BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LXI No. 1/2009	65 - 70	Seria Matematică - Informatică - Fizică

A Ridge Regression Model of the Cracking Process

Cristian Marinoiu

Universitatea Petrol-Gaze din Ploieşti, Bd. Bucureşti 39, Ploieşti, Catedra de Informatică
e-mail: cmarinoiu@upg-ploiesti.ro

Abstract

Although recent theoretical and practical developments have considerably widened the range of modelling instruments, linear regression models still claim a central place in statistical modelling. This fact is largely due to the remarkable characteristics of the least squares approach. However, when the matrix of regressing variables is ill-conditioned, the stability of regression coefficients is in turn affected, and the model thus configured is implicitly unrealistic. Under such circumstances, the ridge regression estimator may prove to be a viable alternative. The present paper deals with the setting up of a ridge regression model for the catalytic cracking of a chemical reactor.

Key words: ridge regression, variance inflation factor, cracking process

Introduction

Catalytic cracking represents mainly the production process of gasolines and, secondarily, ofelines through complex chemical reactions. The whole process can be characterized by the following variables [1]:

- the disturbances of the process highlighted at the level of the raw material by density, medium volumetric temperature and sulphur content;
- the cracking process commands, identified by feedstock flow, output heater feedstock temperature, catalyst temperature in regenerator system and catalyst /feedstock ratio;
- the output of the process: gas productivity and octane number.

In order to model the process several interesting models have been suggested [1]. However, the extreme complexity of these models makes them difficult to use in the control of the catalytic cracking process. An alternative to these models has been elaborated in [2], by using the regression model, and their efficiency in the optimal management of the catalytic cracking process has been emphasized in [3]. The main goal of the present paper is to construct a linear model with regression coefficients stable from a numerical point of view.

Theoretical Aspects of the Model Construction

To construct the model experimental data drawn from [1] have been used. Table 1, taken over from this paper, contains a selection of volume 17, extracted from the observations recorded in a catalytic cracker during a 90 days' span of functioning. The notations used are as follows: octane number (Y_1), gas productivity (Y_2), density (X_1), volumetric temperature (X_2), sulphur content (X_3), feedstock flow (X_4), output heater feedstock temperature (X_5), catalyst temperature in regenerator system (X_6), catalyst/feedstock ratio (X_7).

Table 1. Experimental data for the catalytic cracking process

No. obs.	Y_1	Y_2	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	91.2	52.3	0.9007	442.0	0.38	183.4	316	732.0	4.6
2	90.8	52.8	0.9029	441.5	0.25	183.1	311	730.0	4.5
3	90.6	52.8	0.9028	434.4	0.25	184.3	310	732.0	4.7
4	90.4	51.4	0.9043	448.6	0.29	189.7	310	725.0	4.6
5	90.6	52.4	0.9009	442.5	0.38	183.8	320	731.0	4.7
6	90.6	52.1	0.9039	440.0	0.25	182.4	310	734.0	4.5
7	91.0	52.8	0.9042	445.8	0.38	182.6	312	728.5	4.6
8	90.7	52.2	0.9050	445.0	0.32	183.7	319	733.0	4.5
9	90.5	52.8	0.9007	436.8	0.39	182.8	315	732.0	4.6
10	91.0	51.8	0.9014	440.2	0.28	182.7	316	733.0	4.5
11	91.0	52.3	0.9004	443.2	0.49	187.9	316	726.0	4.5
12	91.0	52.0	0.9020	436.0	0.23	191.1	324	734.0	4.4
13	90.5	53.0	0.9030	441.5	0.25	184.6	311	733.0	4.9
14	91.0	51.3	0.9068	449.6	0.43	182.2	314	727.0	4.6
15	92.0	52.7	0.9033	442.4	0.36	182.7	312	732.0	4.5
16	91.9	43.7	0.9217	438.2	2.14	173.6	314	727.5	4.8
17	92.5	45.4	0.9247	438.4	2,19	188.7	319	727.0	5.0

Let us consider the dependence between the dependent variable y and the independent (regressors) variables $X_1, X_2, \ldots, X_7$ to be of the form:

$$y = \beta_o + \beta_1 X_1 + \ldots\beta_7 X_7 + \varepsilon , \qquad (1)$$

where ε is the additive error.

