GEO India 2018

PaperID **AU320**

Author **Ajit Kumar Sahoo , Reliance Industries Ltd, Indian Institute of Technology, Bombay , India**

Co-Authors **Arun K. Behera, Mukul Srivastava and Vikram Vishal**

Predict the Production: A Data Analytic Approach

Ajit K. Sahoo[1, 2], Arun K. Behera[1], Mukul Srivastava[1] and Vikram Vishal[2]

1. E&P Business, Reliance Industries Ltd, Navimumbai-400701, India.
2. Computation and Experimental Geomechanics Laboratory, Department of Earth Sciences, IIT Bombay, Mumbai-400076, India

Contact Author: ajit.sahoo@ril.com, ajit.sahoo1980@gmail.com

Abstract

Multivariate Analytics (MVA) is becoming favorite of many data analyst in the oil and gas industry. It is becoming a powerful tool to assess the individual impact of geologic, completion, and well design variables on shale gas well performance. In this paper, we have tried to bring out a crisp and handy workflow of building MVA predictive models to predict the production of shale reservoir.

A three segment workflow is proposed in this study. In the first segment we have discussed about the input data preparation, selection, and cleaning. Second segments explains about the model building and its quality check. In the last segment, we have discussed about how to cross validate the model and its stability.

We reviewed 158 wells to extract several geological, production, drilling, and completion information. Geological parameters like thickness, volume of clay, porosity etc., production information like choke size, hydro carbon yield etc. and D&C information like lateral length, stage spacing, cluster spacing, proppant volume etc. are extracted and treated as predictors to predict the normalized 12 month cumulative production through MVA predictive model. Statistically significant predictor are initially identified based t-value and p-value. Outlier and multi-collinearity analysis are used in cleansing the input data. Thereafter on the basis of sound technical understanding, the important set of predictors are finalized and used in the MVA model generation. Linearity of the MVA model is carried out by analyzing the Normal Q-Q, Residuals Vs Fitted and Residual Vs Leverage plots.

In the studied shale play, we found that the geological parameters like reservoir thickness, molybdenum amount, reservoir engineering parameters like choke size, hydrocarbon yield, well head pressure and D&C parameters like cluster spacing, average bpm, average psi are highly correlated with normalized 12 month cumulative production. An equation is derived from the model based on the intercept and coefficients to predict the 12 month cum production. Model is found to be handy and stable on different test data sets.

Key Words: Shale Gas Production, Data Analytics, Workflow of Multivariate Analysis.

Introduction:

Since 1998 unconventional natural gas production has increased nearly 65% in United States. This growth has resulted in unconventional production becoming an increasingly larger portion of total natural gas production in US, increasing from 23% in 2010 to projected 49% of total dry gas production in 2035[U.S. EIA, 2012]. The development of shale gas production was prompted by technological advances particularly concerning horizontal drilling and hydraulic fracturing. While these practices have led to the economical production of natural gases in numerous shale gas reservoirs, the problem of understanding shale gas production has been much involved due to the complicated and unpredicted response of these reservoirs to fluid and proppant injection. Besides, each of the shale gas properties

such as thickness of the productive layer and geomechanical properties of rock vary substantially within the same producing area and this variability of shale gas properties greatly influences the well performance (Esmailli et al, 2015).

Given the complex nature of hydraulic-fracture growth and the very low permeability of the matrix rock in many shale gas reservoirs, in combination with the predominance of horizontal completions, reservoir simulation is commonly the preferred method to predict and evaluate well performance (Cipolla, 2010). But dealing with a reservoir with more than hundreds of wells makes this process inefficient especially when it comes to short-term reservoir management decision making step (Esmailli et al, 2015).

Understanding the geological and engineering drivers behind the performance of such ultra-low permeable formation is challenging. Several studies have addressed the impact of rock properties and the effect of hydraulic fracturing process on well performance through different methodologies. Many authors have used the multivariate analysis (MVA) tool to predict the production. Step by step explanation of the entire workflow is rarely published. In this study, we have attempted to explain the workflow in a detailed manner.

