British Journal of Applied Science & Technology
11(2): 1-7, 2015, Article no.BJAST.19598
ISSN: 2231-0843

SCIENCEDOMAIN *international*
www.sciencedomain.org

SCIENCEDOMAIN

Test to Check the Equality of Regression Models and a Medical Application

Sérgio Ricardo Silva Magalhães[1*]

[1]Universidade Vale do Rio Verde, Av. Amazonas, 1700 – Belo Horizonte /Minas Gerais,
CEP: 30.180-001, Brasil.

Author's contribution

The sole author designed, analyzed and interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/BJAST/2015/19598
Editor(s):
(1) Qing-Wen Wang, Department of Mathematics, Shanghai University, P.R. China.
Reviewers:
(1) Anonymous, Institute of Business Management Karachi, Pakistan.
(2) Anonymous, University of Tsukuba, Japan.
Complete Peer review History: http://sciencedomain.org/review-history/10431

Original Research Article

Received 18th June 2015
Accepted 24th July 2015
Published 5th August 2015

ABSTRACT

In this study, we considered the Model Identity and the Dummy Variables methods used to compare regression models. The adjustment of *h* linear regression equations was considered to verify the equality of the regression models by data simulation. Using features from the Interactive Matrix Language (IML) from the SAS® system, appropriate routines were developed for the methodology of regression models comparison. A data simulation with 100,000 experiments was performed considering different sample sizes (10, 50 and 100 observations). The performances of the two methods were essentially equivalent when comparing the different sample sizes. The results from all cases simulated by the methods had low percentages of Type I and Type II error rates. For larger samples, Type I and Type II error rates were always lower when using the approximate *F* statistics, which must therefore be the method of choice. The Dummy Variables method was the most efficient for all three sample sizes because it exhibited the lowest Type I and Type II error rates.

Keywords: Linear regression models; data simulation; hypothesis tests.

**Corresponding author: E-mail: sergio.magalhaes@unincor.edu.br;*

1. INTRODUCTION

Linear regression models have applications in many different fields of knowledge [1].

A linear model is often used because of its ease in describing the approximate relationship [2].

Regression analysis is often used to determine whether the equations from a set of *h* adjusted equations are identical, i.e., whether the phenomenon studied can be represented by a single equation [3].

In medical data, the dependent variable *Y* and the set of regressive variables X_i, *i* = 1, 2,..., *n* are usually measured in a set composed of different groups to compare how they differ depending on the relationship between X_i and *Y* [4]. This analysis can be performed by developing regression models for each group and then determining whether the corresponding equations are parallel, have a common intercept or are identical [5].

Many authors have reported methods for testing hypotheses concerning the equality of linear models [6,7,8,9,10].

There are many methods of comparing regression equations; among these, the Model Identity [11] and Dummy Variables [12] (binary) methods are the most prominent.

Thus, this study aimed to evaluate the Model Identity and Dummy Variables methods of comparing linear regression equations by data simulation and to determine whether there are differences between these methods and their practical applications.

2. METHODOLOGY AND RESULTS

2.1 Statistical Model

Initially, the fit of the observational data relative to *h* groups was considered. The following linear regression model was fitted to each of them:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \qquad (1)$$

where

y_i : i^{th} value of the response variable, *i*= 1, 2, ..., *N* observations;

y_i x_{ki} : i^{th} value of the k^{th} explanatory variable, *k* = 1, 2, ..., *K* variables;

β_k : model parameters;

ε_i : random errors.

In matrix notation, the model assumes the following form:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \qquad (2)$$

where

$\mathbf{y}$: *N* x 1 vector of observations, *N* being the number of observations;

$\mathbf{X}$: N x (*K* + 1) matrix of explanatory variables, *K* being the number of explanatory variables;

$\boldsymbol{\beta}$:(*K* + 1) x 1 parameter vector, (*K* + 1) being the number of parameters;

$\boldsymbol{\varepsilon}$: *N* x 1 vector of random errors.

To estimate the $\boldsymbol{\beta}$ parameter vector, the least squares and the maximum likelihood methods, which lead to the same estimators, are commonly employed.

According to the error assumptions, there are variations in the method of least squares estimation for the linear regression model, regarding the several forms that the variance and covariance matrix can adopt.

These variations are known as the ordinary, weighted and generalized least squares methods.

