TJMM
2 (2010), No. 2, 165-170

THE OPTIMIZATION OF LINEAR MULTIPLE REGRESSION MODELS THROUGH THE METHOD OF FORWARD SELECTION

FELICIA ZĂVOIANU AND CONSTANTIN ZĂVOIANU

Abstract. This article refers to the problem of optimizing linear multiple regression models through the method of forward selection. Therefore, the techniques for selecting the variables that can be inserted into the model are presented and the optimization algorithm is described. The algorithm is implemented on territorial statistical data that partially characterize the labour market.

1. Techniques For Selecting The Variables That Can Enter The Model

Let there:

$$Y_R = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_{i-1} X_{i-1} + \alpha_{i+1} X_{i+1} + \cdots + \alpha_p X_p + u$$

be a multiple regression model in which $(p-1)$ factorial variables have been inserted (the partial model) and

$$Y_E = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_{i-1} X_{i-1} + \alpha_i X_i + \alpha_{i+1} X_{i+1} + \cdots + \alpha_p X_p + u$$

the model obtained by inserting the factorial variable X_i (the extended model).

We mark up with S_R and S_E the sum of squares SP_{reg} that correspond to the two models and with S_Z the residua variance of the second model, viz.:

$S_R = SP_{reg}^{(R)} = \sum_{i=1}^{n} \left(\hat{y}_i^{(R)} - \bar{y}\right)^2$, a sum with $(p-1)$ degrees of freedom;

$S_E = SP_{reg}^{(E)} = \sum_{i=1}^{n} \left(\hat{y}_i^{(E)} - \bar{y}\right)^2$, a sum with p degrees of freedom;

$$S_Z = \frac{SP_{rez}^{(E)}}{n-p-1} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(E)}\right)}{n-p-1}.$$

The difference $S_E - S_R$ represents the contribution to $SP_{reg}^{(E)}$ of the a_i coefficient on the assumption that all the other terms were in the model and that the model was extended by adding the term $\alpha_i X_i$. With regard to the above statement and taking into account the fact that the sum $S_E - S_R$ has only one degree of freedom, for any i, the difference $S_E - S_R$ can be compared to $SP_{rez}^{(E)}$ using the F test. Such a test is named the *partial F-test* of α_i. Although the test is performed on the α_i coefficient, we shall say that the factorial variable X_i is being tested and therefore, the partial F-test can be used as a criterion for inserting a new factorial variable into a model.

The $S_E - S_R$ statistic has, as it has been previously stated, only one degree of freedom and the $SP_{rez}^{(E)}$ statistic has $(n-p-1)$ degrees of freedom. As such, the $F_{X_i}^* = \frac{S_E - S_R}{SP_{rez}^{(E)}} \cdot \frac{n-p-1}{1}$ statistic has an F distribution with 1 and $(n-p-1)$ degrees of freedom. If $F_{X_i}^* > F_{\alpha;1,n-p-1}$, the $H_0^{(i)} : \alpha_i = 0$ assumption will be invalidated and therefore, the

2010 *Mathematics Subject Classification.* 62j02, 62j05.
Key words and phrases. optimization, multiple regression, selection of factorial variables, algorithm.

factorial variable X_i can be inserted into the model, or, in other words, the extended model is to be preferred to the partial model.

Taking into account the previous notations, the statistic of the partial F-test can also be expressed: $F^*_{X_i} = \frac{SP^{(E)}_{reg} - SP^{(R)}_{reg}}{SP^{(E)}_{rez}} \cdot \frac{n-p-1}{1}$.

One must notice the fact that the statistic of the partial F-test used for checking the assumption $H^{(i)}_0 : \alpha_i = 0$ as to the alternative $H^{(i)}_1 : \alpha_i \neq 0$ can also be determined in another way, which is that it equals the square of the $t^*_i = \frac{a_i}{s(a_i)}$ statistic, i.e. $F^*_{X_i} = (t^*_i)^2 = \frac{a^2_i}{s^2(a_i)}$, $i = \overline{1,p}$.

