Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Comparison of regression models in case of non-normality and Heteroscedasticity

Seray Kahvecioglu, Berna Yazici *

Anadolu University, Science Faculty, Department of Statistics, Eskisehir, Turkey

ARTICLE INFO

Article history:
Received 4 July 2016
Received in revised form
14 August 2016
Accepted 14 August 2016

Keywords:
Generalized p values
Generalized p values in regression
Regression analysis in case of
Assumption violation
Comparison of regression coefficients

ABSTRACT

In regression analysis, in case of comparing two regression models and coefficients where the distribution of variables in question is not known, generalized p values may be used. The generalized p value is an extended version of the classical p value which provides only approximate solutions. Use of approximate methods, generalized p value, has better results, performance with small samples. In this study, the generalized p value –which may be used alternatively when different assumptions aren't fulfilled - is researched theoretically; a simulation is conducted and an application in regression analysis is given. It is concluded that in generalized p value works well for the comparison of regression coefficients both under non-normality and heteroscedasticity.

1. Introduction

Comparing two or more regression lines is a common problem in statistical application studies. Assumption violation has been a popular topic which has been studied by numerous researchers. Generalized p value theory is one of them that are commonly used particularly for non-normality and heteroscedasticity; (Tsui and Weerahandi, 1989; Weerahandi, 2013; Sezer et al., 2015) comparing the regression models is frequently in question when researcher aims to see the change in regression equation or in regression parameters. In this study, generalized p value theory which is used in case of assumption violation is examined. Especially the homogeneity of regression models is examined under heteroscedasticity. A simulation study is conducted in case of assumption violation for comparison of regression models. Also an application is given for a small sample for assumption violation.

2. Data and methods

The concept of generalized test variable and generalized p value is firstly proposed by Tsui and Weerahandi (1989). In that commonly used method, when the distribution of the variables' is unknown the p-value is being calculated via iterative calculations.

2.1. Comparison of two regression models with homogenous variances

For two regression models, while some of the parameters are different, some of them maybe common. Hence let's, take into account the following linear regression model:

$$y_j = U_j c + V_j d_j + W_j t_j + \varepsilon_j, \qquad j = 1,2 \tag{1}$$

Suppose that the normality assumption is valid;

$$\varepsilon_i \sim N\left(0, \sigma_j^2 I_j\right), \qquad i = 1,2,\ldots,n \tag{2}$$

it is assumed that $\sigma_1^2 = \sigma_2^2$ for I_j, $j = 1,2,$. The lengths of vectors y_1 and y_2 are defined by the dependent variables with n_1 and n_2 observations. We will denote the explanatory variables with matrices U_j, V_j and W_j. While the variables are common for the matrixes U_j and W_j, the variables for V_1 and V_2 are different from each other. For the response vectors:

$$y' = (y_1', y_2') \tag{3}$$

and the design matrices:

* Corresponding Author. Tel.: +90 536 922 24 34
Email Addresses: s.mankir@hotmail.com (S. Kahvecioglu), bbaloglu@anadolu.edu.tr (B. Yazici)

http://dx.doi.org/10.21833/ijaas.2016.08.004
2313-626X/© 2016 The Authors. Published by IASE.

$$X_{12} = \begin{bmatrix} U_1 & 0 & V_1 & 0 & W_1 \\ 0 & U_2 & 0 & V_2 & W_2 \end{bmatrix} \quad (n_1 + n_2) \times \left(2p_c + p_{d_1} + p_{d_2} + p_t\right) \tag{4}$$

and

$$X_{1,2} = \begin{bmatrix} U_1 & 0 & V_1 & 0 & W_1 & 0 \\ 0 & U_2 & 0 & V_2 & 0 & W_2 \end{bmatrix} \quad (n_1 + n_2) \times \left(2p_c + p_{d_1} + p_{d_2} + 2p_t\right) \tag{5}$$

1, 2 are used for $t_1 \neq t_2$while, 12 is used for $t_1 = t_2$. Where c is known parameter vector and the length of it is p_c. The lenghts of d_1 and d_2 are $p_{d_1}$and p_{d_2}, respectively. These also include the parameters that are expected to be different. In other words, they are the coefficients of the variables that consist the matrixes of V_1 and V_2 which are coming from different distributions. In order to test the equality of t_1 and t_2 vectors, denoting the lengths of the vectors as p_t.

Under $H_0\colon t_1 = t_2$, we test the equality of coefficients of two different regression models.