Data Preparation:

Shale play development generates huge volume of data of different variety. It is an important step in MVA to know which data needs to be considered and which one needs to be ignored. This process mainly involves:

i) Selection of Wells: A group of producing wells needs to be selected which haven't been affected by any operational issues like depletion, frac hit, and high water cut etc. We have selected 158 wells for this study which haven't experienced any operational issue.

ii) Selection of Production Data: Most of the workers consider estimated ultimate recovery (EUR). Since EUR is itself a predicted information, so it is better to consider the cumulative production instead of EUR. Production data needs to be normalized on the basis of lateral length in the individual wells. In our study we have considered 12 month normalized production as the "Y axis-response" which needs to be predicted for the future wells. Apart from production data, some other important reservoir engineering parameters choke size, tubing head pressure, well head pressure, hydrocarbon yield (CGR) are considered as predictors for this study.

iii) Extraction of Subsurface Information for the producing wells: In shale play, all the producing wells are not pilot wells and don't have rock quality information. So the different rock quality information were extracted from the maps that are prepared using pilot well information. Geological information like Thickness, total organic carbon, porosity, water saturation, hydrocarbon filled porosity, volume of clay, brittleness, and volume of molybdenum were extracted and considered as "X axis-predictors".

iv) Extraction of D&C information: D&C information like effective lateral length, stage spacing, number of clusters, cluster spacing, proppant volume, total fluid volume, average pressure, ISIP, fracture gradient, for all the selected producing wells are reviewed and extracted carefully.

Methodology:

Multivariate analysis is used because of its ability to accommodate many variables. Regression models were developed in the study area to capture the most co-relatable geological and engineering parameters with well productivity (Centurion et al, 2014). The following flowchart shows the steps followed to achieve the multivariate model. We have used 12 month normalized production as response and all other geological & engineering parameters as variables. For multivariate modeling, we have used "R-Studio" statistical tool due to its better QC ability. It is critical to understand the importance of all the above steps, so we have described them in detail.

Step1: Predictors Significance Check with Initial Model

It is always better to initialize the linear model with all the expected predictors against the response and let the statistical process itself finds-out the significant ones. The initial least square linear regression supposed to answer two things: First, is there any overall relationship between the response and predictor at all? Second if yes, then which are those predictors? In statistics, these questions are answered by using hypothesis tests: null hypothesis & alternative hypothesis with the help of F-stat, t-stat and p-value. In general, F-stat value >1, absolute t-stat value >1 & P-stat value < 0.05 reject the null hypothesis & indicates an overall good relationship between response and predictors. The F-stat value in the initial model discussed in this study is greater than 1. The significant predictors with absolute t-value >1 and p-value <0.05 are indicated by '*' symbol. The more the numbers of '*', the more significant is the predictor (Figure-2).

