BRAC University Journal, Vol. II, No. 2, 2005, pp. 35-38

INCORRECT F –STATISTIC TO TEST NONHOMOGENEOUS HYPOTHESIS IN BIVARIATE REGRESSION ANALYSIS

Mohammad Lutfur Rahman
Department of Mathematics and Natural Science
BRAC University, 66 Mohakhali C/A
Dhaka – 1212, Bangladesh

ABSTRACT

In Regression analysis, an F test can be viewed as a comparison between a full and a restricted model. The most general F formula compares the error sums of squares (SSE's) of these two models. This F formula is always correct because the SSE comparison is meaningful in all tests. Other formulas use the corrected model sum of squares (SSM) or the coefficient of determination (R^2) to compare the full and restricted models. This article gives several examples where the SSM's or R^2's of the two models cannot be compared, and hence where the use of F formulas based on SSM or R^2 would be incorrect. This problem usually arises in tests of nonhomogeneous hypotheses, although it may also appear in other situation.

Key words: Coefficient of determination; Full model; Linear model; Reparametrization; Restricted model.

I. INTRODUCTION

This article discusses tests of hypotheses with equality constraints in regression and other fixed linear models, non-linear models under the "usual" (Gauss-Markov-Normal) assumptions. It is instructive to think of a test of this type as a comparison between two models; full model (the unconstrained model) and the restricted model (the model subject to the constraints in the null hypothesis). The appropriate F statistic can be calculated as

$$F=\frac{[SSE(restricted)-SSE(full)]/v_1}{SSE(full)/v_2} \quad (1.1)$$

Where SSE denotes the error or residual sum of squares. The degrees of freedom are

$v_1=df_E$ (restricted)- df_E (full) and $v_2=df_E$ (full),

Where df_E represents the degrees of freedom associated with SSE.

If we use $\sigma\hat{}^2=SSE/df_E$ to denote the usual unbiased estimate of the error variance σ^2, Formula (1.1) can be expressed as

$$F=\frac{[df_E(restricted)*d\hat{\sigma}^2(restricted)-df_E(full)*\hat{\sigma}^2(full)]/v_1}{\hat{\sigma}^2(full)}$$

We can hence interpret the F statistic in (1.1) as comparison between two estimates of σ^2 under two competing models. This is a meaningful comparison in any test of a null hypothesis with equality constraints. If the hypothesis is true, we expect the F in (1.1) to be close to 1.

Let SST, SSM, and R^2 denote, in that order, the corrected total sum of squares, the corrected model sum of squares, and the coefficient of determination (that is, R^2=SSM/SST). Some authors (for example, Kleinbaum, Kupper, and Muller 1988,sec. 9.3.2;Meek and Turner 1983,sec.15.6.3; Myers 1986,sec.3.4) also recommend the formula

$$F=\frac{[SSE(full)-SSE(restricted)]/v_1}{SSE(full)/v_2} \quad (1.2)$$

Yet another formula seen in textbooks (Seber 1977,sec.4.2) is

$$F=\frac{[R^2(full)-R^2\ (restricted)]/v_1}{[1-R^2(full)]/v_2} \quad (1.3)$$

Although correct in the context used by these authors, Formulas (1.2) and (1.3) are not as general as (1.1). Sections 3 summarizes conditions under which formulas (1.2) and (1.3) are equivalent to (1.1).

As explained in Section 2,there may be many equivalent full and restricted models associated

with a given test. Formula (1.1) is invariant to the choice of full and restricted models, but the differences [SSM (full)-SSM (restricted)] and [R^2 (full)- R^2 (restricted)] is not. Hence, formula (1.2) and (1.3) may give incorrect results in this case, a fact that all textbooks should but most fail to emphasize. This problem usually arises in tests of nonhomogeneous hypotheses, although it may also appear in other situations.

II. EXAMPLES

Table 1:Data for Model (2.1a)

i	1	2	3	4	5	6	7	8	9	10
y_i	1	2	3	4	5	6	7	8	9	10
x_{1i}	100	200	300	400	500	600	700	800	900	1000
x_{2i}	1000	3000	1000	5000	8000	7000	8000	9000	10000	

