Application of the permutation test for comparing regression models

Dominika Polko-Zając[1]

Abstract

Regression analysis is one of the most widely used tools for determining the relationship between variables. Such analysis takes place in a situation where the value taken by the dependent variable should be explained by the relationship with other independent variables. In practice there is often need to compare the regression coefficients in different populations. In order to identify differences between the regression models the use of permutation tests are proposed. The properties of the proposed test have been characterized using Monte Carlo study for various variants of the data.

Keywords: *regression analysis, permutation tests, Monte Carlo study*

JEL Classification: C12, C15

1. Introduction and basic notation

In the research may be concerned in comparing the regression function between the same variables in two or more considered groups or populations. In particular maybe interested us verification the hypothesis saying that regression lines have for both population studied the same regression coefficients. This is equivalent to testing whether the regression lines are parallel – have the same slope in the coordinate system. In the case of rejection the hypothesis that two lines are parallel we find that there is different pace of change in the average value of the variable *Y* at identical changes to the variable *X*. Test for examining significance of the existing differences between the compared linear regression functions takes into account the two populations, in which the two-dimensional distribution of observed characteristics *X* and *Y* is normal or close to normal. On the basis of randomly selected samples of the population of sizes respectively n_1 and n_2 hypothesis that both linear regression functions in these populations $y = \beta_{01} + \beta_{11}x$ and $y = \beta_{02} + \beta_{12}x$ have the same regression coefficients (slopes) should be verified, i.e. hypothesis $H_0 : \beta_{11} = \beta_{12}$ against the alternative hypothesis ($H_1 : \beta_{11} \neq \beta_{12}$). The value of the test statistics is calculated from the formula (Greń, 1974):

$$t = \frac{b_{11} - b_{12}}{s_{b_{11}-b_{12}}} \tag{1}$$

[1] Corresponding author: University of Economics in Katowice, Department of Statistics, Bogucicka 14, 40-226 Katowice, Poland, e-mail: dpolko@gmail.com.

where coefficients *b11* and *b12* are regression coefficients (slopes) estimated with method of least squares for each population studied and standard deviation *b11*-*b12* can be estimated from the formula:

$$s_{b_{11}-b_{12}} = \sqrt{\frac{\sum_{i=1}^{n_1}\left(y_{i1}-\hat{y}_{i1}\right)^2+\sum_{i=1}^{n_2}\left(y_{i2}-\hat{y}_{i2}\right)^2}{n_1+n_2-4}\left(\frac{1}{\sum_{i=1}^{n_1}\left(x_{i1}-\overline{x}_1\right)^2}+\frac{1}{\sum_{i=1}^{n_2}\left(x_{i2}-\overline{x}_2\right)^2}\right)}. \tag{2}$$

The critical value of Student's *t* for this test has n_1+n_2-4 degrees of freedom.

In the case of large samples (*n1*, *n2* > 30) formula of standard deviation of difference *b11*-*b12* is advised:

$$s_{b_{11}-b_{12}} = \sqrt{{s_{b_{11}}}^2+{s_{b_{12}}}^2} \tag{3}$$

where $s_{b_{11}}$ and $s_{b_{12}}$ are standard error of estimated parameters of slope of linear functions. Formula is proposed by Clogg et al. (1995) and cited by Paternoster et al. (1998).

In the case there is no basis to reject the null hypothesis that lines are parallel, regression coefficient *b* (slope) common to both regression equations can be calculated with the formula (Tadeusiewicz et al., 1993):

$$b=\frac{\sum_{i=1}^{n_1}\left(x_{i1}-\overline{x}_1\right)\left(y_{i1}-\overline{y}_1\right)+\sum_{i=1}^{n_2}\left(x_{i2}-\overline{x}_2\right)\left(y_{i2}-\overline{y}_2\right)}{\sum_{i=1}^{n_1}\left(x_{i1}-\overline{x}_1\right)^2+\sum_{i=1}^{n_2}\left(x_{i2}-\overline{x}_2\right)^2}. \tag{4}$$