2. The Optimization Algorithm

The algorithm for optimizing linear multiple regression models through the method of forward selection starts with no initial factorial variables included in the model and has the following steps:

(1) ***Selecting the first variable that will be inserted into the model***. The correlation coefficients $r_1, r_2, \ldots, r_p$ between the resulting variable Y and the factorial variables $X_1, X_2, \ldots, X_p$ are calculated, i.e. $r_j = \rho(Y, X_j) = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{n \cdot s_{X_j} \cdot s_Y}$, $\forall j = \overline{1,p}$ and then k is determined so that $|r_k| = \max\limits_{i=1,p} \{|r_i|\}$; in this way, the factorial variable with the strongest correlation to the resulting variable is determined. Obviously, this factorial variable is X_k.

(2) ***Adjusting the model corresponding to the selected variable.*** The model $Y = \alpha_0 + \alpha_k X_k + u$ is adjusted.

(3) ***Verifying the insertion criterion.*** With the help of the $F^*_{X_k} = \frac{a^2_k}{s^2(a_k)}$ statistic the $H^{(k)}_0 : \alpha_k = 0$ assumption is checked as to the alternative $H^{(k)}_1 : \alpha_k \neq 0$. If $F^*_{X_k} < F_{\alpha;1,n-p-1}$ then the assumption $H^{(k)}_0$ will be accepted, a situation in which, the estimate of the resulting variable is in fact its selection mean, viz. $\bar{y}$. If $F^*_{X_k} > F_{\alpha;1,n-p-1}$, the $H^{(k)}_0$ assumption is false and as a result the factorial variable X_k will be inserted into the model and the $Y = a_0 + a_k X_k$ model will be an *optimum partial model.*

(4) ***Selecting a new variable that can be inserted into the model.*** If an optimum partial model exists and there are factorial variables that haven't been included in the model, the partial model is extended by inserting, one at a time, each variable that hasn't been included in this model. If the partial model contains q factorial variables (at the beginning $q = 1$), $(p - q)$ extended models will have to be built. Because the partial model has q degrees of freedom, the extended models will each have $(q+1)$ degrees of freedom. Let there I_q be the set of indices of the factorial variables in the partial model. For each extended model, the $F^*_{X_i}$ statistics are calculated, for those X_i that are not in the partial model, i.e. $F^*_{X_i} = \frac{SP^{(E)}_{reg} - SP^{(R)}_{reg}}{SP^{(E)}_{rez}} \cdot \frac{n-q-1}{1}$, $\forall i \notin I_q$ and the value of k is determined so that $F^*_{X_k} = \max\limits_{i \notin I_q} \left\{|F^*_{X_i}|\right\}$. The X_k variable can be inserted into the model if the insertion criterion is satisfied.

(5) ***Verifying the insertion criterion***. The $F^*_{X_k}$ statistic has an F distribution with 1 and $(n - q - 1)$ degrees of freedom. If $F^*_{X_k} > F_{\alpha;1,n-q-1}$ the $H^{(k)}_0 : \alpha_k = 0$ will be invalidated and as a result the factorial variable X_k can be inserted into the model, or, in other words, the extended model is to be preferred to the partial model. Therefore, if the $H^{(k)}_0 : \alpha_k = 0$ assumption is false, the extended model,

in which the factorial variable X_k has been inserted, will be maintained as it is considered to be an optimum partial model and the procedure described above will be resumed. By contrast, if the $H_0^{(k)} : \alpha_k = 0$ assumption is true, the partial model available at this stage will be considered optimum.

(6) ***Extending the model by inserting the selected variable.*** If the $H_0^{(k)} : \alpha_k = 0$ assumption is false, i.e. $F^*_{X_k} > F_{\alpha;1,n-q-1}$, the optimum partial model from the previous step will be extended by inserting the X_k variable and thus, a new optimum partial model will be obtained.

