SOME APPLICATIONS OF ENTROPY-BASED STATISTICS IN LINEAR REGRESSION ANALYSIS

Dr. Atıf EVREN[1], Dr. Elif TUNA[1]
[1] Yildiz Technical University ,Faculty of Sciences and Literature, Department of Statistics, İstanbul Turkey, aevren2006@gmail.com, elfztrk@gmail.com

ABSTRACT

Statistical entropy is a measure of variation of a distribution especially when the random variable is qualitative. Entropy-based statistics are also used to measure the degree of association between qualitative variables. Two measures of divergence, namely, Kullback-Leibler divergence and Jeffreys' divergence are closely related to loglikelihood function. Thus these two entropy-based measures can be used in hypothesis testing procedures as well. In this study, we discuss how relative entropy measures are applied in testing some hypotheses and how useful they would be in regression analysis especially in determining influential observations.

Keywords: *Entropy, Kullback-Leibler divergence, Jeffreys' divergence, hypothesis testing in regression models*

INTRODUCTION

For Boltzmann, the entropy of a physical system is a measure of disorder related to it (Renyi 2007). The entropy of statistical mechanics is Boltzmann's constant times the natural logaritm of the number of possible states (Pierce 1980). The entropy of a statistical experiment, on the other hand, can be evaluated as a measure of uncertainty before a statistical experiment takes place (Khinchin 1957) So in a statistical sense, entropy and the amount of information are two closely related concepts. Uncertainty is not present after a statistical experiment is conducted. Therefore , entropy can be viewed as the amount of information that can be gathered through sampling (Renyi –b-2007). For some introductory concepts and applications of statistical entropy, one can

also refer to (Ash 1990), (Cover and Thomas 2006), (Evren 2010),(Evren 2011) and (Reza 1994).

2. ENTROPY

Some entropy measures proposed are for a discrete sample space S are as below:

$$H_s(p) = -\sum p_i \log p_i \quad \text{(Shannon entropy 1948)} \tag{1}$$

$$H_\alpha(p) = \frac{1-\sum p_i^\alpha}{\alpha - 1}, \alpha > 0, \alpha \neq 1 \quad \text{(Havrda and Charvát 1967)} \tag{2}$$

$$H_R(p) = \frac{\log \sum p_i^\alpha}{1-\alpha}, \alpha > 0, \alpha \neq 1 \quad \text{(Rényi 1961)} \tag{3}$$

All these measures are nonnegative and are equal to zero whenever any one of probabilities; $p_i, i = 1,2,...k$, is 1.Besides, maximum values of all measures are attained when the conditions $p_i = 1/k, i = 1,2,...,k$ are met simultaneously (Rao 1984). For continuous distributions, the summation operators in (1), (2) and (3) should simply be replaced by integrals to measure entropy. Generalizations on qualitative or multivariate distributions are also straightforward.

3. DIVERGENCE MEASURES BASED ON ENTROPY

Let $f_1(x)$ and $f_2(x)$ be two probability distributions defined by two hypotheses H_1 and H_2. Here the ratio $\log \frac{f_1(x)}{f_2(x)}$ is defined as the amount of information supporting the hypothesis H_1. An average measure is notated by $I(1:2)$.

This statistic can be evaluated as the average amount of information coming from $f_1(x)$, supporting H_1 . $(H_1 : f_1(x) \neq f_2(x); H_2 : f_1(x) = f_2(x))$

$$I(1:2) = \int_{-\infty}^{\infty} f_1(x) \log \frac{f_1(x)}{f_2(x)} dx \tag{4}$$

Similarly one can derive

$$I(2:1) = \int_{-\infty}^{\infty} f_2(x) \log \frac{f_2(x)}{f_1(x)} dx \tag{5}$$

This statistic can also be viewed as the average amount of information from $f_2(x)$ that supports H_1. In literature, these two statistics in (4) and (5) are known as Kullback-Leibler's divergences as well. A symmetric version of I(1:2) is Jeffreys' divergence which is defined as

$$J(1:2) = I(1:2) + I(2:1) \tag{6}$$

4. ASYMPTOTIC PROPERTIES OF I(1:2) AND J(1:2)

The following statistic based on a sample of n observations

$$2n\hat{I} = 2n\left(\int f(x,\theta) \log \frac{f(x,\theta)}{f(x,\theta_2)} d\lambda(x) \right)_{\theta = \hat{\theta}} \tag{7}$$