Variance can be calculated with following formula:

$$\sigma^2(b)=\frac{s^2}{\sum_{i=1}^{n_1}\left(x_{i1}-\overline{x}_1\right)^2+\sum_{i=1}^{n_2}\left(x_{i2}-\overline{x}_2\right)^2}. \tag{5}$$

2. Idea of permutation tests

Parametric tests require the assumptions about the distribution of the characteristic in the population to be fulfilled. Non-parametric tests do not require this assumption, but the power of these tests is usually lower. Alternative approach is to use permutation tests in statistical research (Polko, 2014). These tests do not need fulfill the assumption about conformity with normal distribution however have powers similar to parametric tests. These tests were introduced by R.A. Fisher in 1930's (Welch, 1990). The essence of permutation tests is to determine test statistic and then to evaluate the distribution of this statistics for all permutation of variable. These tests can give results that are more accurate than those obtained with the use of traditional statistical methods (Kończak, 2012). When calculations affect large number of permutations, Monte Carlo study is applied. O'Gorman (2012) describes several different methods of permutations in linear models analysis. Methods that have been proposed are:

- permute the residuals from the reduced model and add these to the predicted values from the reduced model to form a new vector of predicted values;
- permute the independent variable for the coefficient that is to be tested;
- permute the dependent variable;
- permute the residuals from the complete model and add these to the predicted values from the complete model to form a new vector of predicted values.

Permutation tests are computer – intensive statistical methods, but the concept of these tests is simpler than of the tests based on normal distribution. Good (1994) indicates 5 following steps in permutation testing:

1. Identify the null hypothesis and the alternative hypotheses.
2. Choose a test statistic *T*.
3. Compute the test statistic T_0 for the sample data.
4. Determine the frequency distribution of the statistic under the null hypothesis. Perform the permutation of variables *N*-times and then calculate the statistics test value T_i.
5. Make a decision using this distribution as a guide.

The decision concerning a verified hypothesis is made on the basis of *ASL* (*achieving significance level*) value (Efron and Tibshirani, 1993):

$$ASL = P_{H_0}\{T \geq T_0\}\,, \tag{6}$$

for which estimation is obtained on the basis of:

$$A\hat{S}L = \frac{card\{i : T_i \geq T_0\}}{N}\,. \tag{7}$$

This notation applies where the H_0 rejection area is right–sided. In the case of left-sided rejection area in above notation inequality sign should be changed. When *ASL* is lower than the assumed level of significance α, then H_0 is rejected in favor of hypothesis H_1.

3. Comparison of regression models

In a situation where compared regression models have the following form:

$$Y_1 = \beta_{01} + \beta_{11}x_{11} + \varepsilon\,, \tag{8}$$

and

$$Y_2 = \beta_{02} + \beta_{12} x_{12} + \varepsilon \,, \tag{9}$$

to test hypothesis $H_0 : \beta_{11} = \beta_{12}$ against alternative hypothesis $H_1 : \beta_{11} \neq \beta_{12}$ permutation test can be used. It is important to take right decision in setting the test statistics. In the case where we consider hypothesis of no significant difference between slopes of regression lines following statistics are proposed:

$$T^{(1)} = \frac{|b_{11} - b_{12}|}{s_{b_{11}-b_{12}}}, \tag{10}$$

or

$$T^{(2)} = \frac{(b_{11} - b_{12})^2}{s^2_{b_{11}-b_{12}}} \tag{11}$$

where $s_{b_{11}-b_{12}}$ was calculated using formula (3).

It is also possible to identify the differences not only between slopes but also intercepts of models. In the case of testing hypothesis $H_0 : \beta_{01} = \beta_{02} \wedge \beta_{11} = \beta_{12}$ that there is no significant difference between studied regression function versus alternative hypothesis H_1 which is the negation of H_0 following test statistics were proposed:

$$T^{(3)} = \frac{|b_{01} - b_{02}|}{s_{b_{01}-b_{02}}} + \frac{|b_{11} - b_{12}|}{s_{b_{11}-b_{12}}} \tag{12}$$

or

$$T^{(4)} = \frac{(b_{01} - b_{02})^2}{s^2_{b_{01}-b_{02}}} + \frac{(b_{11} - b_{12})^2}{s^2_{b_{11}-b_{12}}} \tag{13}$$

where $s_{b_{01}-b_{02}}$ and $s_{b_{11}-b_{12}}$ were calculated using formula (3).

