A Study on Prediction of Output in Oilfield Using Multiple Linear Regression

Izni binti Mustafar
Department of Electrical and Electronic Engineering
Universiti Teknologi PETRONAS
Bandar Seri Iskandar, 31750 Tronoh, Perak Darul Ridzuan
Tel: +6013-3938787, Email: izni.mustafar@gmail.com

Dr. Radzuan Razali
Department of Fundamental and Applied Sciences
Universiti Teknologi PETRONAS
Bandar Seri Iskandar, 31750 Tronoh, Perak Darul Ridzuan
Tel: +(605) 368 7679, Email: radzuan_razali@petronas.com.my

ABSTRACT

An oilfield is an area with reserves of recoverable petroleum, especially one with several oil-producing wells [1]. The challenge in this project is to find the variable for the output of oilfield because here are numerous factors affecting output in an oilfield. The relationship between field output and one of the affecting factors is unscientific and are not precise, which it is needed to come out with a simple and more accurate. 8 parameters have been identified to predict the oilfield of output. After several screening test using Multiple Linear Regression, 4 parameters have been identified as a most significant parameter. The ordinary least squares method also used to minimize the sum of variables by eliminating the least important variables. And to validate the data, new set of different data used with only the most significant parameters. It validates the method as all the parameters obey the rules of P-value.

Keywords – Oilfield output prediction ; Multiple Linear Regression

INTRODUCTION

The production of oil is very significance as a world energy source. Every year, the increasing of oil production has been by far as the major contribution to the growth in energy production. The oil production is generally from an oilfield. An oilfield is an area of sedimentary rocks under the ground or called as crude oil. Oil is created in a source rock along with water and gas. The oilfields typically extend over a large area, possibly several hundred kilometers across. Therefore, full exploitation entails multiple wells scattered across the area. In addition, there may be exploratory wells probing the edges, pipelines to transport the oil elsewhere and support facilities. The term oilfield is also used as shorthand to refer to the entire petroleum industry. However, it is more accurate to divide the oil industry into three sectors which are upstream, midstream and downstream. Upstream is a crude production from wells and separation of water from oil meanwhile midstream is a pipeline and tanker transport of crude and downstream is a refining and marketing of refined products[2]. For a major reason, it is crucial to predict the oilfield output for oil production. Thus, studies have been making to predict the output using multiple linear regression method.

Model Used To Predict the Output of Oilfield

The oilfield development of predicting the output of an oilfield is the basis of the optimal decision making of oilfield manager [4]. By far, there are many methods to predict the output of oilfield such as Multiple Linear Regression, Artificial Neural Network, Grey Prediction method, and Logistic Curve Method which have different applicable environments and limits [5]. At present time, there are several major models are being used to predict the oilfield output such as logistic model [6], production decline model [7] and logistic model [8]. But the problem is, there are several input variables in the above models and significant factors influencing dynamic system is not considered. Thus, the prediction result was affected and is not accurate. Meanwhile, the Multiple Linear Regression model is more simple and accurate. In the process of predicting the oilfield output using Multiple Linear Regression model, several model factors related to oilfield output are often identified as the model variables. By using this model, the Multiple Linear Regression equation is constructed. Therefore, the most significant factors that influence the oilfield output are determined by using the Multiple Linear Regression model. The model is applied to the actual production and the satisfying predictions are obtained.

Multiple Linear Regressions

Multiple linear regressions are one of the most widely used of all statistical methods. Multiple regression analysis is also highly useful in experimental situation where the experimenter can control the predictor variables. A single predictor variable in the model would have provided an inadequate description since a number of key variables affect the response variable in important and distinctive ways [9]. It attempts to model the relationship between two or more variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for p explanatory variables $x_1, x_2, \ldots, x_p$ is defined to be :

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad (1)$$

This line describes how the mean response μ_y changes with the explanatory variables. The observed values for y vary about their means μ_y and are assumed to have the same standard deviation σ. The fitted values $b_0, \ldots, b_p$ estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ of the population regression line [9].

