Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(3):71-77

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

The linear model analysis of chemical element impaction to hot rolled ribbed steel bar

Chang Jincai*[1,2,3], Chen Liyu[1] and Zhang Yuzhu[1,2,3]

[1]College of Science, Hebei United University, Hebei Tangshan, China
[2]Tangshan Iron and Steel Group Co., LTD, Hebei Tangshan, China
[3]Northeastern University School of Material and Metallurgy, Liaoning Shenyang, China

ABSTRACT

In this paper, we use the least squares estimation and principal component analysis to analyze the performance impact of manganese and chromium to hot-rolled ribbed steel bars, eliminate collinearity and establish the linear regression equation of manganese and chromium with steel yield strength. In the fitting result, the proportion of carbon and manganese is consistent with practical experience, and the correlation of sulfur and the yield strength is still a problem. The fitting result can provide a reference for analyzing the performance of hot-rolled ribbed steel bars. However the nonlinear relationship of manganese and chromium with hot-rolled ribbed steel bars has not been resolved, we need to further analyze.

Keywords: Least Squares, principal component analysis, yield strength

INTRODUCTION

HRB400 hot-rolled ribbed steel bars are widely used in civil construction, bridges, special equipment, chemicals and other areas. Rebar mechanical properties are the basic parameters in much structural safety evaluation. As we all know, there is a objectively certain relationship in the mechanical property indexes of steel, process factors of chemical composition, test conditions and test results[1,2]. For example, the higher of carbon and manganese content in chemical composition, the higher the tensile strength of the steel and the lower the elongation. The relationship between these variables can not be expressed by a determined function, that is to say the one to one relationship does not exist between them, and however, there is a certain correlation. Under some conditions of confidence, the relationship between the variables can also be expressed as a function[3]. Regression analysis is a statistical analysis method to deal with the correlation among variables. As a relatively old method, in the eighteenth century, least squares estimation was first founded by Gauss and successfully applied in astronomical observations and geodetic work. Least squares regression analysis is the more commonly used calculation Method. It is used to solve the problem that how to find a reliable value from a set of measured values. The basic principle is[4,5,6] to find a best fit curve from the measured data which can make sure that the quadratic sum of difference between the measured value and the fitted value at each point on the curve is minimum. The OLSE has been very widely used in many fields such as parameter estimation, system identification and prediction, forecasting, etc.[7,8]

Principal components analysis is an analyzing and simplifying data collection technique. In 1901, it was invented by Karl Pearson for analyzing the data and building mathematical models. The dimensionality reduction method provides a strong theoretical and technical support in the comprehensive evaluation. PCA can transform the problem of high-dimensional space into low-dimensional space. And with the principal component analysis process, it will automatically calculate the weights of each main component which can largely resisted the interference of human factors in the evaluation process [9], therefore, comprehensive evaluation theory based on principal component can be batter to ensure the objectivity of the evaluation results and objectively reflect the real problems

This structure of paper: The first part introduces the principle of least squares and fitting test methods. The second part introduces multicollinearity and principal component analysis. The third part introduces how to estimate the linear regression equation among manganese, chromium and yield strength of hot-rolled ribbed steel in using OLSE

2 The model of Least squares estimation

2.1 theory of Least squares estimation

There totally has p elements $x_1, x_2, \cdots, x_p$, supposing that they have the following linear relationship with y that, $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$, y is a observable random variable, $\beta_1, \beta_2, \cdots, \beta_p$ are unknown parameters, ε is an unobservable random error which meets $E\varepsilon = 0$, $Var(\varepsilon) = \sigma^2$, σ^2 is unknown. In general, we call it multivariate linear regression model, denoted $\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \\ E\varepsilon = 0 , \quad Var(\varepsilon) = \sigma^2 \end{cases}$. The unknown parameters $\beta_1, \beta_2, \cdots, \beta_p$ are the regression coefficients. The explanatory variables or regression variables x_i $(i = 1,2,\cdots,n)$ are often referred as the regression factor or predictor, called factor. In a sense, β_i $(i = 1,2,\cdots,n)$ reflects the contribution regression factor x_i $(i = 1,2,\cdots,n)$ make to observations y .So it is usually referred β_i as the effect of factors x_i