Source: Cambridge Random Number Table,P-21

Table 1gives the observed values of the response and predictor variables of the following regression model:

$$yi = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \qquad (2.1a)$$

(i=1,2,……..10).Assume that we are interested in the parameter $\gamma=(\beta_1+2\ \beta_2-5)/2$; it may be convenient to reparametrize (2.1a) so that γ appears explicitly in the model equation. Substituting β_2 by $(2\gamma-\ \beta_1+5)/2$, in (2.1a), we obtain

$$z_{1i}= \beta_0 + \beta_1 w_{1i} + \gamma x_{2i} + \varepsilon_i \qquad (2.1b)$$

with $z_{1i}=y_i-2.5x_{2i}$ and $w_{1i}=x_{1i}-0.5x_{2i}$. Alternatively, we can substitute β_1 by $(2\gamma-\ 2\beta_2+5)$ in (2.1a) to obtain

$$z_{2i}= \beta_0 + \beta_2 w_{2i} + \gamma w_{3i} + \varepsilon_i \qquad (2.1c)$$

with $z_{2i}= y_i-5x_{1i}$, $w_{2i}=x_{2i}-2x_{1i}$, and $w_{3i}=2x_{1i}$. Models (2.1a),(2.1b),and (2.1c) are reparameterizations of each other (in the sense of Peixoto 1986),although they have different responses. As seen in Table 2, these three models have identical SSE's but different SSM's and R^2's[See Shah (1991) for a discussion on effects of reparametrizations on R^2's].

Consider the test of the nonhomogeneous hypothesis

Ho: $\beta_1 + 2\beta_2 = 5$ (2.2a)

Under the Model (2.1a). This test is equivalent to the test of the homogeneous hypothesis

Ho: $\gamma=0$ (2.2b)

Under Model (2.1b) or (2.1c). Three appropriate restricted model expressions are

$y_i= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

subject to $\beta_1 + 2\beta_2 = 5$ (2.3a)

$z_{1i}= \beta_0 + \beta_1 w_{1i} + \varepsilon_i$ (2.3b)

and $z_{2i}= \beta_0 + \beta_2 w_{2i} + \varepsilon_i$ (2.3c)

Expression (2.3b) is obtained by equating β_2 to $(5-\beta_1)/2$ in (2.1a)[or γ to 0 in (2.1b)], whereas(2.3c) is obtained by equating β_1 to $(5-2\beta_2)$ in (2.1a)[or γ to 0 in (2.1c)].

Models (2.3a), (2.3b) and (2.3c) are reparameterizations of each other and hence have identical SSE's (Peixoto 1986). Table 2 shows a curious result:

Model	Model Sum of Squares (SSM)	Error Sum of Squares (SSE)	Error Degrees of Freedom (df_E)	R^2	Comments
(2.1a)	1.669	18.331	7	.083	[a]
(2.1b)	588,105,002	18.331	7	.999	[a]
(2.1c)	20,625,002	18.331	7	.999	[a]
(2.3a)	-	3094051	8	-	[b], [d]
(2.3b)	585010969	3094051	8	.995	[b]
(2.3c)	17,530,969	3094051	8	.850	[b]
(2.6)	-	36.835	8	-	[c],[d]

Key to comments: [a] full model for Hypothesis(2.2a) and (2.4);[b] restricted model for Hypothesis(2.2a);[c] restricted model for Hypothesis(2.5);[d]SSM and R^2 are not meaningful for (2.3a) and (2.6) since these models do not satisfy equation (2.4)

R^2 is much larger for the restricted models (2.3b) and (2.3c) than for the full model (2.1a) (.850 and .995 versus .083). These three R^2 's cannot be compared, however, since Models (2.1a), (2.3b) and (2.3c) have different responses. We should instead compare the R^2's of models (2.1b) and (2.1c) to those of models (2.3b) and (2.3c). One may think that the R^2 of the original full model [i.e;(2.1a)] could be compared to that of (2.3a),

since (2.1a) and (2.3a) have the same response y_i. Unfortunately, the coefficient R^2 (as usually defined) is not meaningful for model (2.3a), as explained in the following paragraph.

Model (2.3a) has a very peculiar characteristic: the mean parameters β_o, β_1, and β_2 cannot be simultaneously equal to 0.Models with this characteristic are called nonhomogeneous or affine (Peixoto 1992). A constrained model such as (2.3a) can be fitted directly using Lagrangian multipliers (Graybill 1976,sec 6.11.1). The command RESTRICT in the procedure REG in SAS (SAS Institute Inc.1985) uses this technique. If one gives the commands

PROC REG; MODEL $Y=X_1X_2$

RESTRICT X1+2X2=5;

(with the data of table 1), SAS (Version 5) prints the following surprising ANOVA (analysis-of-variance) table:

Source	DF	Sum of Squares	Mean Square	F-Value	PROB>F
Model	1	-3094031.00	-	-	-
Error	8	3094051.00	386756.37	-	-
CTotal	9	20.00			

SAS also prints R^2=-154,701.55 and a cryptic warning about negative sums of squares. How can a sum of squares be negative? The problem is that the corrected ANOVA equation

SST=SSM+SSE (2.4)

is not satisfied in a nonhomogeneous model (because the vector of predicted values is not orthogonal to the residual vector).We would obtain SSM=3,089,844 for model (2.3a) if we used the usual formula SSM= $\sum(\hat{y}_i-\bar{y})^2$. SAS fails to adjust for this anomaly and calculates SSM as SST-SSE. The value of SSE printed, however, is correct and identical to those of models 92.3b) and (2.3c). (See table 2).