β_0 is the mean of y when all x's are 0. Meanwhile, β_j is the change in the mean of Y associated with a unit increase in x_j, holding the values of all the other x's fixed. Coefficient estimated via least squares.

Meanwhile for the confidence and prediction Intervals the below calculation is use :

Variance of mean response at x_0 :

$$Var(\hat{y}_0) = Var\, x'_0 \hat{\beta} = \sigma^2 x'_0 (X'X)^{-1} x_0 = \sigma^2 v_0$$ [10]

(2) Variance of new observation at x_0, $y_0 = \hat{y}_0 + \varepsilon_0$ [10];

$$Var(\hat{y}_0 + \varepsilon_0) = Var(\hat{y}_0) + Var(\varepsilon_0) = \sigma^2 x'_0 (X'X)^{-1} x_0 + \sigma^2 = \sigma^2 (x'_0 (X'X)^{-1} x_0 + 1) = \sigma^2 (v_0 + 1) \qquad (3)$$

An estimate of σ^2 is s^2 = MSE = $\frac{y'(1-H)y}{(n-k-1)}$ (4)

The $(1-\alpha)$ Confidence Interval on Mean Response at x_0 is defined as below [10]:

$$\hat{y}_0 \pm cd \qquad (5)$$

Where;

$c = t_{n-(k+1),\alpha/2}$ and $d = \sqrt[s]{v_0}$ (6)

Meanwhile, the $(1-\alpha)$ Confidence Interval on New Observation at x_0 is defined as below [10] :

$$\hat{y}_0 \pm cd \qquad (7)$$

Where;

$c = t_{n-(k+1),\alpha/2}$ and $d = \sqrt[s]{v_0 + 1}$ (8)

Last but not least, the sum of squares was used. Sum of squares is a concept that permeates much of inferential statistics and descriptive statistics. More properly, it is the sum of squared deviations. Mathematically it is an unscaled , or unadjusted measure of dispersion. When scaled for number of degrees of freedom, it estimates the variance, or spread of the observations about their mean value [10].

Based on sample $i = 1, 2, \ldots, n$ containing n observations;

Sum of Squares Total (SST) :
$\sum_{i=1}^{n} (y_i - \bar{y})^2$ (9)

Sum of Squares for Error (SSE) :
$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$ (10)

Sum of Squares for Regression (SSR) :
$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ (11)

SSR=SST–SSE (12)

To see if there is any linear relationship we test[10] : $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (13)

$H_1: \beta_j \neq 0$ for some j (14)

Compute the equation as below :

SSE=$\Sigma(y_i - \hat{y}_i)^2$ (15)
SST=$\Sigma(y_i - \hat{y}_i)^2$ (16)
SSR=SST–SSE (17)

The F statistic is :

$$\frac{\frac{SSR}{k}}{\frac{SSE}{(n-k-1)}} = \frac{MSR}{MSE} \quad (18)$$

With F based on k and $(n - k - 1)$ degrees of freedom.
Reject H_0 when F exceeds $F_{k,n-k-1(\alpha)}$.

METHODOLOGY

Research Methodology

In order to achieve the aim of the project, some research has been done on several resources from books, technical papers and internet. For the first step, the gathering information needs to be done on the Oilfield, Well Production, Reservoir Behaviour and Multiple Linear Regression method. After all the studies have been done and the parameters have been identified, and the process of constructing a Multiple Linear Regression calculation using Microsoft Excel and obtain the oilfield output begin. The next stage is the simulation stage whereby the calculation will be simulated in order to make it easier to achieve the oilfield output. During this stage, knowledge of MATLAB software is a requirement. Apart from that, the most significant parameters were identified after several screening and validate this model using 2nd set of data with the most significant parameters only.