The sum of squares residual of the matrixes are s_{12}^2, $s_{1,2}^2$ respectively. If there is not common parameters of the regression model, sum of squares residual is calculated by summing the sums of squares residual separately; $s_{1,2}^2 = s_1^2 + s_2^2$.

$$F = \frac{(s_{12}^2 - s_{1,2}^2)/p_t}{s_{1,2}^2/l} \sim F_{p_t,l} \tag{6}$$

$$l = (n_1 + n_2) - \left(2p_c + p_{d_1} + p_{d_2} + 2p_t\right) \tag{7}$$

$$p = 1 - H_{p_t,l}\left[\frac{(s_{12}^2 - s_{1,2}^2)/p_t}{s_{1,2}^2/l}\right] \tag{8}$$

$H_{k,l}$ is the cumulative distribution function of F distribution with degrees of freedom k and l. If $p < \alpha$, then H_0 is rejected.

2.2. Comparison of two regression models without common parameters

While the comparison of two regression models, if the unknown parameter situation is in question, in other words, let us c = 0 in the model. In order to get the unbiased tests depending on the generalized p value, $\tilde{y}_j = a_j^{-1} y_j$, $\tilde{V}_j = a_j^{-1} V_j$ and $\widetilde{W}_j = a_j^{-1} W_j$. $a_1^2 = s_1^2/R$ and $a_2^2 = s_2^2/(1-R)$ (where R is Beta distributed parameter.)

$$\tilde{X}_{12} = \begin{bmatrix} \mathrm{U}_1 & 0 & \tilde{V}_1 & 0 & \widetilde{W}_1 \\ 0 & \mathrm{U}_2 & 0 & \tilde{V}_2 & \widetilde{W}_2 \end{bmatrix} (n_1 + n_2) \times (2p_c + p_{d_1} + p_{d_2} + p_t), \tag{9}$$

and

$$\tilde{X}_{1,2} = \begin{bmatrix} U_1 & 0 & \tilde{V}_1 & 0 & \widetilde{W}_1 & 0 \\ 0 & U_2 & 0 & \tilde{V}_2 & 0 & \widetilde{W}_2 \end{bmatrix} \quad (n_1 + n_2) \times \left(2p_c + p_{d_1} + p_{d_2} + 2p_t\right) \tag{10}$$

In order to test the hypothesis, the p value is obtained as follows (Weerahandi, 2013):

$$p = 1 - E_R\left\{H_{k,l}\left[\frac{l}{k}\left(\tilde{s}_{12}^2\left(\frac{s_1^2}{R}, \frac{s_2^2}{(1-R)}\right) - 1\right)\right]\right\} \tag{11}$$

$$\tilde{s}_{12}^2,\ a_1^2 = s_1^2/R \text{ ve } a_2^2 = s_2^2/(1-R)\ k = p_t,\ l = n_1 + n_1 - \left(2p_c + p_{d_1} + p_{d_2} + 2p_t\right) \tag{12}$$

where $H_{k.l}$ is the cumulative distribution function of F distribution with degrees of freedoms k and l, and the expected values are obtained from Beta distribution as follows (Weerahandi, 2013):

$$R \sim Beta\left(\frac{n_1 - 2p_t - p_{d_1}}{2}, \frac{n_2 - 2p_t - p_{d_2}}{2}\right) \tag{13}$$

Beta distributed R will be used to obtain the p value in order to compare the regression coefficients.

2.3. Comparison of two regression models with heterogeneous variances

Two regression models are defined as follows. Let us assume their variances are not equal ($\sigma_1^2 \neq \sigma_2^2$) and the sample size is moderate.

$$y_j = U_j c + V_j d_j + W_j t_j + \varepsilon_j, \quad j = 1, 2 \tag{14}$$

$$\tilde{X}_{12} = \begin{bmatrix} \widetilde{U}_1 & 0 & \tilde{V}_1 & 0 & \widetilde{W}_1 \\ 0 & \widetilde{U}_2 & 0 & \tilde{V}_2 & \widetilde{W}_2 \end{bmatrix}, \quad (n_1 + n_2) \times (2p_c + p_{d_1} + p_{d_2} + p_t) \tag{15}$$

and

$$\tilde{X}_{1,2} = \begin{bmatrix} \widetilde{U}_1 & 0 & \tilde{V}_1 & 0 & \widetilde{W}_1 & 0 \\ 0 & \widetilde{U}_2 & 0 & \tilde{V}_2 & 0 & \widetilde{W}_2 \end{bmatrix}, \quad (n_1 + n_2) \times (2p_c + p_{d_1} + p_{d_2} + 2p_t) \tag{16}$$

In order to compare the regression coefficients of different regression models ($H_0\colon t_1 = t_2$), the test variable given below is calculated:

$$T = \frac{(\tilde{S}_{12}^2 - \tilde{S}_{1,2}^2)(\sigma_1^2, \sigma_2^2)}{(\tilde{s}_{12}^2 - \tilde{s}_{1,2}^2)(s_1^2 \sigma_1^2 / S_1^2, s_2^2 \sigma_2^2 / S_2^2)} \tag{17}$$