(7) ***Steps 4, 5 and 6 will be repeated for the optimum partial model at hand*** until no new factorial variable can be inserted into the model. The last optimum partial model is in fact the optimum model we have been searching for.

Checking the validness of the optimum model. The $F^* = \frac{SP_{reg}}{SP_{rez}} \cdot \frac{n-q-1}{q}$ statistic, which has an F distribution with q and $(n-q-1)$ degrees of freedom, is computed. With the help of this statistic the $H_0 : \alpha_1 = \alpha_2 = \cdots \alpha_q = 0$ assumption can be verified as to the alternative H_1 assumption: an i exists so that $\alpha_i \neq 0$; testing is not extended for the free term α_o. If $F^* > F_{\alpha;\,q\,,(n-q-1)}$ the H_0 assumption will be invalidated, therefore a significant statistical regression has been obtained meaning that the model is valid and the additional elements of the regression can be determined. In scientific literature, it is recommended to use a regression model as a forecast tool if the F^* statistic is four times larger than the tabled value.

3. The Optimization Of The Multiple Regression Model Regarding The Economic Dependency Rate

After conducting a study regarding the Romanian labour market during the transition period, a study based on representative indicators, an accelerated growth of the values of the economic dependency rate at a district level has been recorded. This growth has obvious negative economic and social consequences. The study has shown the fact that statistical connections, made evident by the values of the correlation coefficients, exist between the *economic dependency rate* indicator and the following indicators: *the percentage of work resources in the total population*, *the labour force employment rate*, *the unemployment rate*, *the percentage of population working in the primary sector*, *the percentage of population working in the secondary sector*. Thus, for the territorial statistical data from 2002, presented in table 2, the correlation coefficients between Y and X_i, $i = \overline{1,5}$, marked up with $r_i = \rho(Y, X_i)$ are shown in table 1.

Table 1.

X_i	X_1	X_2	X_3	X_4	X_5
$r_i = \rho(Y, X_i)$	-0.4423	-0.9292	0.5103	0.5099	-0.4904

In this context, the study regarding the behavior of the economic dependency rate is suited to the employment of multiple regression models. The completely adjusted model, obtained through the least squares method, which has been validated from a statistical point of view is:

$$Y = 739.0814 - 4.5025 * X_1 - 4.7399 * X_2 + 0.0482 * X_3 - 0.0640 * X_4 - 0.0026 * X_5$$

[1] X_1 - the percentage of work resources in the total population (%); X_2 - the labour force employment rate (%); X_3 - the unemployment rate (%); X_4 - the percentage of population working in the primary sector (%); X_5 - the percentage of population working in the secondary sector (%).

[2] Y - the economic dependency rate (number of unemployed persons per 100 employed persons).