In fitting a model by the ordinary least squares method, it is assumed that the average error is null ($E(\varepsilon_i) = 0$); the error variance, ε_i, $i = 1,2,...,n$, is constant and equal to σ^2; and the error of an observation is not correlated with the error of another observation, i.e., $E(\varepsilon_i \varepsilon_j) = 0$, for $i \neq j$ and the errors are random variables with normal distributions [6].

Based on the ordinary least squares method, a $\boldsymbol{\beta}$ vector is estimated under the condition that the residual sum of squares is minimized. The quadratic function *Z*, which is the residual sum of the squares, is

$$\mathbf{Z} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \boldsymbol{\beta}\mathbf{X})' \ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \qquad (3)$$

By taking the partial derivative relative to $\boldsymbol{\beta}$, the following system of normal equations is obtained [7]:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \qquad (4)$$

As the matrix $\mathbf{X}$ has full column rank, $\mathbf{X'X}$ is a positive definite matrix and, thus, $\mathbf{X'X}$ is nonsingular. Therefore, the inverse matrix $(\mathbf{X'X})^{-1}$ exists, and the solution for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y} \tag{5}$$

This unique solution corresponds to the non-linear unbiased estimator with the minimum variance for $\boldsymbol{\beta}$.

2.2 Model Identity Method

The Model Identity method is a very general test that verifies the equality of two linear regressions. Its algorithm proceeds through the following steps [11]:

1. Given the following linear relationships:

$$\begin{aligned} y_{1i} &= a_1 + b_1 x_{1i} + e_{1i} \qquad i = 1,\ldots,n_1 \\ y_{2i} &= a_2 + b_2 x_{2i} + e_{2i} \qquad i = 1,\ldots,n_2 \end{aligned} \tag{6}$$

relative to two observation sets.

2. All the $n_1 + n_2$ observations are combined, and the least squares estimates of a and b are calculated in the combined regression $y = a + bx + e$. From this equation, the residual sum of squares (S_1) is obtained with $n_1 + n_2 - p$ degrees of freedom, in which p is the number of parameters to be estimated. In this case, p = 2.

3. The residual sum of squares for each of the two equations, i.e., S_2 and S_3, is obtained with $n_1 - p$ and $n_2 - p$ degrees of freedom, respectively. These two residual sums of squares are added, i.e., $S_4 = S_2 + S_3$, as are their degrees of freedom, i.e., $n_1 + n_2 - 2p$.

4. $S_5 = S_1 - S_4$ is obtained.
5. The F statistic is calculated as follows:

$$F_c = \frac{S_5 / p}{S_4 / (n_1 + n_2 - 2p)} \tag{7}$$

with p and $n_1 + n_2 - 2p$ degrees of freedom.

If F_c > the *F* value from the table for a given α significance level, the hypothesis that $a's$ and $b's$ parameters are the same for both observation sets is rejected.

2.3 Dummy Variables Method

The inclusion of additive or multiplicative dummy variables makes it possible to determine whether two linear equations differ in intercept, slope or both.

Given the following relation, relative to two sets [12],

$$\begin{aligned} & y_i = a_0 + a_1 D + a_2 x_i + a_3 (Dx_i) + e_i \\ & i = 1,\ldots,(n_1 + n_2) \end{aligned} \tag{8}$$

where *D* = 1 for observations from the first set (*n_1* observations), and *D* = 0 for observations from the second set (*n_2* observations).

The binary variables were introduced as additive and multiplicative. The *a_1* and *a_3* coefficients are the differences in the intercepts and the slopes, respectively.

If H_0: *a_1* = 0 is rejected, i.e., *a_1* is significant, then the intercept value of the first set is obtained by *a_1* + *$a_{0.}$* In this case, *a_0* is the intercept of the second set. If H_0: *a_1* = 0 is not rejected, i.e., *a_1* is not significant, then *a_0* represents the common intercept for both sets.

If H_0: *a_3* = 0 is rejected, then the slope value from the first set is *a_2* + *a_3*. In this case, *a_2* is the slope of the second set.

If H_0: *a_3* = 0 is not rejected, then *a_2* represents the common slope for both sets.

2.4 Methods Simulation

A data simulation composed of 100,000 experiments, each one with 10, 50 or 100 observations, was performed.

For each experiment, simple linear regression models were developed in which the values of the independent variables were obtained on a closed interval from 0 to 10, randomly, by the RANUNI function of the SAS® system [13].