Determining Factors Affecting Oilfield Production

In oil production, there are two major factors affecting oilfield production which are geological factors and human factors. Therefore, these two factors are being considered to predict the output of oilfield. Considering the geological factors, the oil wells are the utmost important element in predicting oilfield's output directly determines the yield of oilfield [3].

Next, the water content of oil also is considered as major factor that affect the oilfield production. These due to some of oil well in our country are non self spraying. Thus, the respective oil wells need steam or injecting water to drive oil. It also can be used to increase pressure and thereby it will stimulate production of oilfield. The available oil reserve is also a factor because an underground reserve of oil is basically unchanged [3].

The basic method is to establish the linear relationship between oil output and the influencing factors such as moisture content. Then the linear system is established according to the experience. To predict future output of an oilfield, the influencing factors combined with actual production are selected and analysed deeply. Eight factors are selected as follows [3]:

1. The total numbers of wells
2. The startup number of wells
3. The number of new adding wells
4. The injected water volume last year
5. The oil moisture content of previous year
6. The oil production rate of previous year
7. The recovery percent of previous year
8. The oil output of previous year

RESULTS

A list of data parameters from China's Oilfield were obtains. Please refer to ***Table 1***. Based from data in ***Table 1***, the calculation was constructed using Microsoft Excel and Matlab based on Multiple Linear Regression (MLR) model where:

$x1$ = The total numbers of wells
$x2$ = The start up number of wells
$x3$ = The number of new adding wells

$x4$ = The injected water volume last year
$x5$ = The oil moisture content of previous year
$x6$ = The oil production rate of previous year
$x7$ = The recovery percent of previous year
$x8$ = The oil output of previous year
y = The oil output

From basic MLR equation $y = x\beta$, the basic form MLR can be expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$
(19)

Table 1 - Parameters Data From China Oilfield

year	y	x1	x2	x3	x4	x5	x6	x7	x8
1983	1442800	689	612	311	2375900	41.80%	1.45%	9.07%	1421900
1984	1417200	855	720	351	2305000	42.33%	1.53%	9.54%	1442800
1985	1466100	1028	874	426	2765900	42.93%	1.60%	9.49%	1417200
1986	1454500	1268	1087	472	3306400	46.21%	1.55%	10.25%	1466100
1987	1489400	1446	1197	652	3981400	45.80%	1.49%	9.35%	1454500
1988	1559200	1705	1417	486	4551000	47.80%	1.43%	9.08%	1489400
1989	1652300	1892	1524	458	5269100	49.30%	1.31%	9.31%	1559200
1990	2024600	2113	1761	473	6020400	52.15%	1.37%	10.13%	1652300
1991	2175900	2372	1903	506	7406200	55.46%	1.26%	10.88%	2024600
1992	2606400	2640	2123	705	8676500	59.83%	1.18%	11.54%	2175900
1993	3025300	3090	2574	689	9879800	60.87%	1.11%	12.07%	2606400
1994	3493100	3603	2826	964	11108700	63.39%	1.11%	12.96%	3025300
1995	3725800	3987	2878	1073	11832700	63.12%	1.20%	13.57%	3493100
1996	4037600	4530	3002	1003	13091800	64.79%	1.20%	14.76%	3725800
1997	4200500	4872	3172	1044	14063100	67.45%	1.07%	14.59%	4037600
1998	4398200	5110	3260	854	15760600	68.89%	1.01%	14.88%	4200500
1999	4649700	5400	3375	686	16760300	70.12%	0.95%	15.40%	4398200
2000	4712500	5524	3497	758	16519000	71.88%	0.88%	15.82%	4649700
2001	5205000	5653	3704	891	18083400	71.88%	0.91%	16.46%	4712500
2002	6115500	6958	5523	1043	19267300	72.95%	0.83%	17.22%	5205000
2003	7158700	8680	7805	1181	19580500	72.83%	0.83%	17.74%	6115500
2004	8109500	9864	8263	1319	25365000	72.28%	0.89%	17.71%	7158700
2005	9051000	11805	9522	1946	30032000	72.01%	0.84%	16.98%	8109500
2006	9623000	12314	11092	2347	32987000	72.31%	0.85%	17.20%	9051000