Now we use matrix to discuss the solution of issues. This method has many advantages, the more important point is the solution can be used for any regression problems once the problem is given in matrix form and obtain a matrix form solution. No matter how many terms in the regression equation, all the related formulas are in the same forms. Set up there are n independent observations $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$, $(i = 1,2,\cdots,n)$ and

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (i = 1,2,\cdots,n) \\ \qquad E\varepsilon_i = 0 , Var(\varepsilon_i) = \sigma^2, \end{cases} ,$$

ε_1, ε_2, $\cdots$, ε_n are independent. And set

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} , X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} , \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

We can get $\begin{cases} Y = X\beta + \varepsilon \\ E\varepsilon = 0 , \quad Var(\varepsilon) = \sigma^2 I_n \end{cases}$. And the matrix form of multiple linear regression, this condition can be expressed as $\varepsilon \sim N(0, \sigma^2 I_n)$. With the above assumptions and nature of the multivariate normal distribution, random vector y can be consider as an expectation vector which obey the 5-dimensional normal regression model. As we know $E(Y) = X\beta$, $Var(y) = \sigma^2 I_n$,so we can get $y \sim N(X\beta, \sigma^2 I_n)$. We use the least squares estimator to find the estimated value $\hat{\beta}$ of β .And $\hat{\beta}$ meets the equation $Q(\hat{\beta}) = \min_{\beta} Q(\beta)$ .Then calculate the partial derivative and make it zero, we can get the equation

$\sum_{i=1}^{n} (y_i - \sum_{i=1}^{p} x_{ti} \beta_i) x_{ti} = 0 \quad (k = 0,1,\cdots,p)$, $\sum_{i=1}^{p} (\sum_{i=1}^{n} x_{ti} x_{tk}) \beta_i = \sum_{i=1}^{n} y_t x_{tk} \quad (k = 0,1,\cdots,p)$ .Write in matrix form, replace β with $\hat{\beta}$,then we can get the normal equation $X'X\beta = X'Y$. Solve the normal equations and get the solution of the equation $\hat{\beta} = (X'X)^{-1} X'Y$,the one $\hat{\beta}$ is the least squares estimation of β .For other non-linear distribution function such as exponential functions, logarithmic functions can be transformed into a linear problem and solved.

2.2 Regression testing
Then we introduce three different test methods

1) F-Test (the regression equation test)
Set the overall regression coefficient is $\beta_j (j = 1,2,\cdots, p)$, then null hypothesis H_0 and alternative hypothesis H_1 of F-test can be written as

H_0: $\beta = 0$, scilicet $\beta_1 = \beta_2 = \cdots = \beta_p = 0$;

$H_1 : \beta \neq 0$, scilicet there is at least one component of β is not zero $\beta_j \neq 0 (j = 1,2,\cdots, p)$

Make the two sides of the equation $y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$ squared, then we can get $\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$, it can be abbreviated denoted as $SST = SSE + SSR$. $SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$, $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, $SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$. SSR is called the regression sum of squares, SSE is called the residual sum of squares and SST is called the total sum of squares. Construct F –statistic $F = \frac{SSR}{SSE/(n-p-1)} \sim F(p, n-p-1)$,for α test level，check the F distribution table, get the critical value $F_\alpha (p, n-p-1)$ of the rejection region. If $F \leq F_\alpha (p, n-p-1)$, then we accept the null hypothesis H_0, otherwise, reject the null hypothesis H_0.

2) T-Test(the coefficient test regression)
From the t distribution, we can get the equation

$$\frac{\sqrt{\sum_{i=1}^{p} (x_i - \overline{x})^2} (\hat{\beta}_1 - \beta_1)}{\sqrt{\frac{SSE}{n-2}}} = \frac{\sqrt{\sum_{i=1}^{p} (x_i - \overline{x})^2} (\hat{\beta}_1 - \beta_1)}{\hat{\sigma}_e} \sim t(n-p-1)$$

So when we get a given level of significance α , we can use the t-test statistics to test H_0 .the rule is if $|T| > t_{\frac{\alpha}{2}} (n-2)$, then we reject the null hypothesis H_0, otherwise, accept the null hypothesis H_0.