(under the assumption $H_1 : \theta \neq \theta_2$) follows asymptotically a chi-square distribution with k (the elements of parameter vector) degrees of freedom. Here $f(x,\theta)$ is a general multivariate probability distribution having multiple parameters $\hat{\theta}$. It is also assumed to

be a consistent, asymptotically multivariate normal, efficient random vector estimator of θ. The statistic

$$n\hat{J} = n\left(\int \left(f(x,\theta) - f(x,\theta_2) \right) \log \frac{f(x,\theta)}{f(x,\theta_2)} d\lambda(x) \right)_{\theta = \hat{\theta}} \tag{8}$$

is a chi-square variate with k degrees of freedom, asymptotically (Kullback 1996).

5. MULTIVARIATE NORMAL DISTRIBUTIONS AND ENTROPY BASED DIVERGENCE MEASURES

Consider two k-variate normal populations, $Normal(\mu_i, \Sigma_i)$, with $\mu_i' = (\mu_{i1}, \mu_{i2}, ..., \mu_{ik})$, i=1,2 the vectors of mean values, and $\Sigma_i = (\sigma_{irs})$, i=1,2; r,s=1,2,…,k, the covariance matrices . The densities are

$$f_i(x_1, x_2, ..., x_k) = \frac{1}{\left| (2\pi)^k \Sigma_i \right|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) \right) \tag{9}$$

from which one can get

$$\mathrm{I}(1:2) = \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2} tr\Sigma_1 \left(\Sigma_2^{-1} - \Sigma_1^{-1} \right) + \frac{1}{2} tr\Sigma_2^{-1} (\mu_1 - \mu_2)(\mu_1 - \mu_2)' \tag{10}$$

$$\mathrm{J}(1:2) = \frac{1}{2} tr(\Sigma_1 - \Sigma_2)\left(\Sigma_2^{-1} - \Sigma_1^{-1} \right) + \frac{1}{2} tr(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2)(\mu_1 - \mu_2)' \tag{11}$$

6. LINEAR REGRESSION MODELS

In matrix form, the linear regression model is written as $y = X\beta + \varepsilon$ where y is an nx1 vector of dependent- variable values, X is an nxp matrix of regressors, including the constant regressor of 1's as its first column, β is a px1 vector of regression parameters; ε is an nx1 vector of errors. The least-squares estimates b of β are

$$b = (X'X)^{-1} X'y \tag{12}$$

Assuming $\varepsilon \sim Normal(0, \sigma^2 I)$ one concludes b~ $Normal_{k+1}(\beta, \sigma^2 (X'X)^{-1})$ (13)

A 100(1-α)% ellipsoidal joint confidence region for the regression coefficients are given by

$$(b-\beta)'(X'X)(b-\beta) \le ps^2 F_{\alpha,p,n-p}$$ (Griffiths et all.1993) (14)

where $s^2 = e'e/(n-p)$ is an unbiased estimator of σ^2, and $F_{\alpha,p,n-p}$ is the critical α value of an F distribution with p and n-p degrees of freedom (Fox 1991).

7. TESTING A LINEAR HYPOTHESIS

Let the two hypotheses H_1 and H_2 be given as $H_1 : \beta = \beta^{(1)}$ and $H_2 : \beta = \beta^{(2)}$

Then under the assumptions of classical linear regression model,

$$J(1:2) = (\beta^{(1)} - \beta^{(2)})' X'X (\beta^{(1)} - \beta^{(2)}) / \sigma^2 \tag{15}$$

In particular, for the common hypothesis $H_2 : \beta^{(2)} = 0$

$$\hat{J}(1:2) = \left(\beta^{(1)}\right)' X'X\left(\beta^{(1)}\right) / \hat{\sigma}^2 \tag{16}$$