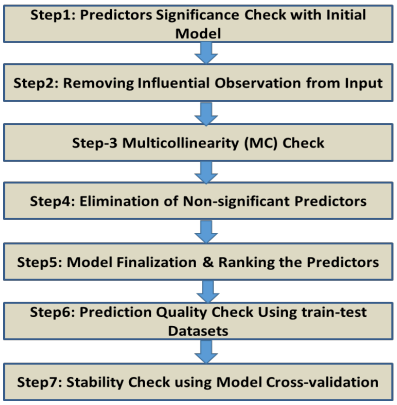


Figure-1: Steps involved in multivariate model

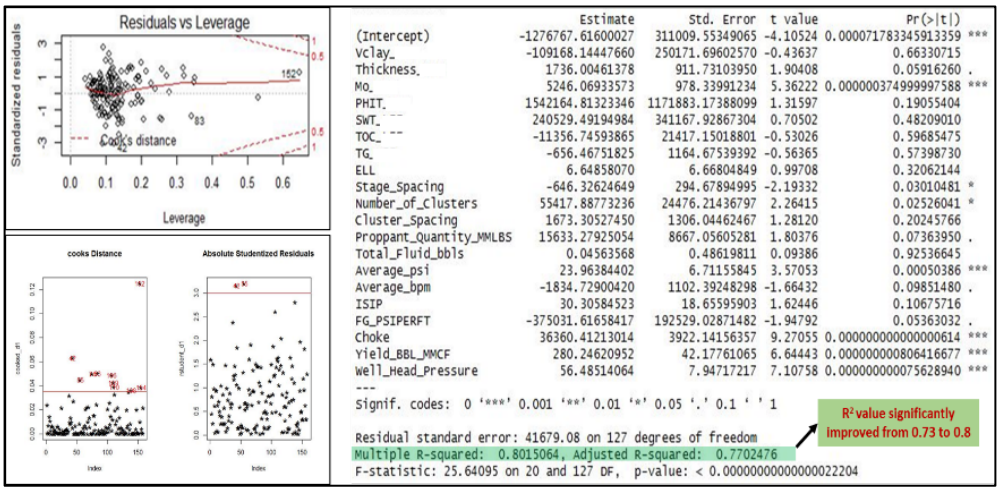
```
Coefficients:
                          Estimate        Std. Error  t value             Pr(>|t|)
(Intercept)          -1588825.17875137   317377.57217392 -5.00610 0.000001678093497756 ***
Vclay                   56166.56209189   284373.22091239  0.19751           0.84372118
Thickness                1838.13859779     1014.66525512  1.81157           0.07224290 .
Mo                       4808.13993852     1128.84954679  4.25933 0.000037859890503485 ***
PHIT                  1524318.22741450  1311855.76944543  1.16196           0.24727376
SWT                     64814.48025841   388481.13285944  0.16684           0.86774136
TOC                      5679.13638748    24032.59474222  0.23631           0.81354526
TG                       -983.70956951     1307.01828657 -0.75264           0.45295971
ELL                        12.19292184        7.72143039  1.57910           0.11661895
Stage_Spacing            -697.55959000      329.33090449 -2.11811           0.03597131 *
Number_of_Clusters      63197.97924158    26431.47689923  2.39101           0.01816112 *
Cluster_Spacing          1587.99774831     1381.87588967  1.14916           0.25249193
Proppant_Quantity_MMLBS 13917.53824693     9192.64580156  1.51399           0.13233300
Total_Fluid_bbls           -0.06660907        0.51608799 -0.12907           0.89749523
Average_psi                27.31825148        7.57940648  3.60427           0.00043714 ***
Average_bpm              -661.04071186     1054.19623322 -0.62706           0.53166574
ISIP                        9.74163799       13.67864404  0.71218           0.47756524
FG_PSIPERFT           -116672.46966317   188450.56451934 -0.61911           0.53686912
Choke                   36907.37070345     4417.48333674  8.35484 0.000000000000064917 ***
Yield_BBL_MMCF            253.33634106       48.23131600  5.25253 0.000000559979260535 ***
Well_Head_Pressure         60.30708467        8.56284881  7.04288 0.000000000083143384 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49230.59 on 137 degrees of freedom
Multiple R-squared:  0.7363527, Adjusted R-squared:  0.697864
F-statistic: 19.13168 on 20 and 137 DF,  p-value: < 0.00000000000000022204
```

Figure-2: Outcome of Step-1 indicating the initial relationship between response and predictors

Step2: Removing Influential Observation from Input

Influential observations/Outliers detection and their removal is an important step in statistical modelling as they may alter the t-stat, P-stat values as well as the regression coefficients significantly. In this case study, we have used residual vs. leverage, cook's distance and absolute studentized residual plots to identify the outliers. In the residuals vs leverage plot, no sample point fell beyond the cook's distance > 0.5 contour. Then, we tried one step ahead. We followed the criteria: cook's distance of single observation > 4* mean value of cook's distances of all observation and absolute value studentized residual >3. The observations are marked for further analysis and masked in the next level of regression

model in order to see the improvements. R2 value of the model significantly improved. Some more predictors became significant (Figure-3).