The p-value can be obtained using Beta distributed variable:

$$p = 1 - E_{R_1,R_2}\left\{H_{k,l}\left[\frac{l}{k} R_2 \tilde{s}^2 \left(\frac{s_1^2}{R_1}, \frac{s_2^2}{(1-R_1)}\right)\right]\right\}$$

$$R_1 \sim Beta\left(\frac{n_1 - p_c - p_t - p_{d_1}}{2}, \frac{n_2 - p_c - p_t - p_{d_2}}{2}\right),$$

$$R_2 \sim Beta\left(\frac{n_1 + n_2 - 2p_c - 2p_t - p_{d_1} - p_{d_2}}{2}, \frac{p_c}{2}\right).$$

$$y_j = U_j c + V_j d_j + W_j t_j + \varepsilon_j, \quad j = 1,2 \tag{18}$$

Generalized p value is calculated in the comparison of regression coefficients, if the distributions of the variables which are used in the regression models are not known or when the homoscedasticity assumption is not valid. For the simulation studies, R is used. In order to conduct simulations, 8-unit samples are chosen from certain distributions. Those samples are used to obtain generalized p values with 500 replications using the design matrix. The obtained p values are given in the following tables. R is used to conduct the simulation study.

Let us assume two regression models as given below and we are interested in comparing them. While V_1 in the first model is Normal (0, 2.5), and W_1 is Gamma (2, 2) distributed, V_2 in the second model is Weibul (1, 1.5), and W_2 is Gamma (2, 2) distributed.

$$y_j = U_j c + V_j d_j + W_j t_j + \varepsilon_j, \quad j = 1,2$$

We suppose that those two models do not have the common parameters. The generalized p values after 500 replications are given in Table 1.

When Table 1 is examined, it can be seen that, only 2 values are smaller or closed to nominal value 0.05. The null hypothesis can be accepted for remaining 98 values. In other words, according to Table 1, there is no significant difference between the two regression models.

The simulation scenario is changed for the comparison of two regression models. As can be seen from Table 2 that, in the first model V_1 is Normal (0, 2.5) distributed and W_1 is Gamma (9, 0.5) distributed. On the other hand in the second model, similarly V_2 is Normal (0, 2.5) and W_2 is Gamma (9, 0.5) distributed.

When the simulation results are examined for the second scenario in Table 2, it can be seen that only 2 values are smaller than 0.05 or closed to it. As the result of the others, null hypothesis H_0 should be accepted. In other words, the difference the regression models constructed for Table 2 are not statistically significant.

In order to construct the table to see the power of the test, the first models' variables V_1 and W_1 are chosen from Gamma(2,2) and the second models' variables V_2 and W_2 are chosen from Beta(2,5) distributed populations. The two regression models, with 8 units and without any common parameter, are compared with 10,000 replications using R. The results are summarized in Table 3.

It can be seen from Table 3 that almost all values are closed to 1, which means that, the test to compare the regression models from different distributions is powerful.

3. Application

A dietitian is applying a program for 3 months. After the program, the weight loss (kg.) of the subjects' are noted and modeled also taking into account their ages (year), heights (cm.) and gender. 16 subjects, 8 males and 8 females, are randomly selected after the weight loss diet. The data set is shown in Table 4.

Model form for two genders:

$$Y = \alpha + \beta Heigth + \gamma Age + \varepsilon$$

The least squares estimates of the parameters of the two regressions estimated using the above data are respectively:

$$\hat{\alpha}_M = -10.441 \,, \hat{\beta}_M = 0.110 \,, \hat{\gamma}_M = 0.031$$

and

$$\hat{\alpha}_F = -39.969 \,, \hat{\beta}_F = 0.293 \,, \hat{\gamma}_F = -0.037$$

Table 1: Number of observations 8, Variable $V_1 \sim$ N (0, 2.5), $V_2 \sim$ Weibull (1, 1.5), $W_1$and $W_2 \sim$ Gamma (2, 2) (Generalized p values for 500 replicates)

Generalized p values for 500 replicates				
0.8563	0.9484	0.8493	0.9247	0.9790
0.5150	0.3005	0.6902	0.1937	0.4306
0.5561	0.9829	0.7998	0.7484	0.3430
0.7751	0.4706	0.9430	0.3357	0.9444
0.6093	0.5218	0.1002	0.1016	0.2897
0.0837	0.7206	0.3605	0.4529	0.4019
0.7548	0.4253	0.5839	0.6500	0.7650
0.9367	0.5483	0.7634	0.3068	0.1265
0.6241	0.5555	0.3008	0.4745	0.6148
0.0863	0.9463	0.3758	0.1375	0.9144
0.6310	0.8814	0.2994	0.4900	0.9682
0.0368	**0.0186**	0.8868	0.7659	0.1246
0.1015	0.6196	0.6317	0.2656	0.5202
0.7911	0.4672	0.3707	0.6429	0.3557
0.5453	0.6247	0.6601	0.9529	0.8344
0.8907	0.3795	0.6434	0.7684	0.8592
0.4738	0.7820	0.6650	0.8782	0.0842
0.2432	0.4913	0.8121	0.9177	0.7930
0.9265	0.2534	0.4391	0.3391	0.7036