To generate the residues for each model, their variance was estimated [14]. Setting the R^2 coefficient of determination to 90%, and knowing the relationship $R^2 = \frac{\delta^2_{\mathrm{mod}el}}{\delta^2_{\mathrm{mod}el} + \delta^2_{error}}$, in which $\delta^2_{\mathrm{mod}el}$ corresponds to the variance values of the dependent variables, the variance of residuals δ^2_{error} was estimated [15].

Once the variance of residuals δ^2_{error} was estimated, the RANNOR function of the SAS® system generated the random residuals for each model. These residuals are supposedly independent and normally distributed, with a zero mean and common variance, i.e., ε_{hi} ~ NID (0, δ^2_{error}) [16].

Based on the regression models considered and setting the parameters of each model for each of the situations described above to compare the methods, the Model Identity and Dummy Variables methods were computationally implemented by the IML module of the SAS® system.

2.5 Results

To evaluate the methods, four linear regression cases were considered: these were represented by (a) the most general case, when all coefficients are different; (b) parallel regressions, wherein the slopes are equal but the intercepts are different; (c) concurrent regressions, wherein the intercepts are equal but the slopes are different; and (d) coincident regressions, wherein all the lines coincide.

The results were analyzed based on the FREQ procedure of the BASE module from the Statistical Analysis System (SAS), and the frequencies of the results were determined for the nominal significance levels. These results were found for the values of the *F* test in the models for sample sizes of 10, 50 and 100, respectively.

The evaluation of the Model Identity and the Dummy Variables methods was based on the 5% nominal level for the rates of Type I error, which lies in the rejection of a hypothesis H_0 regarded as true, and on the rates of Type II error, which lies in the non-rejection of an initial hypothesis H_0, regarded as false.

Table 1 illustrates all simulated situations using both of the methods under study, showing a combination of frequencies of Type I and Type II errors.

The Model Identity and the Dummy variables methods indicate very similar results due to the low rates of Type I and Type II errors.

In general, higher rates of the combination of Type I and Type II errors were perceived when the sample size was 10 observations, with an apparent advantage to the Model Identity method.

Reduced Type I and Type II error rates were expected with an increased number of observations. This expectation usually occurred, showing better efficiency of the methods for larger sample sizes. For example, for the Dummy Variables method, lower rates were found with a sample size of 100 observations. In general, samples with 100 observations showed lower error rates, but these values are not much different from those of the other sample sizes.

In all of the cases studied, evidence that the three methods studied have good accuracy was observed, given the low percentages of the Type I and Type II error rates. However, it should be noted that a lower probability of Type I and Type II errors was obtained for the Dummy Variables method.

2.6 Example with Real Data

It was deemed necessary and appropriate to present a numerical example to illustrate the results obtained in this study.

Table 1. Frequency distribution of errors type I and type II for the methods used

	Methods					
Cases	**Identity of models (Number of observations)**			**Dummy variables (Number of observations)**		
	10	**50**	**100**	**10**	**50**	**100**
a	2811	1126	1219	2415	1001	1003
b	1342	487	219	1083	308	115
c	1312	448	371	1084	487	242
d	1053	1002	18	3101	185	12
	6518	3063	1827	7683	1981	1372
		11408			9055	

Thus, based on the method used in Table 1 calculations were performed to illustrate the methods. The data analyzed were collected between 2009 and 2010 from a sample of blood donors from the Blood Center of the Mário Penna University Hospital from the University of Vale do Rio Verde in Belo Horizonte. The donors included both males and females.

To compare the proposed methodologies, regression lines were fit for systolic blood pressure versus age, for a sample of 1,500 men and 1,500 women, to determine whether these variables have a similar linear relationship for both sexes.

In industrialized countries, the average blood pressure of the population increases with age. After age 50, systolic pressure tends to rise with increasing age, resulting in systolic hypertension. Thus, increased systolic pressure is well established as a cardiovascular risk factor [17].

Therefore, this application attempted to confirm the theories that systolic blood pressure increases continually with age in both sexes [18].

Through the SAS® program for statistical analysis, the following cases were considered:

a) Different intercepts and equal slopes;
b) Equal intercepts and different slopes;
c) Different intercepts and slopes; and
d) Equal intercepts and slopes.

The two methods for comparing linear regression models were explored, and hypothesis tests were applied to identify the above situations.

To apply the Model Identity method, the lines were first fit for each sex:

Male: $\hat{Y}_{mal} = 99.81 + 0.48x$

Female: $\hat{Y}_{fem} = 105.14 + 0.37x$

and the parameter estimates for both sexes were recorded in Table 2.