Note that $\hat{J}(1:2)$ in (16) is the quadratic form in the exponent of the multivariate normal distribution of the β_i 's with the covariance matrix replaced by an unbiased estimate with n-p degrees of freedom. $\hat{J}(1:2)$ is thus Hotelling's generalized Student ratio (Hotelling's T^2). Under the assumptions of classical regression analysis

$$(n-p)s^2 / \sigma^2 \sim \chi^2_{n-p} \tag{17}$$

$$F_{p,n-p} = \frac{\chi^2_p / p}{\chi^2_{n-p} / (n-p)} = \left(\beta^{(1)} - \beta^{(2)}\right)' X'X\left(\beta^{(1)} - \beta^{(2)}\right) / ps^2 \tag{18}$$

8. ANALYSIS OF RESIDUALS IN LINEAR REGRESSION MODELS

Let $b_{(-i)}$ denote the vector of least squares regression coefficients calculated with the ith observation deleted from data set. Cook's D_i is the F value for testing the "hypothesis" that $\beta = b_{(-i)}$

$$D_i = \frac{\left(b - b_{(-i)}\right)' X'X\left(b - b_{(-i)}\right)}{ps^2} = \hat{J}(1,2) / p \tag{19}$$

An alternative way of calculating Cook's D_i can be done by the formula given below (Rawlings 1988):

$$D_i = \frac{\left(\hat{Y}_{(i)} - \hat{Y}\right)'\left(\hat{Y}_{(i)} - \hat{Y}\right)}{ps^2} \tag{20}$$

$\hat{Y}_{(i)}$ values are estimated by the new regression function obtained by deleting the ith observation from the data set at each iteration.

9. APPLICATION

In the following regression model, the dependent variable is taken to be the net migration to Italy. The independent variables are population of Italy, Italian GDP per capita and Italian unemployment rates for the years between 1977 and 2004 as shown in the following table:

Table 1. Migration Data of Italy between 1978-2004

Year	Net Migration (thousand)(C2)	GDP per capita (current $) (C3)	Unemployment Rate (%) (C4)	Total Population (thousand) (C5)
1978	0.1	5.406	7.2	55446
1979	0.1	6.735	7.7	55602
1980	0.1	8.148	7.6	55657
1981	-0.3	7.344	7.9	55774
1982	1.9	7.281	8.5	55995
1983	2.4	7.547	9.4	56228
1984	1.6	7.457	10.1	56344
1985	-0.5	7.699	10.3	56498
1986	-0.4	10.902	11.1	56576
1987	-0.2	13.715	12	56664
1988	0	15.71	12	56763
1989	0.2	15.788	12	56837
1990	0.2	19.983	11.4	56737
1991	0.1	21.059	10.9	56760
1992	3.2	22.286	10.5	56859
1993	3.2	17.965	9.7	56442
1994	2.6	18.54	10.6	56623
1995	1.6	19.809	11.2	56745
1996	2.6	22.152	11.2	56826
1997	2.2	20.957	11.3	56941
1998	1.6	21.386	11.3	57040

1999	1.8	21.096	11	57078
2000	3.1	19.269	10.2	57189
2001	2.2	19.609	9.1	57348
2002	6.1	21.326	8.6	57474
2003	10.6	16.164	8.5	57478
2004	7.9	29.699	8.1	57553

The summarizing statistics for the linear regression model obtained by NCSS are as follows:

Table 2-a Summarizing statistics for regression model (full model)

Run Summary Section		
Parameter	**Value**	**Parameter**
Dependent Variable	C2	Rows Processed
Number Ind. Variables	3	Rows Filtered Out
Weight Variable	None	Rows with X's
Missing	1	
R2	0.7379	Rows with Weight
Missing	0	
Adj R2	0.7038	Rows with Y
Missing	0	
Coefficient of Variation	0.7125	Rows Used in
Estimation	27	
Mean Square Error	2.030653	Sum of Weights
Square Root of MSE	1.42501	Completion Status
Ave Abs Pct Error	172.069	