The corresponding linear regression model may be written in a matrix form as follows:

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

where:

- $\boldsymbol{y}$ (17 x 1) is the vector of y_i observations for the dependent variable Y_1 or Y_2;
- $\boldsymbol{X}$(17 x 8) is the matrix of $x_{1i}, x_{2i}, \ldots, x_{7i}$ observations, respectively for the regressors $X_1, X_2, \ldots, X_7$, the elements in the first column of the matrix being all equal to 1;
- $\boldsymbol{\beta}$ (8 x 1) is the vector of unknown parameters $\beta_o, \beta_1, \ldots, \beta_7$;
- $\boldsymbol{\varepsilon}$ (17 x 1) is the vector of errors, with the mean $E(\boldsymbol{\varepsilon}) = 0$ and a variance-covariance matrix $Cov(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{I}_{17}$, σ^2 being the unknown variance of errors, and $\boldsymbol{I}_{17}$ the 17×17 unit matrix.

When the matrix $\boldsymbol{X}$ has the columns linearly independent, the Ordinary Least Squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ for parameter $\boldsymbol{\beta} = (\beta_o, \beta_1, ..., \beta_7)$ is as shown beneath [4]:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} . \quad (3)$$

The OLS estimator has remarkable properties: it is the best linear unbiased estimator ($E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$) in the class of the linear estimators in the observations of the dependent variable $\boldsymbol{y}$. On the other hand, the numerical stability of the ORE can be affected under certain circumstances. Thus, if the columns of the matrix $\boldsymbol{X}$ are linearly dependent or almost linearly dependent, the matrix $\boldsymbol{X}$ is rank deficient; this is termed multicolliniarity or near multicolliniarity, respectively, and the matrix $\boldsymbol{X}$ is ill-conditioned [4]. The degree of conditioning of the matrix $\boldsymbol{X}$ is given by the so-called condition number, which registers values higher or equal to 1. The higher the values of this number, the worse-conditioned the matrix will be. Statistically speaking, this situation occurs when the regressors are strongly correlated. The unpleasant consequence is that the matrix determinant $\boldsymbol{X}'\boldsymbol{X}$ is equal to 0 or is almost 0, which can affect the accuracy of the values of the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and implicitly of the estimated regression coefficients $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_7$. An indicator of the presence of colliniarity is the VIF Variance Inflation Factor(VIF). It is recognized that a VIF value much higher than 1 clearly indicates instability issues of the corresponding coefficients [5].

Formally, the Ordinary Ridge Estimator (ORE) differs from the Ordinary Least Squares (OLS) estimator by an arbitrary constant k ($0 \le k \le \infty$) added to the diagonal of the correlation matrix of the regressors X_1, X_2, …, X_7. In other words, if we define $\boldsymbol{Z}$ to be the matrix of 17×7 order obtained from $\boldsymbol{X}$ by canceling the first column and standardizing the other columns, the ORE estimator for our model is defined as follows [4]:

$$\hat{\boldsymbol{\beta}}_k = (\boldsymbol{Z}'\boldsymbol{Z} + k\boldsymbol{I}_{17})^{-1}\boldsymbol{Z}'\boldsymbol{Y} , \quad (4)$$

where $0 \le k < \infty$.

For $k = 0$ the OLS estimator can be obtained, provided that we consider that the data of the matrix $\boldsymbol{X}$ were previously standardized. The resulting model is still linear, but the ORE, unlike the LS estimator is biased, and the extent of the bias depends on the vector of unknown parameters $\boldsymbol{\beta}$. Also, when $k \to \infty, \hat{\boldsymbol{\beta}}_k \to \boldsymbol{0}$, namely, the ORE shrinks the estimates towards zero. From a practical point of view, if the matrix $\boldsymbol{Z}$ is ill-conditioned, for the values of the constant k strictly higher than 0, the determinant of the matrix that is reversed $\boldsymbol{Z}'\boldsymbol{Z} + k\boldsymbol{I}_{17}$ will be non zero. The direct consequence is obtaining regression coefficients stable from a numerical point of view.

Practical Aspects of the Model Construction

To solve the model (2), SAS software has been used [6]. Solving the model means above all estimating the regression coefficients $\beta_o, \beta_1, ..., \beta_p$. To start with, it was sought to obtain the OLS estimator according to the formula (3) with the help of the REG procedure of SAS. We considered the case when y stands for Y_1 (octane number).