3) R-Test
If SSR / SST is close to 1, then the effect of the regression equation is remarkable. Again look at the ratio $\frac{SSR}{SST} = \frac{\hat{\beta}_1^2 L_{xx}}{L_{yy}} = (\frac{L_{xy}}{L_{xx}})^2 \frac{L_{xx}}{L_{yy}} = \left[\frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} \right]^2 = \left[\frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \right]^2$,we can get the sample correlation coefficient r , $r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$ which is referred as the correlation coefficient.

This suggests that, when the SST fixed, the more $|r|$ is close to 1, the less the SSE. Specially, when $|r|$ is equal to 1, then SSE is equal to 0and SSR is equal to SST. That is to say the change of y is caused by the linear relationship of y and x . So statistics r can be used to weigh the close degree of the linear correlation of y and x .This test method is called the r-test.

Generally, the hypothesis test of regression equation includes two aspects: one is the regression model test which can test if there is a linear model can show the relationship of the independent variable and dependent variable. This is done by f-test; another test is about the regression parameters test the t-test. When the model is through the inspection, we need the specific inspection to test the influence of each independent variable on the dependent variable.

3 Principal component analysis

3.1 The multicollinearity

The original meaning of multicollinearity is some of the independent variables is linear related in linear regression model. Now the multicollinearity contains two cases the completely linear and approximate linear. That is to say some independent variables in linear regression model have a completely or approximate linear relationship. Generally multicollinearity elimination methods are principal component analysis, partial least squares, ridge regression, etc.

1)The dangers of multicollinearity

(1)In the condition of completely multicollinearity parameter estimators may not exist. (2)In the condition of approximate multicollinearity the OLS estimators are not effective and the parameter estimation is unstable. (3)It can't correctly reflect the influence of each explanatory variable to the explained variable which may cause the actual meaning of the parameter estimator is not reasonable.

2) The causes of multicollinearity

(1) The limitation of sample data and the number of sample data is not enough. (2) In the model, the setting of explain variables is error and there has internal relationship between the variables. (3) Variables have common trend

3) Methods for identifying multicollinearity

Multicollinearity shows there have the relevant relationship between variables, so the main test method for identifying multicollinearity are statistical methods.

(1) Variance inflation factor method (VIF)

define $VIF_j = \left(1 - R_j^2\right)^{-1}$, R_j^2 is the multiple determination coefficient of X_j .Generally speaking, if the biggest VIF_j is larger than 10,then there may exist multicollinearity. The fact show that if $VIF_j = \left(1 - R_j^2\right)^{-1} > 10$, then $1 - R_j^2 < 0.1, R_j^2 > 0.9$.

(2) Characteristic root decision method

According to the nature of the matrix determinant, the value of matrix determinant is equal to it's the product of characteristic root. When $\left|X^{'}X\right| \approx 0$, at least there is one characteristic root is zero, there must exist multicollinearity in column vectors. According to the condition number $K_i = \sqrt{\frac{\lambda_m}{\lambda_i}}$, λ_m is the biggest characteristic root, λ_i is the others. It is commonly believed if $0 < k < 10$, there is no multicollinearity, or if $k > 10$ there has multicollinearity.

3.2 Principal component analysis

Principal component analysis is a method of dimension reduction. It has the characteristics which can keep the largest contribution to the variance. Through linear transformation, it can combine the original multiple indicators into a few indicators which can fully reflect the overall information. On the premise of not losing important information, it can avoid problem of multicollinearity and do further analysis.