	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1276767.61600027	311009.55349065	-4.10524	0.000071783345913359	***
Vclay_	-109168.14447660	250171.69602570	-0.43637	0.66330715	
Thickness_	1736.00461378	911.73103950	1.90408	0.05916260	.
Mo_	5246.06933573	978.33991234	5.36222	0.000000374999997588	***
PHIT_	1542164.81323346	1171883.17388099	1.31597	0.19055404	
SWT_	240529.49194984	341167.92867304	0.70502	0.48209010	
TOC_	-11356.74593865	21417.15018801	-0.53026	0.59685475	
TG_	-656.46751825	1164.67539392	-0.56365	0.57398730	
ELL	6.64858070	6.66804849	0.99708	0.32062144	
Stage_Spacing	-646.32624649	294.67894995	-2.19332	0.03010481	*
Number_of_Clusters	55417.88773236	24476.21436797	2.26415	0.02526041	*
Cluster_Spacing	1673.30527450	1306.04462467	1.28120	0.20245766	
Proppant_Quantity_MMLBS	15633.27925054	8667.05605281	1.80376	0.07363950	.
Total_Fluid_bbls	0.04563568	0.48619811	0.09386	0.92536645	
Average_psi	23.96384402	6.71155845	3.57053	0.00050386	***
Average_bpm	-1834.72900420	1102.39248298	-1.66432	0.09851480	.
ISIP	30.30584523	18.65595903	1.62446	0.10675716	
FG_PSIPERFT	-375031.61658417	192529.02871482	-1.94792	0.05363032	.
Choke	36360.41213014	3922.14156357	9.27055	0.000000000000000614	***
Yield_BBL_MMCF	280.24620952	42.17761065	6.64443	0.000000000806416677	***
Well_Head_Pressure	56.48514064	7.94717217	7.10758	0.000000000075628940	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41679.08 on 127 degrees of freedom

Multiple R-squared: 0.8015064, Adjusted R-squared: 0.7702476

F-statistic: 25.64095 on 20 and 127 DF, p-value: < 0.00000000000000022204

Figure-3; Outcome of Step-2 showing outlier detection and the impact on R^2 after removal.

Step3: Multi-collinearity (MC) Check

One of the important assumptions of multiple linear regression model is that the predictors are linearly independent. Collinearity between the predictors can affect R^2 and prediction quality. So, it is essential to find out the collinear predictors and then to decide the predictors to go ahead in the next level.

In this case study, MC is analyzed using Variance Inflation Factor (VIF) and co-relation co-efficient. VIF of a single predictor is the ratio of variance in the model with all the predictors, divided by variance in the model with that single predictor. More importance is given to VIF value of each predictor. VIF >5 indicates the existence of multicollinearity (Figure-4).

	VIF
Vclay_	2.2903305
Thickness_	13.2595345
Mo_	3.5946996
PHIT_	8.9222177
SWT_	4.4809199
TOC_	2.7218480
TG_	27.0344515
ELL	2.8695111
Stage_Spacing	23.3698593
Number_of_Clusters	22.3103217
Cluster_Spacing	17.2499336
Proppant_Quantity_MMLBS	9.4691260
Total_Fluid_bbls	12.3955184
Average_psi	1.5623154
Average_bpm	1.9721874
ISIP	4.6814620
FG_PSIPERFT	3.2693869
Choke	2.6456663
Yield_BBL_MMCF	3.7624495
Well_Head_Pressure	3.1185646
FIRST12_BOE	5.0379455

Following Parameters removed from the model due to multicollinearity:

- PhiT
- Total Gas (TG)
- No. of Clusters
- Total fluid volume

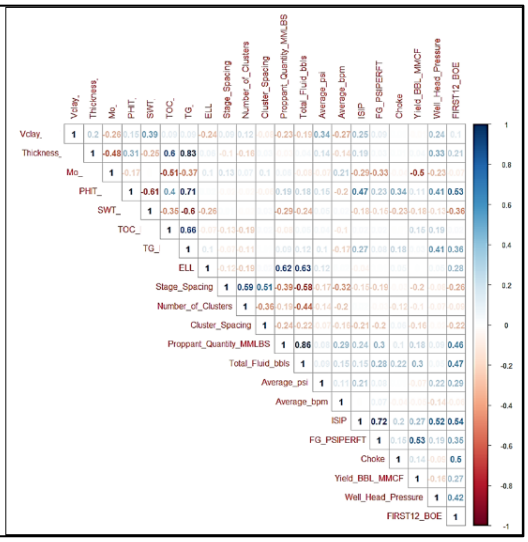


Figure-4: Multicollinearity check using VIF and Co-relation co-efficient

Step4: Elimination of Non-significant Predictors

Once the multi collinearity is managed by removing appropriate predictors, least square regression modeling is carried out using the remaining predictors and response. As the number of predictors

reduces in the process, a decrease in R2 value is expected. Now, based on the t-stat & P-stat values, the non-significant predictors are removed from the model one by one. Each time modeling is carried out and changes in the results are observed.

Step5: Predictor Finalization

Following the above four steps, a final regression model is run with the remaining predictors and number of observations. R2, adjusted R2 and residual standard errors are analyzed.

But before finalizing the model, one should be aware of the assumptions behind the algorithm used in the process and validate through proper QC procedures. Some of the important assumptions of multiple linear regression model and QC validations are discussed below (Figure-5):

1. *Residual Vs Fitted Plot:* If the model predicts the response with 100% accuracy, all the observations will fall on zero residual red line as shown in the plot. But in practice, if observations are distributed uniformly across the zero-residual line & randomly, then the above assumption is validated.
2. *Normal quantile-quantile (Q-Q plot):* A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. A straight line mostly indicate a linear relationship between the response and the predictor.
3. The residuals are equally distributed along all the predictors. This assumption is validated by plotting the fitted values against square root of standardized residuals. The observation should be randomly distributed on both side of a straight line trend to residual axis.
4. No outliers/Influential observation: This is validated by plotting leverage against standardized residual as discussed in step2. The standardized residual should be less than 3. Observations beyond cook's distance greater than 0.5 should be taken care for better modelling result.

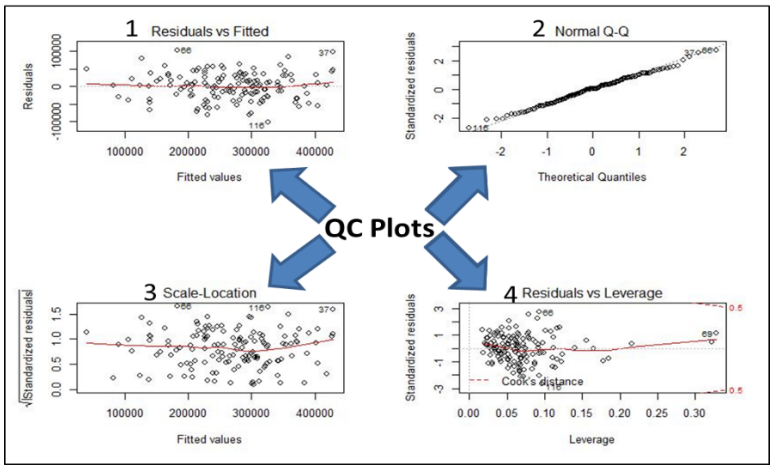


Figure-5: QC plots for linearity check.