Unfortunately, there are very tight correlations among the variables of the system: for example, $corr(X_1, X_2) = 0.96$ and $corr(X_1, X_7) = 0.66$. The consequence of these tight correlations is multicolliniarity or near multicolliniarity, which is indicated by the exaggerated size of the VIF value for the estimators $\hat{\beta}_1$ and $\hat{\beta}_3$ (Figure 1). Moreover, the presence of multicolliniarity is demonstrated by the fact that no regression coefficient is significant – see column ($Pr > |t|$).

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	34.71798	64.29647	0.54	0.6023	0
x1	x1	1	-53.60850	69.69988	-0.77	0.4615	21.78116
x2	x2	1	0.05890	0.04788	1.23	0.2499	3.65251
x3	x3	1	2.09482	1.08354	1.93	0.0852	40.50799
x4	x4	1	0.05413	0.04341	1.25	0.2439	2.61192
x5	x5	1	-0.04139	0.04331	-0.96	0.3642	2.84952
x6	x6	1	0.11993	0.08302	1.44	0.1825	5.53674
x7	x7	1	-1.48531	0.98520	-1.51	0.1659	2.24411

Fig. 1. OLS coefficients affected by multicolliniarity

The clear conclusion is that there are serious reasons for doubt concerning the correctness of the obtained estimations (see Figure 1) and that the ridge regression must be used as an alternative. The ORE has been obtained according to the formula (4) by means of the same REG procedure of SAS. The graphic representation of the VIF values of the regression coefficients is given in Figure 2 for the range of values of k between 0 and 0.2 with a step of 0.02.

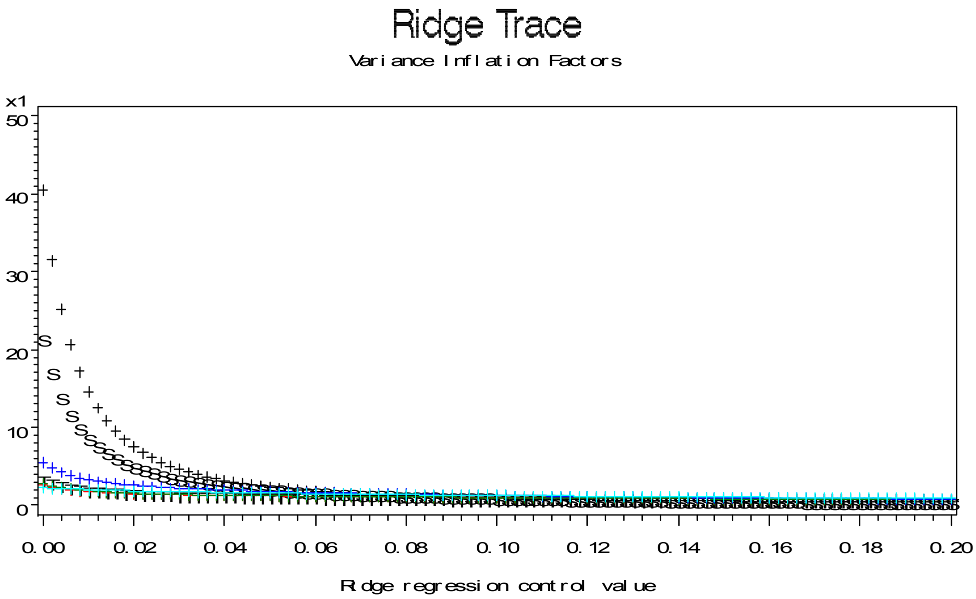

Fig. 2. The VIF values ploted against k

Figure 3 represents the ridge curves that offer an enlightening view over the stability of the regression estimators depending on parameter k which varies between 0 and 0.20 with step 0.02. It can be seen that while for the variables X_2, X_3, …, X_7 the values of the regression coefficients estimators become stable for small values of k, the value of the regression coefficient estimator of the variable X_1 becomes stable for much higher values.