Linear regression method was used to calculate the regression coefficients with 8 independent variables. The regression coefficients from β_0 to β_8 are respectively given as follows:

β_0 = 2019687.48
β_1 = 177.71
β_2 = 218.255
β_3 = 193.70
β_4 = 0.077
β_5 = -5450242.21
β_6 = -98346111.91
β_7 = 27192743.26
β_8 = 0.026

The mathematical regression model is obtained as:

$$y = 2019687.48 + 177.71x\_1 + 218.255x\_2 + 193.70x\_3 + 0.077x\_4 - 5450242.21x\_5 - 98346111.91x\_6 + 27192743.26x\_7 + 0.026x\_8$$
(20)

The less significant variables are rejected one by one based on P-value. The significant indicator $\alpha = 0.1$ is considered as the screening index. When P – value > 0.1, the item is the less significant items and should be removed. Otherwise, the result is opposite. For instance, the P – value for $x8$ which is oil output last years is 0.9008 in the first round screening and P – value > 0.1, therefore, this item should be rejected. After 5 rounds screening, the variables rejected are as follows:

$x1$ = The total numbers of wells
$x3$ = The number of new adding wells
$x6$ = The oil production rate of previous year
$x8$ = The oil output of previous year

From the calculation, the P – value of all variables left satisfy the significance requirements of $\alpha = 0.1$. After the screening of P – value, the four most important factors which affect the oilfield output are determined. They are (in most significant order) :

$x2$ = The start up number of wells
$x7$ = The recovery percent of previous year
$x4$ = The injected water volume last year
$x5$ = The oil moisture content of previous year

Therefore, the new mathematical model is obtained with the screening of P – value. The model may written as

$$y = -259910 + 352.2079\,x_2 + 0.123018\,x_4 - 3660564\,x_5 + 27702445\,x_7 \quad (21)$$

After the result obtained, two kinds of model (four parameters model and eight parameters model) were compared to see the error differences. (refer to ***Table 2***)

Table 2 - The Comparison Of Prediction Results Of Two Models

Year	Actual Output	Four Parameters Output	Error (%)	Eight Parameters Output	Error (%)
2000	4712500	4755212	0.413101	4819842	1.191005
2001	5205000	5197864	0.069019	5180056	0.276769
2002	6115500	6155542	0.387279	6169154	0.595312
2003	7158700	7146255	0.120367	7195551	0.408875
2004	8109500	8030987	0.759365	7966909	1.582103
2005	9051000	8856198	1.884093	8956653	1.046812
2006	9623000	9822647	1.93096	9752501	1.436868
Average Total			5.564183		6.537743

Thus, to verify this method the author obtain new list data parameters obtain from another China’s oilfield (Please refer to ***Table 3***) but by only using 4 parameters that have the most significant value in calculation that were made using the first data. After the screening of P – value in the new set of data, the P- value for all parameters still satisfy the significant requirement $\alpha = 0.1$ which is P – value > 0.1.

Thus, from the result obtain, the author calculate the percentage error from the latest model. We can see that from ***Table 4*** that the total percentage error is less than 4.57%. This validate that the MLR method can be use to forecast oilfield data.