Regression Equation Section				
	Regression	**Standard**	**T-Value**	
Independent	**Coefficient**	**Error**	**to test**	**Prob**
Variable	**b(i)**	**Sb(i)**	**H0:B(i)=0**	**Level**
Intercept	-210.8484	48.6969	-4.330	0.0002
C3	0.0066	0.0769	0.085	0.9327
C4	-1.1382	0.2047	-5.560	0.0000
C5	0.0040	0.0009	4.487	0.0002

Table 2-b Summarizing statistics for regression model (full model)

Normality Tests Section			
Test	**Test**	**Prob**	**Reject H0**
Name	**Value**	**Level**	**At Alpha = 20%?**
Shapiro Wilk	0.952	0.239748	No
Anderson Darling	0.4739	0.241655	No
D'Agostino Skewness	0.1765	0.859913	No
D'Agostino Kurtosis	1.9456	0.051698	Yes
D'Agostino Omnibus	3.8167	0.148328	Yes

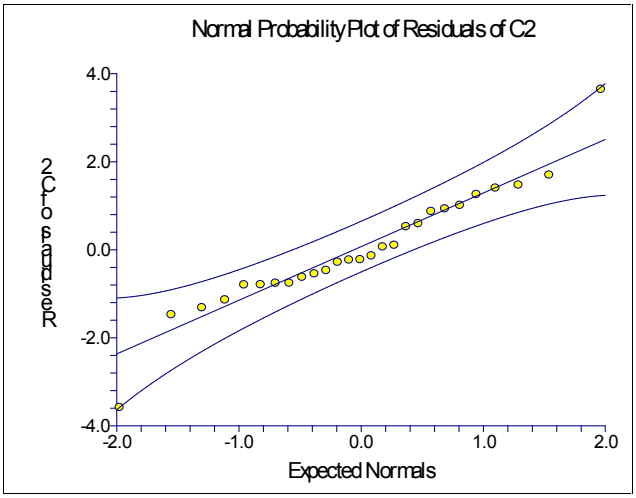


Figure 1. Normal probability plot of residuals

By Table 2-b and Figure 1, one can conclude that the residuals distribute normally roughly (at α=0.20). So I(1:2) and J(1:2) statistics can roughly be applied to data.

The Impact of GDP per Capita on Net Migration to Italy

If we look at the results more cautiously, we can see that the coefficient of GDP per capita falls the neighbourhood of zero which is a problem that needs obvious remedy. Let the two hypotheses to investigate the impact of GDP per capita on migration be formulated by the following vectors

$$H_1 : \beta = \beta^{(1)} = \begin{bmatrix} -210.84 \\ 0.006 \\ -1.13 \\ 3.95 \end{bmatrix} \quad \text{and} \quad H_2 : \beta = \beta^{(2)} = \begin{bmatrix} -210.84 \\ 0 \\ -1.13 \\ 3.95 \end{bmatrix}$$

The related I(1:2) and J(1:2) statistics are 0.08 and 0.165 respectively. The critical chi-square variable (p=4, α=0.05) is 9.48 Thus the effect of GDP per capita on migration to Italy seems to be marginal. To support this point of view, one may drop GDP per capita from the set of independent variables and hence get the summarizing statistics for this new (reduced) model as below:

Table 3. Summarizing statistics of "reduced model" obtained by deleting GDP per capita as an explanatory variable

Run Summary Section				
Parameter	**Value**	**Parameter**		
Dependent Variable	C2	Rows Processed		
Number Ind. Variables	2	Rows Filtered Out		
Weight Variable	None	Rows with X's		
Missing	1			
R2	0.7379	Rows with Weight		
Missing	0			
Adj R2	0.7160	Rows with Y		
Missing	0			
Coefficient of Variation	0.6976	Rows Used in		
Estimation	27			
Mean Square Error	1.946658	Sum of Weights		
Square Root of MSE	1.395227	Completion Status		
Ave Abs Pct Error	170.139			
Regression Equation Section				
	Regression	**Standard**	**T-Value**	
Independent	**Coefficient**	**Error**	**to test**	**Prob**
Variable	**b(i)**	**Sb(i)**	**H0:B(i)=0**	**Level**
Intercept	-214.1653	28.7205	-7.457	0.0000
C4	-1.1363	0.1993	-5.701	0.0000
C5	0.0040	0.0005	7.721	0.0000

By checking the summarizing statistics of this "reduced" linear model , we can still assert that impact of GDP per capita on migration can be neglected. So these two results do not conflict each other.

