CADERNOS DO IME – Série Estatística

Universidade do Estado do Rio de Janeiro - UERJ
ISSN impresso 1413-9022 / ISSN on-line 2317-4536 - v.40, p.34 - 45, 2016
DOI: 10.12957/cadest.2016.27711

COST ESTIMATING RELATIONSHIPS IN ONSHORE DRILLING PROJECTS

Ricardo de Melo e Silva Accioly
Departamento de Estatística – IME/UERJ
raccioly@ime.uerj.br

Abstract

Cost estimating relationships (CERs) are very important tools in the planning phases of an upstream project. CERs are, in general, multiple regression models developed to estimate the cost of a particular item or scope of a project. They are based in historical data that should pass through a normalization process before fitting a model. In the early phases they are the primary tool for cost estimating. In later phases they are usually used as an estimation validation tool and sometimes for benchmarking purposes. As in any other modeling methodology there are number of important steps to build a model. In this paper the process of building a CER to estimate drilling cost of onshore wells will be addressed.

Keywords*: Cost Estimating Relationships. Onshore wells. Upstream projects.*

1. Introduction

Cost estimating plays an important role in any project, especially when it involves large amounts of money like in upstream projects. An onshore upstream project usually involves two main scopes of work: facilities and wells. In this paper, we will focus on wells part.

In the initial phases of any project there is little information available, but it is still necessary to have cost estimates. Cost estimating relationships (CERs) are the best choice in these cases, since they are based in historical data of similar projects. Through a careful choice of input variables, it is possible to create a multivariable regression model that could give good estimates for a specific scope. If the developed model includes all relevant drivers, it is possible to use it not only for initial cost estimates but also for estimation validation and benchmarking purposes.

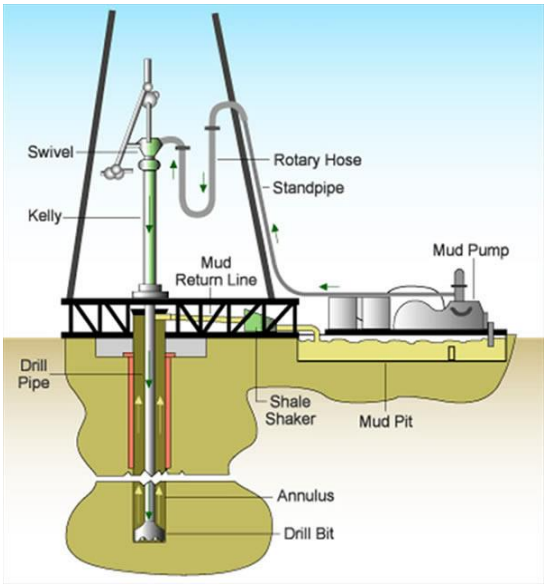
The second section will discuss drilling costs drivers of an onshore well. Ideally, every key characteristic from an onshore well is a potential candidate for testing and verification if it is a statistically significant cost driver or not. Of course, such things are not feasible because it is always difficult to have enough reliable data to use. In this section, will also be treated data collection, cleaning and normalization. When collecting data to build a model a lot of care should be taken. Since we are going to use historical data we need to be sure that they can be used as a basis of future costs. Data normalization is another important issue, especially when involving cost data, which requires the use of a specific cost index.

Section three will discuss some import issues when building a cost estimating relationship (CER). In section four will discuss the development of drilling cost CER for an onshore well. Section five will show some conclusions.

2. Cost Drivers Identification, Data Collection, Cleaning and Normalization

The main cost driver in onshore drilling is the rig, which usually represents a daily cost. Other important elements that have influence in the cost are well depth, number of casing strings and well geometry. In Figure 1 you can see a schematic of an onshore rig drilling a well.

Figure 1 - Onshore Drilling Rig



Source: Petroleum Online (2016).

Since drilling rig costs are daily rates, drilling duration is the main concern when estimating drilling costs. The other costs can be estimated in a well length basis (US$/m). For this paper purpose, i.e. to build a simple model, only two cost drivers from onshore wells will be addressed: well depth and the number of casing strings.

Data characterization is very important so it is necessary to identify exploratory wells and development ones. Exploratory wells are the first wells drilled in one region, so they are intrinsically different, because they have different objectives and knowledge about the area. Development wells are designed with the information gained from the exploratory ones and in this sense, they are optimized versions. Another important aspect is the well geometry: vertical, directional and horizontal. These types of wells are increasingly difficult to drill requiring special tools and services. Only directional wells will be used in this paper.