Table 2

DISTRICT	Factorial variables[1]					Dependent variable[2]
	X_1	X_2	X_3	X_4	X_5	Y
BC	69.4399	47.9541	9.4024	34.9341	34.6791	200.3073
BT	65.4500	54.2547	11.0231	56.8408	19.2164	181.6132
IS	66.5811	55.6712	9.6944	38.1770	27.9062	169.7853
NT	70.0171	53.1350	10.6590	49.5395	24.8182	168.7911
SV	67.0140	55.2078	10.3021	52.0220	20.7303	170.2925
VS	66.3902	49.3201	15.8682	56.5772	22.0134	205.4020
BR	69.0932	49.5272	9.9607	38.2146	33.5944	192.2271
BZ	66.7633	54.5143	9.3437	49.3355	23.0343	174.7586
CT	71.9322	53.3023	8.6895	29.0664	26.0029	160.8136
GL	70.7192	47.1533	14.7788	36.6409	28.9932	199.8819
TL	69.8556	48.1655	9.6411	44.4959	26.3036	197.2097
VR	67.7449	55.8642	5.9338	52.6244	22.6312	164.2345
AG	69.8072	57.7946	6.6070	32.7383	38.8910	147.8637
CL	64.6445	48.1303	10.6392	58.7129	18.4158	221.4030
DB	67.3099	55.7780	8.7745	42.8220	28.6627	166.3535
GR	62.4604	49.0207	7.2813	62.6096	12.8289	226.5998
IL	66.0298	50.8104	12.0434	55.7789	18.3920	198.0623
PH	69.6668	50.0176	10.2162	27.4896	39.3499	186.9796
TR	64.4626	61.9056	10.1555	61.5517	17.2414	150.5891
DJ	66.9291	56.1645	7.0501	48.4783	21.9203	166.0257
GJ	67.9834	55.1832	10.7675	33.1727	37.7839	166.5575
MH	67.5476	55.7942	8.7796	52.5087	22.9239	165.3391
OT	67.8240	52.9765	9.9044	55.1763	21.1604	178.3129
VL	67.9826	59.3374	11.6676	42.4115	25.8548	147.8986
AR	68.1029	61.4640	5.0463	29.2809	32.8505	138.8986
CS	69.7448	53.5277	9.7635	41.1576	27.5723	167.8609
HD	71.4594	56.4124	9.7966	25.9959	39.3769	148.0654
TM	68.7057	64.7954	3.9187	29.4235	32.9026	124.6276
BH	68.6918	66.5017	3.2214	39.0591	30.4158	118.9081
BN	68.9960	53.2016	10.0360	47.6399	22.2028	172.4274
CJ	69.2392	61.3874	9.9914	30.2310	31.8045	135.2712
MM	69.2884	56.3592	6.5350	45.5321	26.6566	156.0793
SM	70.9027	57.4090	3.9956	45.5518	29.3645	145.6729
SJ	67.2354	58.4094	7.3422	43.9425	26.5914	154.6355
AB	68.9876	66.8439	10.8496	33.8810	34.2210	116.8538
BV	72.5434	56.0495	11.9184	16.7850	43.9666	145.9407
CV	70.1302	56.0242	9.2090	33.5240	35.9268	154.5183
HG	69.7430	57.5781	7.7101	39.8473	31.3740	149.0244
MS	68.8891	60.1785	6.4324	36.2126	33.5133	141.2172
SB	70.5134	56.0577	7.2508	21.2358	41.6917	152.9838
B-IF	72.1201	56.4481	3.2723	5.5494	35.1500	145.6374

Source: Computations based on the 2003 Statistical yearbook, National Institute of Statistics.

The optimization algorithm

I. The correlation coefficients between Y and X_i, $i = \overline{1,5}$, marked up with $\rho(Y, X_i)$ are:

X_i	X_1	X_2	X_3	X_4	X_5
$r_i = \rho(Y, X_i)$	-0.4423	-0.9292	0.5103	0.5099	-0.4904

Because the maximum absolute value is $r_{\max} = |r_2| = 0.9292$, the variable that can be inserted into the models is X_2, and the initial selected model is $Y = 437.4658 - 4.9042 * X_2$. This models is an optimum partial model because $F^*_{X_2} = 246.4831 > 4.0913 = F_{0.05\,;\,1,39}$ (the insertion criterion is satisfied).

II. The extended models obtained by inserting the otherX_i, $i \neq 2$ variables, as well as their associated statistics $F^*_{X_i}$, are:

X_i	The extended model obtained by inserting the $X_i$variable	$F^*_{X_i}$
X_1	$Y = 717.0428 - 4.7222 * X_2 - 4.2282 * X_1$	405.8348
X_3	$Y = 425.1689 - 4.7618 * X_2 + 0.4925 * X_3$	0.5806
X_4	$Y = 392.0058 - 4.5066 * X_2 + 0.5719 * X_4$	42.6947
X_5	$Y = 443.7159 - 4.5382 * X_2 - 0.9318 * X_5$	35.4563

$F_{\max} = F^*_{X_1} = 405.8348$; so the X_1 variable can be inserted into the model. Because $F_{0.05:1,38} = 4.0982$, we observe that $F_{\max} > F_{0.05:1,38}$ and as a result the insertion criterion is satisfied. Consequently, the X_1 variable is inserted into the model and $Y = 717.0428 - 4.7222 * X_2 - 4.2282 * X_1$ becomes the new optimum partial model.