Table 2: V1 ~ Normal (0, 2.5), W1 ~ Gamma (9, 0.5) V2 ~ Normal (0, 2.5), W2 ~ Gamma (9, 0.5) (Generalized p values for 500 replicates)

Generalized p values for 500 replicates				
0.5855	0.8378	0.1099	0.5307	0.9569
0.9666	0.1453	0.3514	0.4958	0.5149
0.5183	0.6842	0.4408	0.6349	0.7684
0.3429	0.2207	0.4353	0.1939	0.5287
0.7931	0.1277	0.3449	0.1648	0.5794
0.2190	0.5311	0.0818	0.3802	0.6908
0.6637	0.5502	0.6817	0.8831	0.6387
0.9748	0.3979	0.5839	0.1657	0.1899
0.4438	0.4581	0.9608	0.6348	0.2271
0.6384	0.2654	0.2339	0.3753	0.6532
0.5094	0.5385	0.5070	0.1233	0.5852
0.7218	0.2046	0.5413	0.7129	0.1634
0.2456	0.3525	0.7753	0.9282	0.4027
0.3627	0.8557	0.6640	**0.0402**	**0.0460**
0.8611	0.1410	0.5164	0.3807	0.6450
0.2513	0.7480	0.9542	0.1612	0.9121
0.3221	0.4979	0.1229	0.4538	0.7499
0.4753	0.5583	0.2122	0.1015	0.7901
0.9552	0.5857	0.2674	0.7621	0.8563
0.3709	0.2110	0.1607	0.4079	0.9283

Table 3: V1~ Gamma (2, 2), V2~ Beta (2, 5), W1~ Gamma (2, 2) and W2~Beta (2, 5) (Power of the test for 10000 replicates)

Power of the test for 10000 replicates				
0.6500	0.7083	0.5832	0.8457	0.8565
0.9693	0.8446	0.9105	0.8455	0.9759
0.8386	0.9040	0.9690	0.9178	0.9780
0.8052	0.8184	0.9995	0.8123	0.9943

Table 4: Data set for weight loss application

	MALE							
Age	33	47	21	65	27	57	45	25
Height	180	176	196	180	179	176	177	165
Weigth Loss	11.3	15.6	11.8	11	8.9	8.8	8.7	8.1
	FEMALE							
Age	39	44	19	23	33	27	31	69
Height	162	164	164	164	154	162	164	158
Weigth Loss	6.5	7.4	8.1	8.7	4.2	4.9	5.4	6

The residual sums of squares from these regressions are $s_1^2 = 36.308$ and $s_2^2 = 8.743$, respectively. since the two regressions have no common parameters,

$$s_{1,2}^2 = s_1^2 + s_2^2 = 36.308 + 8.743 = 45.051.$$

Under null hypothesis;

$$H_0 \colon \hat{\beta}_M = \hat{\beta}_F,$$

the residual sum of squares s_{12}^2 is to be computed by concatenated regressions of Y on X_{12}.

The estimated parameters of this regression are

$$\hat{\alpha}_M = -15.823,\ \hat{\alpha}_F = -15.163,\ \hat{\beta} = 0.0343,\ \hat{\gamma}_M = -0.0346,\ \hat{\gamma}_F = 0.139.$$

$$s_{12}^2 = 47.707$$

The observed value of the F-statistic appropriate for testing the null hypothesis can be computed as

$$F = \frac{2.656/1}{45.051/10}$$

Its p-value is $p = 1 - H_{1,10}(0.589) = 0.4604$ to be computed by R language program, and therefore the data in Table 4 do not provide sufficient evidence to suspect the validity of the null hypothesis. The difference in weight loss for male and female is not significantly different.

4. Conclusion

In this study three different cases; comparison of regression models with common parameters, comparison of regression models without common parameters and comparison of parameters in case of heteroscedasticity are examined. For all cases, generalized p value method has given good results even in case of heteroscedasticity. It can be concluded that generalized p value can be used in cases of assumption violations; for both nonnormality and heteroscedasticity in order to compare the regression models.

References

Sezer A, Ozkip E and Yazici B (2015). Comparison of confidence intervals for the behrens-fisher problem. Communications in Statistics-Simulation and Computation. DOI: 10.1080/03610918.2015.1082587.

Tsui KW and Weerahandi S (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. Journal of the American Statistical Association, 84(406): 602-607.

Weerahandi S (2013). Exact statistical methods for data analysis. Springer Science & Business Media, New York, USA.