriculture, Dire Dawa, Ethiopia.

HOERL, A.L. & KENNARD, R.W., 1970. Ridge regression: Application to non-orthogonal problems. *Technometrics* 12, 69-82.

ISAK, S., 1982. The effect of different soils on the growth and yield of some tef cultivars. M.Sc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

KASSA, L., HAMITO, D. & KOROTO, T., 2001. Multivariate assessment of environmental effects on grain yield and component characters of tef [*Eragrostis tef* (zucc.) *Trotter*] genotypes. I. Principal components analysis of grain yield and component characters. *Ethiopian Journal of Statistical Association*, 11, 43-52.

KASSA, L., HAMITO, D. & KOROTO, T., 2003. Multivariate assessment of environmental effects on grain yield and component characters of tef (*Eragrostis tef Trotter*). Part II: Canonical correlation and canonical variate analysis of grain yield and components characters. *Ethiopian Journal of Statistical Association*, 13, 17-25.

KASSA, L.D., SMITH, M.F. & FUFA, H., 2006. Stability analysis of grain yield of tef (*Eragrostis tef*) using the mixed model approach. *S. Afr. J. Plant Soil*, 23, 38-42.

KATIYAR, R.P., 1979. Correlations and path analysis of components in chickpea. *Indian J. Agric. Sci.* 49, 35-38.

KEBEBEW, A., ARNULF, M. & TEFERA, H, 2003. Multivariate analysis of diversity of tef (*Eragrostis tef* (Zucc.) Trotter) germplasm from western and southern Ethiopia. *Hereditas*, 138, 228-236.

KETEMA, S., 1993. Tef (*Eragrostis tef*) Breeding, genetic resources, utilization and role in Ethiopia agriculture. Institute of Agricultural Research, Addis Ababa, Ethiopia.

KETEMA, S., 1997. *Eragrostis tef* (Zucc.) Trotter. Promoting the conservation and use of underutilized and neglected crops. Bioversity International Publication No. 12, Gatersleben/International Plant Genetic Resources Institute, Rome, Italy.

MURPHY, M.F., 1968. A report on the fertility status and other data on some soil of Ethiopia. Exper. Sta. Bull. No. 44, HISU. College of Agri. Dire Dawa, Ethiopia.

NATIONAL RESEARCH COUNCIL, 1996. Lost crops of Africa. Volume 1: Grains. National Academy Press, Washington DC.

NETER, J., KUTNER, M.H., WASSERMAN, W. & NACHTSHEIM, C.J., 1996. Applied linear statistical models, 4th edn, McGraw-Hill/Irwin, Boston, Chicago.

SAS Institute, 2004. SAS User's Guide: Statistics. Version 9.1 edn. SAS Inst., Cary, North Carolina.

VAVILOV, N.I., 1951. The origin variation, immunity and breeding of cultivated plants. Ronald press, New York.

WOLDE, T., 1974. Agroclimatology of tef. *In:* Agroclimatology of the highlands of eastern Africa. World Meteorological Organaization (WMO) Report No. 389. Geneva, Switzerland.