4. Monte Carlo study

Simulation study to confirm the effectiveness of the proposed procedure for comparing two regression models was executed considering the data generated according to the following method.

Data set of n_1=50 points was generated according to $Y = X + \varepsilon$, where $X \sim U(0;1)$ and $\varepsilon \sim N(0;0.1)$. Then next n_2=50 points were introduced according to the models:

a) $Y = (\delta + 1)X + \varepsilon$, where $\delta = 0, 0.05, 0.1, 0.15$ (Fig. 1);

b) $Y = -0.5\delta + (\delta + 1)X + \varepsilon$, where $\delta = 0.1, 0.2, 0.3$ (Fig. 2).

These simulated datasets are similar to Durio and Isaia (2010) proposal.

The steps in simulation test according to presented plan in order to compare regression function were as following:

1. Level of significance $\alpha = 0.05$ was assumed.
2. Sample S_1 (n_1 elements) with values compliant with first model was generated and sample S_2 (n_2 elements) with values compliant with second model.
3. Regression lines coefficients for each sample were determined.
4. Value of statistics was calculated according to formula (1).
5. Value of statistics T_0 defined with formula (10) or (11) was calculated at the same time.
6. Variables Y were permuted N times and each time value of statistics T_i was calculated.
7. Value of statistics T_0 defined with formula (12) or (13) was calculated too and in this case variables Y were permuted N times and each time value of statistics T_i was calculated.
8. Steps 2-7 were repeated 1000 times.

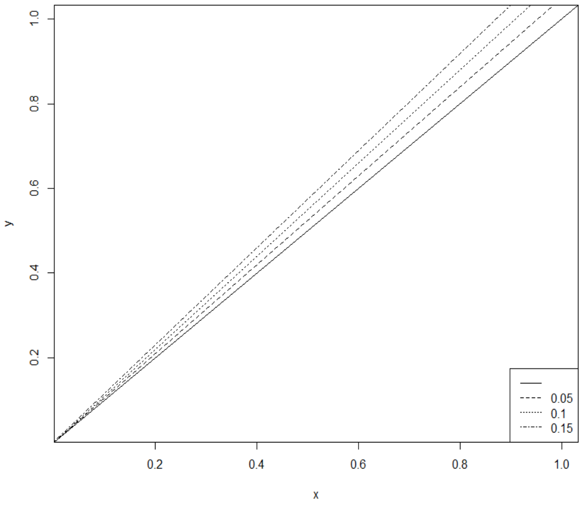

Fig. 1. Considered variants of compared linear regressions (models (a)).

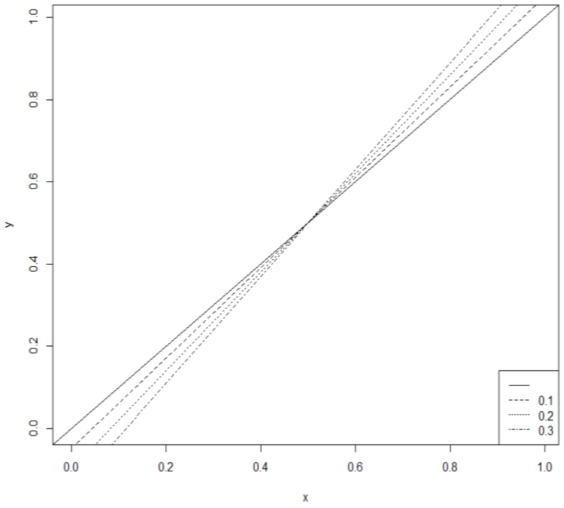

Fig. 2. Considered variants of compared linear regressions (models (b)).