Table 3 - Parameters Data From China Oilfield 2

year	y_{new}	$x2_{new}$	$x4_{new}$	$x5_{new}$	$x7_{new}$
1983	1352300	407	1564500	40.96%	8.92%
1984	1326700	515	1493600	41.49%	9.39%
1985	1375600	669	1954500	42.09%	9.34%
1986	1364000	882	2495000	45.37%	10.10%
1987	1398900	992	3170000	44.96%	9.20%
1988	1468700	1212	3739600	46.96%	8.93%
1989	1561800	1319	4457700	48.46%	9.16%
1990	1934100	1556	5209000	51.31%	9.98%
1991	2085400	1698	6594800	54.62%	10.73%
1992	2515900	1918	7865100	58.99%	11.39%
1993	2934800	2369	9068400	60.03%	11.92%
1994	3402600	2621	10297300	62.55%	12.81%
1995	3635300	2673	11021300	62.28%	13.42%
1996	3947100	2797	12280400	63.95%	14.61%
1997	4110000	2967	13251700	66.61%	14.44%
1998	4307700	3055	14949200	68.05%	14.73%
1999	4559200	3170	15948900	69.28%	15.25%
2000	4622000	3292	15707600	71.04%	15.67%
2001	5114500	3499	17272000	71.04%	16.31%
2002	6025000	5318	18455900	72.11%	17.07%
2003	7068200	7600	18769100	71.99%	17.59%
2004	8019000	8058	24553600	71.44%	17.56%
2005	8960500	9317	29220600	71.17%	16.83%
2006	9532500	10887	32175600	71.47%	16.61%

Table 4 - Coefficients Table For Latest Model

Year	Actual Output	Latest Model	Error (%)
2000	4712500	4660515.787	0.408306
2001	5205000	5108470.348	0.06392
2002	6115500	6067627.998	0.4519
2003	7158700	7058753.792	0.100139
2004	8109500	7968840.953	0.531737
2005	9051000	8815560.913	1.536501
2006	9623000	9672381.567	1.482886
Average Total			4.57539

CONCLUSION

As for the conclusion, the variables that affecting the performance of oilfield's output has been identified and the full calculation were already constructed in order to find the value for regression coefficient, β and to predict the output of oilfield. By implementing this method, output of oilfield can also obtained by using MATLAB simulation. Since there are too many variables that affecting the performance of oilfield's output, the author used the ordinary least squares method to minimize the sum of variables by eliminating the least important variables. The author also uses different set of data with the most significant parameters that have been identified in first calculation using first set of data to do some comparison and verified the validity of this method. From the result and discussion it shown that the percentage error of predicted y value from the actual output is only 4.57%. This validate that this method can be implement to forecast the oilfield output.

REFERENCES

Princeton University. (2003). *The Definition Of an Oilfield.* Retrieved September 19, 2010, from http://www.thefreedictionary.com/oilfield

Daniel Yergin. 1999. *"The Prize: The Epic Quest for Oil, Money, and Power",* United Sates, Simon Schuster

Ling Guo and Xianghui Dei , 2009, "Application of Improved Multiple Linear Regression Method in Oilfield Output Forecasting," *International Conference on Information Management, Innovation Management and Industrial Engineering.*

Larry W. Lake. 2007. "Petroleum Engineering Handbook Volume VII",United States, Society of Petroleum Engineers.

Changjun Zhu and Xiujuan Zhao, 2009, "Application of Artificial Neural Network in the Prediction of Output in Oilfield," *International Joint Conference on Artificial Intelligence*

Chen Yuanqian, Hu Jianguo and Zhang Dongjie. 1999. *Derivation of Logistic Model and Its Self-Regression Method*, China, Citic Press.

Wang Junkui. 1993. *A Theoretical Discussion of Production Decline Curve of Oil and Gas Reservoirs*, China, Petroleum Exploration and Development.

Wang Tao, Chen Xiang Guang and Li Yu Feng. 2006. *Optimization of Multivariate Model in Oilfield Output Prediction*, New York, Computer Simulation.

Massachusetts Institutes of Technology. (2006). Retrieved September 19, 2010, from http://ocw.mit.edu/courses/sloan-school-of-management/15-075-applied-statistics-spring-2003/lecture-notes/lec14and15_chap11.pdf.

Kutner, Natchsheim and Neter. 2004. *Applied Linear Regresion Model*,(4), New York, Mc Graw Hill