For the Dummy Variables method, a regression model of the whole set was fitted and then separated, producing a model for each sex through the inclusion of dummy variables.

$$D = \begin{cases} 0, & \text{If the individual is male.} \\ 1, & \text{If the individual is female.} \end{cases} \tag{9}$$

Main line:

$$\hat{Y} = 100.11 + 0.52x + 12.67D - 0.04xD$$

Adjusted line - Male:

$$\hat{Y}_{mal} = 100.01 + 0.52x \;\; (D = 0)$$

Adjusted line - Female:

$$\hat{Y}_{fem} = 113.41 + 0.49x \;\; (D = 1)$$

Table 2. Estimate for the parameters of the estimated models, age versus systolic pressure

Group	$\hat{\beta}_0$	$\hat{\beta}_1$	$\bar{x}$	s_x^2	$s_{Y/X}^2$
Masculine	99,81	0,48	31,08	105,21	328,25
Feminine	105,14	0,37	31,05	115,44	254,81

Table 3 shows the analysis of variance of the line fit for this situation.

For the Model Identity method, cases in which the estimated lines fit the parallelism and equality tests of the parameters were identified and are detailed below.

a) Parallelism test:

$$H_0 : \beta_{1mal} = \beta_{1fem}$$

$S_{P,Y/X}^2 = 301.25$ and $S^2_{\hat{\beta}_{1mal} - \hat{\beta}_{1fem}} = 0.04$.

Table 3. ANOVA by the dummy variables method for the variable age versus systolic pressure

Variation source	GL	SQ	QM	F
Regression (x)	1	77071,12	7707,12	20,60
Residual	3005	1124445,00	374,19	
Regression (x,D)	2	926547,00	463273,5	245,81
Residual	3004	1428954,00	475,70	
Regression (x,d,xD)	3	155768,00	51912,67	165,84
Residual	3003	926158,00	308,41	

The test statistic was $T = 0.61$. For this statistic, the critical bilateral value given by the p-value was $2P\ (T \geq |0.61|) = 0.55$. Considering an α nominal significance level of 5%, it was observed that the *p*-value > α. Therefore, the null hypothesis was not rejected; i.e., there was sufficient sample evidence not to reject the parallelism hypothesis.

b) Intercept equality test:

$$H_0 : \beta_{0mal} = \beta_{0fem}$$

$S^2_{P,Y/X} = 301.25$ and $S^2_{\hat{\beta}_{0mal} - \hat{\beta}_{0fem}} = 5.01$.

The test statistic was $T = -5.61$. For this statistic, the critical bilateral value given by the *p*-value was $2P\ (T \geq |-5.09|) \cong 0$. Therefore, the null hypothesis was rejected for all α nominal significance levels. There was strong sample evidence that the hypothesis of equality of intercepts is not true. Fig. 1 shows that women had higher systolic blood pressure regardless of age, considering the estimated line parallelism.

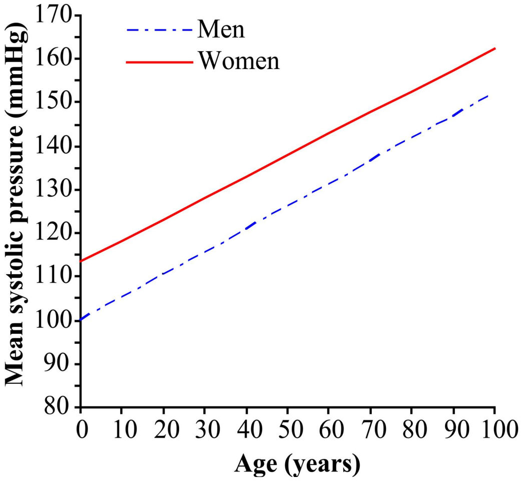

Fig. 1. Sistolic Arterial blood pressure rate (mmHg) versus age in years

This result is consistent with a large study in Paris that involved 77,023 men and 48,480 women and correlated the risk of systolic and diastolic hypertensions with the patients' age. One of the conclusions of this study was that women had higher systolic hypertension than men [18].