Specific steps of principal component regression:

(1) Do standardized processing to the original sample data and get the correlation coefficient matrix R of

explanatory variables.
(2)Calculate the eigenvalues of the R , $\lambda_1 > \lambda_2 > \cdots > \lambda_k$ and the standardization feature vector $u_1, u_2, \cdots, u_k$.
(3) Use eigenvalue to test multicollinearity. At least there is one characteristic root is zero, there has multicollinearity. Then set $\lambda_{m+1}, \lambda_{m+2}, \cdots, \lambda_k$ approximate to zero, this shows that there have $k-m$ linear correlation among the variables.
(4) Setting the multiple linear model is $Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_K X_K + \mu$ and we can get K principal components of the explanatory standardized variables $\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_k$

$$\begin{cases} z_1 = u_{11}\mathrm{X}_1 + u_{12}\mathrm{X}_2 + \cdots + u_{1k}\mathrm{X}_k \\ z_2 = u_{21}\mathrm{X}_1 + u_{22}\mathrm{X}_2 + \cdots + u_{2k}\mathrm{X}_k \\ \vdots \\ z_k = u_{k1}\mathrm{X}_1 + u_{k2}\mathrm{X}_2 + \cdots + u_{kk}\mathrm{X}_k \end{cases}$$

z_i are unrelated and $z_{m+1}, z_{m+2}, \cdots + z_k$ approximate to zero. Get the regression of the standardized variables Y and the principal component $z_1, z_2, \cdots, z_m$

$$\mathrm{Y} = \hat{\mathrm{a}}_1 \mathrm{z}_1 + \hat{\mathrm{a}}_2 \mathrm{z}_2 + \cdots + \hat{\mathrm{a}}_\mathrm{m} \mathrm{z}_\mathrm{m}$$

4 The example analysis

In the construction industry, the HRB400 hot rolled ribbed steel bars instead of HRB335 hot rolled ribbed steel bar is widely used. It not only can improve the seismic resistance of buildings but also reduce the usage of steels. As we all know, there is a certain relationship between the steel composition and performance, such as the higher content of carbon and manganese is, the higher tensile strength and the lower elongation of steels. Reflected in the math, this relationship is the interrelation among the variables. But now the effect of chromium and manganese element to the performance of hot rolled ribbed steel bar is still not clear. It needs to do further research such as statistical analysis, function fitting and ingredient optimization research. The following is data fitting with SPSS software.

Table 1 Anova [b]

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	75004.483	11	6818.589	46.514	.000a
Residual	462209.230	3153	146.593		
Total	537213.712	3164			

a. Predictors: (Constant), ALT, MN, MO, P, S, C, V, CU, Si, Cr, Ni
b. Dependent Variable: Yield strength

Table 2 Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	370.351	8.197		45.184	.000		
C	91.206	20.178	.078	4.520	.000	.917	1.091
MN	15.789	6.233	.049	2.533	.011	.715	1.398
S	260.819	58.696	.076	4.444	.000	.938	1.066
P	-255.107	49.464	-.094	-5.157	.000	.814	1.228
Si	-9.286	9.005	-.019	-1.031	.303	.764	1.308
Cr	146.245	32.450	.135	4.507	.000	.306	3.267
V	1079.564	74.698	.263	14.452	.000	.827	1.210
Ni	-105.392	64.801	-.051	-1.626	.104	.277	3.612
CU	476.754	86.254	.106	5.527	.000	.743	1.346
MO	-95.940	180.933	-.009	-.530	.596	.862	1.160
AL	5.290	21.281	.004	.249	.804	.994	1.006

a. Dependent Variable： Yield strength

Table 2 has given the regression coefficient of the linear regression model and some corresponding statistics. From this table ,we can get constant and coefficient of various elements about the linear regression model. In addition, we can get t value of the constant and coefficient of the linear regression model from Table 2. We can get the fitting result

$$\mathrm{Y} = 370.351 + 91.206\mathrm{C} + 15.789\mathrm{Mn} + 260.819\mathrm{S} - 255.107\mathrm{P} - 9.286\mathrm{Si} +$$
$$1079.564\mathrm{V} + 146.245\mathrm{Cr} - 105.392\mathrm{Ni} + 476.754\mathrm{Cu} - 95.940\mathrm{Mo} + 5029\mathrm{ALT}$$

Table 3 Collinearity Diagnostics

Dimension	1	2	3	4	5	6
Eigenvalue	10.974	0.548	0.233	0.141	0.035	0.023
Condition Index	1	4.473	6.867	8.815	17.829	21.724
Dimension	7	8	9	10	11	12
Eigenvalue	0.019	0.015	0.007	0.003	0.001	0
Condition Index	23.911	26.653	40.737	65.829	95.258	170.847