III. The extended models obtained by inserting the other X_i, $i \notin \{2, 1\}$ variables, as well as their associated statistics $F^*_{X_i}$, are:

X_i	The extended model obtained by inserting the $X_i$variable	$F^*_{X_i}$
X_3	$Y = 716.2092 - 4.7153 * X_2 - 4.2248 * X_1 + 0.0245 * X_3$	0.0162
X_4	$Y = 740.0086 - 4.7528 * X_2 - 4.5022 * X_1 - 0.0610 * X_4$	0.9946
X_5	$Y = 728.1021 - 4.7421 * X_2 - 4.4025 * X_1 + 0.0697 * X_5$	0.5158

$F_{\max} = F^*_{X_4} = 0.9946$; so the X_4 variable can be inserted into the model. Because $F_{0.05:1,37} = 4.1055$, we observe that $F_{\max} < F_{0.05:1,37}$ and as a result the insertion criterion is not satisfied. Consequently, the X_4 variable is not inserted into the model and the **optimum model** is model obtained in the previous step, viz.:

$$Y = 717.0428 - 4.7222 * X_2 - 4.2282 * X_1$$

The additional elements of the regression are:
The standard error values for the coefficients:
Es(a0)= 14.6116
Es(a1)= 0.2071
Es(a2)= 0.0919
The coefficient of determination: R^2= 0.9886
Residual quadratic mean deviation: Su= 2.8062
The F* statistic: F*=1645.5981
The number of degrees of freedom for SPrez : ng=38
The regression sum of squares : SPreg= 25916.8947
The residual sum of squares: SPrez= 299.2353
The optimum model is valid from a statistical point of view because
F* > 4*Finv(0.05; 2,38) i.e. 1645.5981 > 12.9792.

4. Conclusions

(1) The conclusion I have reached after conducting this study was that the optimization through the method of forward selection offers the advantage of successively inserting into the model only those variables that have statistically important coefficients of regression, thus progressively building an optimum model from a previous optimum partial model.

(2) This method of optimization is an "excellent remedy" for reducing the multicollinearity phenomenon that can in fact be regarded as being "omnipresent" because of the multiple interdependencies that exist in the economy.

References

[1] Cojocaru N., Clocotici V., Dobra D., *Metode statistice aplicate în industria textilă*, Editura Didactică şi Pedagogică, Bucureşti, 1986.
[2] Craiu V, *Verificarea ipotezelor statistice*, Bucureşti, 1972.
[3] Isaic-Maniu Al., Mitru C., Voineagu V, *Statistică*, Editura Universitară, Bucureşti, 2003.
[4] Rencher A.C., *Linear Models in Statistics*, Wiley-Interscience, 1999.
[5] Hans von Storch, Zwiers F.W., *Statistical Analysis in Climate Research*, Cambridge University Press, Cambridge, UK, 2002.
[6] ***, Internet: http://psych.ucs.edu/faculty-zurbrigg-psy214a.
[7] ***, *The 2003 Romanian Statistical Yearbook*; National Institute of Statistics, 2003.
[8] Zavoianu Felicia, *Models for evaluating the impact of economic restructuring at a territorial level*, Doctoral Thesis, A.S.E. Bucureşti, 2007.

University of Petroşani
Department of Mathematics and Computer Science
Universităţii 20, 332006 Petroşani, România
E-mail address: fzavoianu@yahoo.com

University of Petroşani
Department of Mathematics and Computer Science
Universităţii 20, 332006 Petroşani, România
E-mail address: constantin.zavoianu@yahoo.com