In contrast, using the Dummy Variables method on cases in which the estimated lines fit the parallelism test, the parameter equality test and the coincidence test were identified and are detailed below.

a) Parallelism test:

$$H_0 : \beta_3 = 0$$

The test statistic was $F(XD/X,D) = 0.52$. The *p*-value with 1 and 3,003 degrees of freedom was equal to 0.46. Therefore, the null hypothesis H_0 was not rejected for any nominal values of α, and there was no sample evidence for rejecting the hypothesis of parallelism of linear regressions.

b) Intercept equality test:

$$H_0 : \beta_2 = 0$$

The test statistic was $F(D/X,XD) = 253.25$. The *p*-value with 1 and 3,003 degrees of freedom was approximately zero. The null hypothesis H_0 was therefore rejected for any nominal values of α different from zero. Thus, sample evidence was found that the hypothesis of equal intercepts for the linear equations of both sexes was not true.

c) Coincidence test:

$$H_0 : \beta_2 = \beta_3 = 0$$

The test statistic was $F(D/X) = 121.68$. The *p*-value with 2 and 3,003 degrees of freedom was < 0.001. Therefore, the null hypothesis H_0 was rejected for any nominal values of α different from zero. Therefore, sample evidence for the hypothesis of the coincidence of estimated linear regressions for both sexes were not found.

3. CONCLUSIONS

The sample data for systolic blood pressure and age, subjected to both methods under study, have shown that the estimated lines for males and females were not coincident. They were parallel, with different intercepts and had the form $Y = \beta_0 + \beta_1 x + \varepsilon$.

The application of the Model Identity method was equivalent to the application of the Dummy Variables method.

However, for the simulated situations for each of the three sample sizes, the Dummy Variables method proved to be more efficient than the Model Identity method because the former had the lowest percentage of Type I and Type II errors.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Armitage P, Berry G. Statistical methods in medical research. 6th ed. Blackwell. Oxford; 2011.
2. Hoffmann R, Vieira S. Regression analysis: An introduction to econometrics. 5rd ed. São Paulo. Hucitec; 2009.
3. Seber GAF. Linear regression analysis. New York. John Wiley; 2007.
4. David CS, Hall DB. A computer program for the regression analysis of ordered categorical repeated measurements. Computer Methods and Programs in Biomedicine. 2011;51:153-169.
5. Copenhaver MD, Holland BS. Computation of the distribution of the maximum studentized range statistic with application to multiple significance testing of simple effects. Journal of Statistical Computing and Simulation. 2012;30:1-15.
6. Chow GC. Tests of equality between sets of coefficients in two linear regressions. Econometrica. 1960;28:591-605.
7. Cordeiro GM, Paula GA. Regression models for univariate data analysis. Impa. Rio de Janeiro; 1989.
8. Neter J, Kutner MH, Nachtsheim CJ, Wasseman W. Applied linear statistical models. 4th ed. Richard D. Irwin. Chicago; 2011.
9. Gujarati D. Use of dummy variables in testing for equality between sets of coefficients in two linear regressions: A note. The American Statistician. 1970;24: 50-52.
10. Ratkowsky DA. Nonlinear regression modeling: A unified practical approach. Marcel Dekker, New York; 2010.
11. Graybill FA. Theory and application of the linear model. Duxbury Press. Belmont; 1976.
12. Gujarati D. Use of dummy variables in testing for equality between sets of coefficients in two linear regressions: A note. The American Statistician. 1970; 24:50-52.
13. SAS® Institute. SAS Procedures guide for computers. 13th ed. SAS® Institute. Cary, NC. 2012;3.
14. Verbeke G, Molenberghs G. Linear mixed models in practice: A SAS-oriented approach. SAS Institute. Cary, NC; 2011.
15. Littell RC, Henry PR, Ammerman CB. Statistical analysis of repeated measures data using SAS procedures. The Journal of Animal. Science. 2010;76:1216-1231.
16. Brown RL, Durbin J, Evans JM. Techniques for testing the constancy of regression relationships over time. Journal of the Royal Statistical Society. Series B, Statistical Methodology. 1975;37:149-192.
17. Sesso HD, Stampfer MJ, Rosner B, Hennekens CH, Gaziano JM, Manson JE, Glynn RJ. Systolic and diastolic blood pressure, pulse pressure, and mean arterial pressure as predictors of cardiovascular disease risk in men. Hypertension. 2010;36:801-807.
18. Franklin SS, Khan SA, Wong ND, Larson MG, Levy D. Is pulse pressure useful in predicting risk for coronary heart disease? The Framingham Heart Study. Circulation. 2010;100:354-360.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/10431