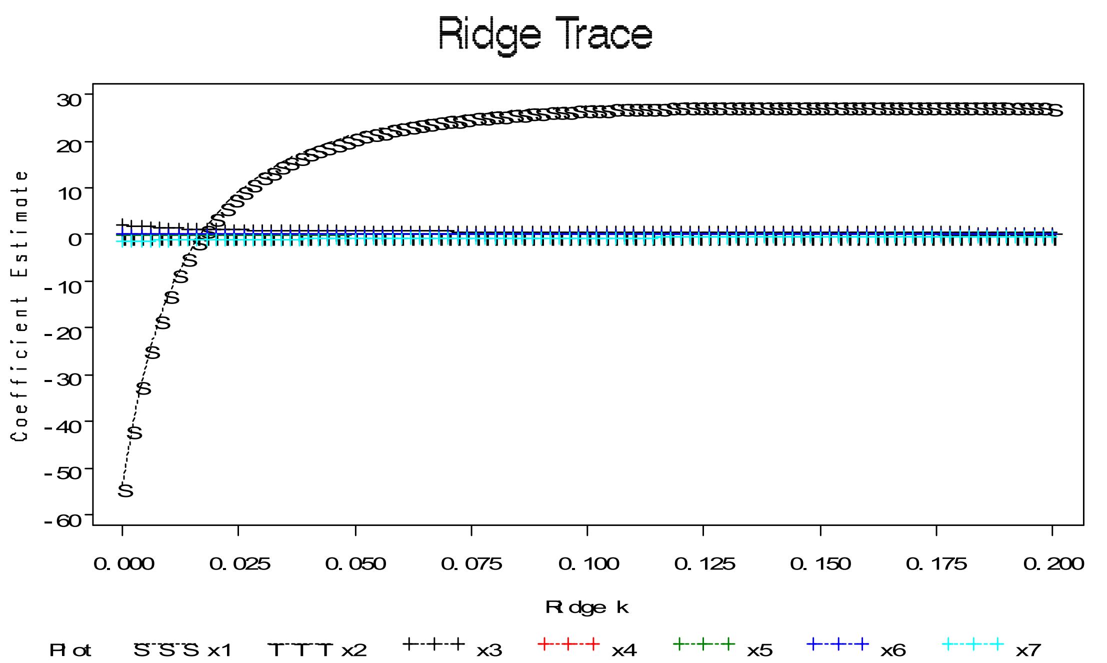

Fig. 3. The values of the estimated regression coefficients ploted against k

For theoretical reasons [5], we must choose the value of parameter k as the value that produces a VIF value that is approximately equal to 1 for all the estimated regression coefficients.

Obs	k	VIF1	VIF2	VIF3	VIF4	VIF5	VIF6	VIF7
104	0.102	1.27644	1.07078	0.85263	0.96912	0.97014	1.24090	1.2456
106	0.104	1.25415	1.06200	0.83028	0.96257	0.96349	1.22729	1.2357
108	0.106	1.23261	1.05337	0.80900	0.95614	0.95695	1.21398	1.2259
110	0.108	1.21179	1.04491	0.78871	0.94981	0.95052	1.20096	1.2163
112	0.110	1.19165	1.03659	0.76936	0.94358	0.94420	1.18822	1.2068
114	0.112	1.17215	1.02842	0.75088	0.93746	0.93799	1.17575	1.1974
116	0.114	1.15326	1.02040	0.73322	0.93143	0.93187	1.16354	1.1881
118	0.116	1.13495	1.01251	0.71634	0.92550	0.92585	1.15159	1.1790
120	0.118	1.11720	1.00476	0.70019	0.91965	0.91992	1.13987	1.1700
122	0.120	1.09998	0.99713	0.68472	0.91390	0.91408	1.12839	1.1612
124	0.122	1.08326	0.98963	0.66990	0.90823	0.90833	1.11713	1.1524
126	0.124	1.06702	0.98225	0.65568	0.90264	0.90267	1.10610	1.1437

Fig. 4 . The tabulated values of the VIF. Note that for $k = 0.12, VIFi \approx 1$.

Further, around this value both the RMSE (Root Mean Square Error) for each coefficient and the very values of the coefficients have to undergo insignificant changes. Looking to the Figure 2 to Figure 5 it can be noted that a convenient value is 0.12.