The study was performed to compare the efficacy of the proposed method involving the use of the permutation test in comparison with the results obtained using the classical method of comparing slopes of linear regression function. The values of probability being results of permutation test for comparing two regression models (slopes and intercepts) were presented too. In simulations N = 1,000 permutations of variables were assumed. To evaluate *ASL* values 1,000 simulations have been executed. Estimated probabilities of rejecting the null hypothesis are presented in Table 1.

For economic data, the relationship between variables is rarely linear and different regression models should be considered (e.g. logarithmic or exponential). In Table 2 results of proposed permutation tests for different possible forms of regression functions were presented.

Generating the data (compared models)		**Classic test**	**Permutation test**			
		t	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$
a) $Y = X + \varepsilon$	a$_0$) $Y = X + \varepsilon$	0.044	0.048	0.064	0.049	0.044
	a$_1$) $Y = 1.05X + \varepsilon$	0.128	0.111	0.136	0.078	0.058
	a$_2$) $Y = 1.1X + \varepsilon$	0.325	0.295	0.299	0.120	0.143
	a$_3$) $Y = 1.15X + \varepsilon$	0.593	0.566	0.582	0.208	0.311
b) $Y = X + \varepsilon$	b$_1$) $Y = -0.05 + 1.1X + \varepsilon$	0.302	0.288	0.320	0.262	0.279
	b$_2$) $Y = -0.1 + 1.2X + \varepsilon$	0.795	0.795	0.830	0.767	0.789
	b$_3$) $Y = -0.15 + 1.3X + \varepsilon$	0.977	0.984	0.988	0.979	0.976

Table 1. The probability of rejecting the null hypothesis for linear models.

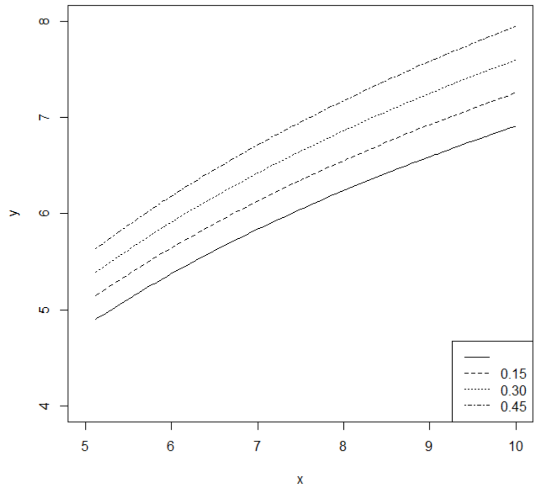

Fig. 3. Variants of compared logarythmic regression models.

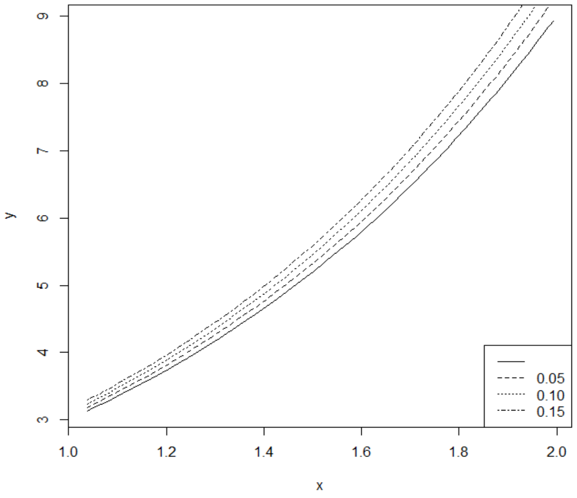

Fig. 4. Variants of compared exponential regression models.

Data generated according to two scenarios were considered in the following way:

1. Set of n_1 = 50 points according to $Y = 3\log X + \varepsilon$ where $X \sim U(5;10)$ and $\varepsilon \sim N(0;0.1)$. Then next n_2 = 50 points according to model $Y = (\delta + 3)\log X + \varepsilon$ where $\delta = 0{,}0.15{,}0.30{,}0.45$ (Fig. 3).
2. Set of n_1 = 50 points according to $Y = 3^X + \varepsilon$ where $X \sim U(1;2)$ and $\varepsilon \sim N(0;0.1)$ was generated. Then next n_2 = 50 points according to $Y = (\delta + 3)^X + \varepsilon$ where $\delta = 0{,}0.05{,}0.10{,}0.15$ (Fig. 4).