10. THE IMPACT OF UNEMPLOYMENT RATE

As a second application, we try to determine the affect of unemployment rate on the dependent variable (the net migration to Italy). The null and alternative hypotheses are as follows:

$$H_1 : \beta = \beta^{(1)} = \begin{bmatrix} -210.84 \\ 0.006 \\ -1.13 \\ 3.95 \end{bmatrix} \quad \text{and} \quad H_2 : \beta = \beta^{(2)} = \begin{bmatrix} -210.84 \\ 0.006 \\ 0 \\ 3.95 \end{bmatrix}$$

I(1:2) and J(1:2) are found to be 876 and 1752 respectively. Since these are very far away from critical chi-square value, even without checking critical values in probability tables, we can say that the divergence between these two hypotheses can not be neglected. Hence the impact of unemployment on migration to Italy is not marginal. To check this, one can analyse the statistics of another reduced model by dropping unemployment rate from the set of explanatory or independent variables..

Table 4. The Summarizing statistics of reduced model obtained by dropping unemployment rate

Run Summary Section		
Parameter	**Value**	**Parameter**
Dependent Variable	C2	Rows Processed
Number Ind. Variables	2	Rows Filtered Out
Weight Variable	None	Rows with X's
Missing	1	
R2	0.3858	Rows with Weight
Missing	0	
Adj R2	0.3346	Rows with Y

Missing	0		
Coefficient of Variation	1.0679	Rows Used in Estimation	27
Mean Square Error	4.561221	Sum of Weights	
Square Root of MSE	2.135701	Completion Status	
Ave Abs Pct Error	403.228		

Regression Equation Section

Independent Variable	**Regression Coefficient b(i)**	**Standard Error Sb(i)**	**T-Value to test H0:B(i)=0**	**Prob Level**
Intercept	-177.8188	72.4382	-2.455	0.0217
C3	-0.0391	0.1146	-0.341	0.7361
C5	0.0032	0.0013	2.441	0.0224

All these figures show that this model is inadequate. So all these findings agree with I(1:2) and J(1:2) statistics.

11. COOK'S DISTANCES AND J(1:2) STATISTICS

Theoretically by (19), Cook's D and J(1:2) values are functionally dependent. Inverse of variance-covariance matrix that appears in (15) is a weighted average of the inverses of variance covariance matrices specified by two hypotheses (one hypothesis stating that the regression equation summarized in Table 3 is the most appropriate one and the other one obtained by deleting one specific observation at current iteration). Thus by removing one observation at each iteration one can obtain 27 "different" regression equations and 27 variance- covariance matrices. Yet a more convenient way of obtaining Cook's distances is to use (20) without dealing with the complicated calculations of inverses of variance-covariance matrices for all iterations. To summarize we have calculated Cook's distances by both ways; one by (20) and the other one by NCSS. Table 5 reflects the idea that calculated J(1:2) values Cook's D values obtained

by NCSS are highly correlated as expected a priori By (19) the ratio J(1:2) to D is 3 since there are 3 parameters in the model. The average of J(1:2)/D values in the following table is 3.00256 . Finally, we have to note that for the 10th iteration we couldn't get any ratio since Cook's D is zero. So we omitted this observation from consideration in calculating average ratio of Jeffreys' divergences to Cook's distances.