Step6: Prediction Quality Check Using train-test Datasets

The measure of standard residual error & R2 are not sufficient enough to comment about the prediction quality of the model, as all the observations are used in the exercise to minimize the residuals. The ultimate test of prediction quality is to use any blind datasets not used in modeling and predict the response. A common approach is to divide the whole datasets into two parts randomly: train data & test data. Modeling is carried out using train data and prediction is done using test data over the model. Then Mean Square Error (MSE) is analyzed.

In this case study, all the observations are randomly split into train and test data with 80% & 20% ratio respectively. Modeling is done using train data & all QC steps are followed. Using the model, the responses are predicted using test data and compared with the true responses. MSE is calculated. The model predicts the response with an average error of 12.8% and a very good correlation coefficient of 80% (figure-6).

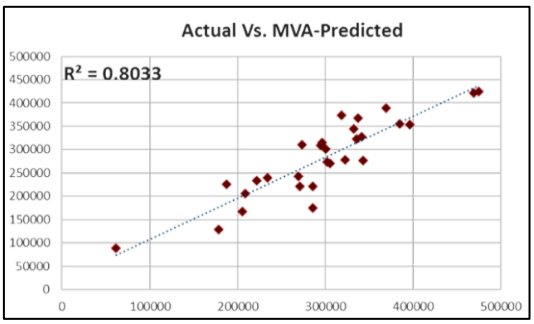


Figure-6: Actual vs. Predicted 12 month production.

Step7: Stability Check using Model Cross-validation

As the model fitting depends on the chosen observations, randomly dividing the data into train and test sets and then to get minimum average error in prediction should not be a matter of chance. So, stability of the model should be checked through cross-validation. The observations should be split into train and test sets a number of times and the process in step 6 should be followed. In this case study, models are generated with five randomly selected train datasets. Predictions are done using corresponding test sets. Average MSE is calculated and compared with that from step6 and found to be nearly equal which indicates that the multivariate model generated for this data set is stable.

Modeling Results & Summary:

Initially we took 20 variable to predict the 12 month cumulative production, but after finalizing the model we found only nine parameters are driving the well performance and can be used to predict the production. Finally we came up with the following equation for the study area which can be used to predict the 12 month cumulative production.

12 month (BOE)= 1183*Thickness + 5674*Molybdenum - 1377*Cluster spacing + 19412 Proppant volume + 26* Treating pressure – 3403*BPM + 40607* Choke + 276* CGR + 62* WH pressure-995384

Geological parameters like, reservoir thickness, molybdenum (indicator of TOC) are found to have direct correlation with production. Completion parameters like proppant volume, treating pressure have positive relation whereas cluster spacing, average BPM are showing negative relationship with production. Reservoir parameters like choke size, CGR, and well head pressure have direct and positive impact on well performance. If multivariate analysis is carried in a proper way, it is not only going to be useful to predict the production, but also identifies the static and dynamic drivers that affects the shale well productivity.

Acknowledgement: Authors are grateful to Mr. Neeraj Sinha for his critical inputs and the management of Reliance Industries Ltd for giving permission to publish this paper.

References:

Centurion Sergio, Junca-Laplace Jean-Philippe, and Randall Cade; Lessons Learned From an Eagle Ford Shale Completion Evaluation, SPE-170827-MS, presented at the SPE Annual Technical Conference and Exhibition held in Amsterdam, The Netherlands, 27–29 October 2014.

Cipolla, C., Mack M., and Maxwell, S., Reducing Exploration and Appraisal Risk in Low Permeability Reservoirs Using Microseismic Fracture Mapping, Part 2. SPE 138103 presented at the SPE Latin American & Caribbean Petroleum Engineering Conference, Lima, Peru, and December 2010.

Esmailli S, Mohaghegh. S.D, Dahaghi A.K., Shale Asset Production Evaluation by Pattern Recognition, SPE 174061, presented at the SPE Western Regional Meeting held in Garden Grove, California, USA, 27-30 April, 2015.