Table 1 is the variance analysis result of the regression equation, by the F test $F = 46.514$, Significance level $\alpha = 0.001$, $F_\alpha(p, n-p-1) = F_{0.001}(11, 3174-11-1) \approx 3$,we can get $F = 46.514 > F_\alpha(p, n-p-1) \approx 3$, so the overall regression model is through f test. By the t test, $t_{\alpha/2}(n-p-1) = t_{0.0005}(3162) = 3.291$, most of the elements can be through the regression test. From colinearity test results of Table 1 and Table 2, we can see in the variance expansion factors of the corresponding parameters estimator are small, but by the eigenvalues and condition index, some characteristics roots tend to zero and there is significant collinearity. And by the condition number in the table, there has significant collinearity. We nee to further study and to eliminate collinearity.

Table 4 Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.547	23.152	23.152	2.547	23.152	23.152
2	1.656	15.051	38.203	1.656	15.051	38.203
3	1.179	10.718	48.922	1.179	10.718	48.922
4	1.021	9.281	58.203	1.021	9.281	58.203
5	.959	8.719	66.922	.959	8.719	66.922
6	.849	7.718	74.640	.849	7.718	74.640
7	.807	7.340	81.980	.807	7.340	81.980
8	.746	6.784	88.764	.746	6.784	88.764
9	.605	5.504	94.268			
10	.472	4.288	98.556			
11	.159	1.444	100.000			

Extraction Method: Principal Component Analysis.

By the principle of the cumulative contribution rate can not be less than 85% , In this paper, we get eight principal component. First of all we can get linear regression of the principal components $F_1, F_2, \cdots F_8$ and the yield strength.

$$Y = 453.85 + 3.547F_1 + 1.58F_2 + 0.533F_3 + 0.343F_4 - 0.195F_5 - 1.366F_6 + 1.132F_7$$
$$-1.977F_8$$

Then do linear fitting of each principal components $F_1, F_2, \cdots F_8$ and chemical elements. Then we can get the fitting results

$$Y = 372.15 + 98.68C + 14.09Mn + 298.72S - 210.56P - 1.32Si + 969.98V + 55.55Cr$$
$$+183.56Ni + 256.15Cu - 363.73Mo + 76877Al$$

Because we have selected the large number of principal components, the results after eliminating collinearity and initial results have no obvious difference.

CONCLUSION

In the regression analysis, the least square method is a widely used regression methods. After fitting regression equation, the collinearity inspection is necessary. If there exist the multicollinearity, we can use principal component analysis, partial least squares, ridge regression and other methods to eliminate collinearity. In this paper, we use the least square method and principal component analysis to build the linear regression equation of chromium manganese element and rebar yield strength. The results are in conformity with practical experience.

Acknowledgements

This work was supported by National scientific and technological support projects (No.2012BAE09B00), the National Natural Science Foundation of China (No.51274270) and the Hebei iron and steel co., LTD., Tangshan Branch Project(No. JS201212)

REFERENCES

[1] HeXiaoqun. Application of regression analysis [M]. Renmin University of China Press **2007**.
[2] GaoHuixuan. Multivariate statistical analysis [M]. Peking University Press. **2004**.
[3] WuXiangbo, YeAzhong. *Statistics and Decision*,**2007**,(08)
[4] FangKaitai. Practical regression analysis [M]. Science Press, **1988**
[5] Jan Kmant .Elements of Econometrics. Second Edition [M].Macmi l -l ,New York,**1986**.
[6] Nickey J.Messick, John H.Kalivas, Patrick M.Lang. *Microchemical Journal*, 55(**1997**),200-207.
[7] V.Choulakian *Statisties&Probability Letter*[J]. **2001**(37):135-150
[8] LIJinghua, GuoYaohuang. *Journal of Industrial Engineering/Engineering Management*,**2002**(1):39-43
[9] WangLu. *Statistics & Information Forum*,**2003**(3):55-57