Table 5. Cook's distances and J(1:2) statistics

ith observation	**J(1,2)**	**Cook's D**	**Ratio(J(1,2) to Cook's D)**
1	0.012374	0.0041	3.018001726
2	0.012738	0.0042	3.032901292
3	0.056818	0.0189	3.006259452
4	0.145723	0.0486	2.998417686
5	0.044359	0.0148	2.997204754
6	0.068205	0.0227	3.004634336
7	0.027061	0.009	3.006739128
8	0.05869	0.0196	2.994376136
9	0.027745	0.0092	3.015728872
10	7.47E-05	0	#DIV/0!
11	0.002036	0.0007	2.908629732
12	0.004913	0.0016	3.070549
13	0.013113	0.0044	2.98033631
14	0.050247	0.0167	3.00882502
15	0.022036	0.0073	3.018696833
16	0.068671	0.0229	2.998711948
17	0.050823	0.0169	3.007306684
18	0.013278	0.0044	3.017657519
19	0.059069	0.0197	2.998410149
20	0.010989	0.0037	2.969871694
21	0.009641	0.0032	3.012889408
22	0.022605	0.0075	3.014052547
23	0.029863	0.01	2.986290209
24	1.360074	0.4534	2.999721716
25	0.11457	0.0382	2.999218021
26	2.568166	0.8561	2.999843557
27	0.005102	0.0017	3.001344649
cor(J(1,2);Cook's D)	0.999		
Average of Ratio (J(1,2) to Cook's D)	3.00256		

12. DISCUSSION

Entropy based divergence measures can easily be applied to some statistical problems. These measures can be useful in a)measuring qualitative variation b)determining the degree of association between qualitative variables, c) In Bayesian methodology, assessing the sample information especially in comparing prior and posterior conjugate distributions, c) determining the degree of goodness of fit, and e)in hypothesis testing procedures. In this study we tried to show that entropy-based statistics like Kullback-Leibler divergence and Jeffreys' divergence can also be engaged to linear regression analysis especially in hypothesis tests on population parameters and in residual analysis to determine which observations are influential on the regression equation as a whole.

REFERENCES

ASH, R.B., "*Information Theory*", Dover Publications, NY , 1990.
COVER, T.M.; THOMAS, J.A., "*Elements of Information Theory*", Wiley Interscience (Second Edition), Hoboken, New Jersey, 2006.
EVREN, A., "*Entropinin İstatistik'teki Bazı Uygulamaları*", II. Ulusal Konya Ereğli Kemal Akman Meslek Yüksek Okulu Tebliğ Günleri, 13-14 Mayıs 2010, Sayı 2, 2010, s 414-428, 2010.
EVREN, A.; "*İstatistik'te Entropiye Dayalı Uyum Ölçülerinin Diğer Uyum Ölçüleri ile Kıyaslanması*", 7. İstatistik Günleri Sempozyumu, Bildiri Tam Metinleri Kitabı, Orta Doğu Teknik Üniversitesi, Mart 2011, Ankara, s58-67 , 2011.
FOX, J., "*Regression Diagnostics*", Series : Quantitative Applications in the Social Sciences", 79, A Sage University Paper, pp 81-84 ,1991.
GRIFFITHS, W.E.,HILL, R.C., JUDGE,G.G., "*Learning and Practicing Econometrics*", John Wiley & Sons, USA, pp 361-363, 1993.
KHINCHIN, A.I., "*Mathematical Foundations of Information Theory*", Dover Publications, NY, p7 ,1957.
KULLBACK, S., "*Information Theory and Statistics*", Dover Publications, NY , p6, pp97-100, 1996.
PIERCE, J.R., "*An Introduction to Information Theory Symbols, Signals and Noise*", Dover Publications, Second, revised edition, NY , pp-21-23, p 202 , 1980.
RAO, C.R., "*Convexity Properties of Entropy Functions and Analysis of Diversity*", Proceedings of Symposium on Inequalities in Statistics and Probability, Institute of Mathematical Statistics ,Lecture Notes-Monograph Series, Shanti S.Gupta ,Series Editor, Volume 5, October 27-30 , 1982, Lincoln, Nebraska, pp 68-77 ,1984

RAWLINGS, J.O,, "*Applied Regression Analysis A Research Tool* ", Wadsworth&Brooks/Cole, Statistics/Probability Series, California, 268, 1988.
RENYI, A, "*Probability Theory*", Dover Publications, NY, p 554,2007.
RENYI, A., "*Foundations of Probability*", Dover Publications,NY, p23 ,2007.
REZA,F.M. ; "*An Introduction to Information Theory*", Dover Publications, NY, 1994.
http://stats.oecd.org/Index.aspx