Obs	_RIDGE_	_RMSE_	Intercept	x1	x2	x3	x4	x5	x6	x7
105	0.102	0.46779	51.9867	27.4953	0.004342465	0.54797	.006037023	0.007428	0.015885	-0.67677
107	0.104	0.46816	52.2392	27.5714	0.004164223	0.54371	.005853939	0.007594	0.015491	-0.67053
109	0.106	0.46852	52.4882	27.6418	0.003991017	0.53958	.005675573	0.007755	0.015108	-0.66439
111	0.108	0.46888	52.7338	27.7069	0.003822629	0.53557	.005501732	0.007911	0.014734	-0.65833
113	0.110	0.46923	52.9760	27.7669	0.003658853	0.53168	.005332235	0.008063	0.014369	-0.65237
115	0.112	0.46958	53.2149	27.8223	0.003499495	0.52789	.005166914	0.008211	0.014013	-0.64648
117	0.114	0.46992	53.4506	27.8731	0.003344373	0.52421	.005005606	0.008354	0.013666	-0.64068
119	0.116	0.47026	53.6831	27.9198	0.003193317	0.52063	.004848160	0.008494	0.013327	-0.63495
121	0.118	0.47059	53.9124	27.9625	0.003046162	0.51714	.004694432	0.008630	0.012996	-0.62931
123	0.120	0.47092	54.1387	28.0014	0.002902757	0.51374	.004544287	0.008763	0.012673	-0.62374
125	0.122	0.47124	54.3619	28.0368	0.002762957	0.51043	.004397595	0.008891	0.012357	-0.61824
127	0.124	0.47156	54.5822	28.0688	0.002626624	0.50721	.004254234	0.009017	0.012049	-0.61281

Fig. 5. The estimated regression coefficients for some values of k (_RIDGE_)

Results and Conclusions

We can obtain the estimated regression coefficients for the chosen value *k=0.12* (Figure 5). This leads us to the following regression model:

$$Y_1 = 54.14 + 28 \times X_1 + 0.003 \times X_2 + 0.51 \times X_3 + 0.004 \times X_4 + 0.08 \times X_5 + 0.012 \times X_6 - 0.62. \times X_7 \,.$$

The model underlines the fact that the octane number (Y_1) significantly depends on density (X_1) and to a much lesser extent on sulphur content (X_3), catalyst /feedstock ratio (X_7) and the other variables of the system. Similarly, we obtain the regression model for the second case (Y_2 represents gas productivity):

$$Y_2 = 95.57 - 160.76 \times X_1 + 0.05 \times X_2 - 1.73 \times X_3 + 0.08 \times X_4 + 0.06 \times X_5 + 0.11 \times X_6 + 0.44. \times X_7 \,.$$

The ORE has been adopted as an alternative to the OLS, with a view to obtaining regression coefficients stable from a numerical point of view.

References

1. Belitzianis, M. - *Conducerea evoluată a sistemului reactor-regenerator din instalația de cracare catalitică*, Teză de doctorat, Universitatea Petrol-Gaze din Ploieşti, 1992
2. Pătrăşcioiu, C., Marinoiu, C. - Soluții numerice pentru modelarea statistică a procesului de cracare catalitică, *Revista de Informatică Economică*, vol. 3, nr. 10, Bucureşti, 1999
3. Pătrăşcioiu, C., Marinoiu, C. - Optimal Control System of FCC Plant, *Buletinul Universității Petrol-Gaze din Ploieşti*, vol. LII, nr. 1, 2000
4. Vinod, H., Ullah, A. - *Recent advances in regression methods*, Marcel Dekker, New York, 1981
5. Marques de Sa, J. - *Applied Statistics using SPSS, STATISTICA, MATLAB,* Springer-Verlag Berlin Heidelberg, 2007
6. *** - *SAS 9.1.3*, SAS Institute Inc., USA, http://www.sas.com/, accessed on 10 June 2009

Un model de regresie ridge al procesului de cracare catalitică

Rezumat

Deşi dezvoltările teoretice şi practice ale ultimilor ani au lărgit considerabil paleta instrumentelor de modelare, totuşi, modelele de regresie liniară continuă să ocupe un loc central în modelarea statistică. Acest lucru se datorează în mare măsură proprietăților remarcabile ale estimatorului prin cele mai mici pătrate. Totuşi, atunci când matricea variabilelor regresoare este rău condiționată, stabilitatea coeficienților de regresie este afectată, şi implicit modelul obținut poate fi nerealist. În această situație, estimatorul ridge de regresie poate fi o alternativă bună. Lucrarea de față se ocupă de construcția unui model de regresie ridge pentru procesul de cracare catalitică a unui reactor chimic.