The F statistic for the test (2.2a) under model (2.1a) can be obtained from (1.1). The calculated value of (1.1) is 1,181,493.Note that this value is invariant to the choice of full model [(2.1a), (2.1b) or (2.1c)] or restricted model [(2.3a), (2.3b) or (2.3c)].

Formulas (1.2) and (1.3) are correct for some choices of full and restricted models, but incorrect- and not equivalent to each other-for other choices. The calculated value of (1.2), for example, is –223,396,256 when (2.1a) and (2.3b) are taken as the full and restricted models. The corresponding calculated value of (1.3) is –6.960.These two F values are obviously meaningless. On the otherhand, Formulas (1.2) and (1.3) are correct and equivalent to (1.1) if we use (2.1b) and (2.3b) as full and restricted model expressions.

For another example where Formulas (1.2) and (1.3) are inadequate, consider the test of

Ho: $\beta o=0$ (2.5)

Under model (2.1a)[or (2.1b) or (2.1c)]. An appropriate restricted model for this test is

$Y_i=\beta_1 x_{1i}+\beta_2 x_{2i}+\varepsilon_i$ (2.6)

Model (2.6) does not satisfy the corrected ANOVA equation (2.4), as it is the case with most models without an intercept term. [We have that $\sum(y_i-\bar{y})^2$=20, $\sum(\hat{y}_i-\bar{y})^2$=20.872, and $\sum(y_i-\hat{y})^2$=36.835 for model (2.6)]. Thus the usual definitions of SSM and $R^2$are not meaningful for this model and we cannot use Formulas (1.2) and (1.3). We can nevertheless still use Formulas (1.1) to obtain the correct F statistic.

III. SUMMARY AND CONCLUSIONS

Formulas (1.2) and (1.3) are correct iff the following two conditions are both satisfied:

(C1) the corrected ANOVA Equation (2.4) is valid for the full and restricted models, and

(C2) SST (full) =SST (restricted).

Moreover, the following three conditions, which are easier to verify, are sufficient for (C1) and (C2):

(C3) the full and restricted models are homogeneous

(C4) the full and restricted models include an unconstrained intercept parameter, and

(C5) the full and restricted models have the same response variable

A homogeneous linear model is one where all the mean parameters can be 0.Model (2.3a) shows that tha ANOVA equation may not be valid for

nonhomogeneous models. Since SST and SSM are corrected for the mean, models without intercept may also fail to satisfy(C1).Condition (C5) obviously implies(C2),and section 2 gives several examples of full and restricted models with different responses where SST(full)≠SST (restricted). [Peixoto (1992) gives additional details conditions(C1)-(C5)]

The test of a nonhomogeneous hypothesis is the most common application where either condition (C1) or (C2) is violated. However, formulas (1.2) and (1.3) may also be incorrect even if the hypothesis is homogeneous, as seen with Hypothesis (2.2b) and (2.5).

In summary,(1.1) is the most general formula for the F statistic in regression analysis. It is always correct (under the Gauss-Markov-normal assumptions), even when conditions (C1) and (C2) are not satisfied. This simple F formula is the one that every instructor and textbook should emphasize and every program should use. Formula (1.2) and (1.3) may also be useful, but students should be warned that these formulas are incorrect in some applications especially in experimental data analysis.

REFERENCES

[1] Graybill, F.A. (1976), Theory and Application of the Linear Model, North Scituate, MA: Duxbury Press.

[2] Kleinbaum, D.G.Kupper, L.L, and Muller, K.E. (1988), Applied Regression Analysis and Other Multivariate Methods, Boston: PWS-Kent.

[3] Meek, G.E; and Turner, S.J. (1983), Statistical Analysis for Business Decisions, Boston: Houghton Mifflin.

[4]Myers, R.H. (1986), Classical and Modern Regression with applications, Boston: Duxbury press.

[5]Peixoto, J.L. (1986), Testable Hypotheses in singular Fixed Linear Models, Communications in Statistics, Part A-Theory and Methods; 15, 1957-1973.

[6] SAS Institute Inc. (1985) SAS User's Guide: Statistics

[7] Seber, G.A.F.(1977), Linear Regression Analysis, New York: John Wiley.

[8] Shah, A.K. (1991), Relationship Between the coefficients of Determination of Algebraically related Models.