Model	**Method of generating data (compared models)**		**Permutation test**			
			$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(4)}$
logarythmic	b) $Y = 3\log X + \varepsilon$	b0) $Y = 3\log X + \varepsilon$	0.057	0.044	0.052	0.042
		b1) $Y = 3.15\log X + \varepsilon$	0.298	0.320	0.122	0.157
		b2) $Y = 3.30\log X + \varepsilon$	0.809	0.815	0.312	0.555
		b3) $Y = 3.45\log X + \varepsilon$	0.994	0.991	0.898	0.991
exponential	a) $Y = 3^X + \varepsilon$	a0) $Y = 3^X + \varepsilon$	0.074	0.062	0.082	0.078
		a1) $Y = 3.05^X + \varepsilon$	0.198	0.208	0.133	0.125
		a2) $Y = 3.10^X + \varepsilon$	0.629	0.584	0.350	0.230
		a3) $Y = 3.15^X + \varepsilon$	0.896	0.903	0.788	0.487

Table 2. The probability of rejecting the null hypothesis.

The results obtained in simulation confirm the efficacy of the proposed method of comparing regression models. Permutation tests allow comparing populations where assumptions for classic test using Student’s *t* statistics are not fulfilled. Tests also allow comparing complete models of regression functions (comparison includes also intercept of the model). In the case of linear and logarithmic regression models test sizes are near assumed significance level α. In the case of exponential regression model test sizes are slightly larger than assumed significance level α.

The advantage of using permutation test is the ability to use this method even if assumptions about the distribution of the characteristic in the population are not met. It is possible to use any test statistics without knowing of the distribution. Permutation tests in economic research lead to effective identification of differences between statistical populations also when these populations do not meet particular assumptions and there is no possibility of using parametric methods.

Conclusion

Regression analysis is a commonly used tool to determine the relationship between variables. In order to identify differences between the regression models permutation test were proposed. Permutation tests have similar power to parametric tests. The paper presents the results of computer simulation of classical test comparing the regression coefficients (slopes) and tests built on permutation tests. Permutation tests allow comparing populations where assumptions for classic test using Student's *t* statistics are not fulfilled. Tests also allow comparing complete models of regression functions (comparison includes also intercept of the model). The tests concerned various forms of the regression functions. Simulations showed that permutation tests can be used to study various relationships between variables – not only linear regression models.

In economic research these tests leads to effective identification of differences between statistical populations. The properties of the proposed tests have been characterized using a computer simulation in R program.

References

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology, 100* (5), 1261-1293.

Durio, A., & Isaia, E. D. (2010). Clusters Detection in Regression Problems: A Similarity Test Between Estimated Models. *Communications in Statistics – Theory and Methods, 39,* 508-516.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Good, P. (1994). *Permutation Tests: A practical guide for testing Hypotheses*. New York: Springer–Verlag.

Greń, J. (1974). *Statystyka matematyczna*. Warszawa: PWN.

Kończak, G. (2012). On testing multi-directional hypotheses in categorical data analysis. In *Proceedings of 20th International Conference on Computational Statistics*, 427-436.

O'Gorman, T. W. (2012). *Adaptive Tests of Signiffcance Using Permutations of Residuals with R and SAS*. New Jersey: John Wiley and Sons.

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for equality of regression coefficients. *Criminology, 36*(4), 859-866.

Polko, D. (2014). On testing the similarity of multivariate populations structures. In *Proceedings of 32nd International Conference Mathematical Methods in Economics*. Olomouc: Palacký University, 813-818.

Tadeusiewicz, R., Izworski, A., & Majewski, J. (1993). *Biometria*. Kraków: Wydawnictwo AGH.

Welch, W.J. (1990). Construction of Permutation Tests. *Journal of the American Statistical Association. Theory and Methods, 85*(411), 693-698.