Of course, there are other cost drivers that could be analyzed, but the main objective of this paper is to present the ideas behind this methodology. For a more extensive analysis of drilling costs the reader should consult Kaiser and Pulsipher (2007) and Augustine *et al.* (2006).

Data collection is a critical part of every data modeling problem. When dealing with cost data, extra precaution should be taken, cost should always walk hand by hand with its realization (or estimation) date. When you have a relatively large data set

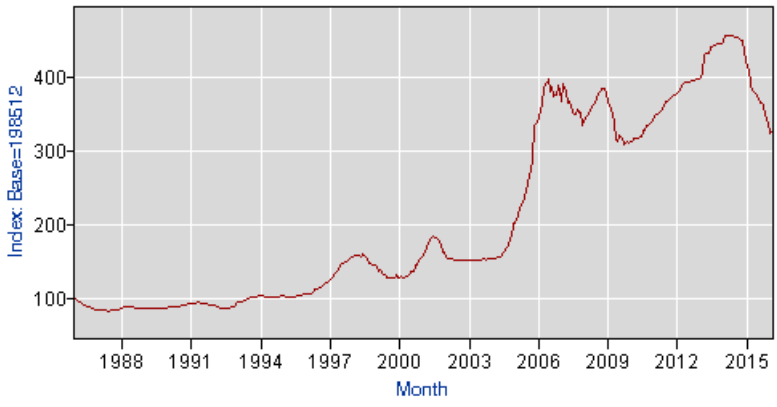
sometimes is possible to identify trends over different drilling programs, which will bring valuable information about its drivers.

Raw data always have some problems that need to be identified and diagnosed. Some missing information or a very different behavior of a particular data point, in each case there is a statistical procedure that could help us identify these problems. Summary statistics, box-plots, histograms and Q-Q plots will always be part of the toolset necessary to do this job, for an extensive description of techniques see Chambers *et al.* (1983).

The process of making the data comparable and consistent within different drilling programs used in model building is called data normalization. Two types of data normalization were used in this work: normalization by cost and by well geometry grouping.

Cost normalization should bring every cost to the same data base and unit. So, it is necessary to normalize the inflation into or out of the data, depending which date you choose to work. That process is called escalate or de-escalate and should be made with a proper index. One cannot use any inflation index to de-escalate, it is necessary to work with a sectorial inflation index, in this case the Producer Price Index (PPI) for drilling oil and gas wells that can be obtained from Bureau of Labor Statistics site. In Figure 2 below it is possible to see the huge variation from 1985 to 2016, showing clearly the need to de-escalate cost data before building a model.

Figure 1 – PPI index from drilling oil and gas wells.



Source: Bureu of Labor Statistics (2016).

The other normalization was just selecting a group of comparable wells, for example, similar deviation, depths, rigs, so that we have a homogenous sample of onshore wells.

3. Regression Models and Model Selection

There are several types of regression models that could be used to build a CER. The set ranges from simple regressions through the origin to multivariate linear regression models. Only ordinary least squares (OLS) will be considered in model fitting, but there are other ways to estimate the regression parameters (Draper and Smith, 1998).

A starting point to begin is to assume that relationship between the independent and dependent variable is linear, which means that we will be using the simple regression equation as follows (1):

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

Where:

- y represents the dependent variable;
- χ represents the independent variable;
- β_0 is linear coefficient;
- β_1 is the slope of the regression;
- ε is the error term with $E(\varepsilon) = 0$ is additive.

This simple regression model could be directly extended to a multivariate version that includes other explanatory (independent) variables.

When fitting a model a common situation is the use of transformed variables to respect some models assumptions like linear relationship and random errors. Two useful simple nonlinear relationships, that could be linearized, are the exponential growth model and the power model.

The functional form of an exponential model is (2):

$$y = \beta_0^* e^{\beta_1 x} u \tag{2}$$

Where $\beta_0^* = e^{\beta_0}$ e $u = e^{\varepsilon}$ with a linear form (3):

$$\ln(y) = \beta_0 + \beta_1 x + \varepsilon \tag{3}$$

The functional form of a power model is (4):

$$y = \beta_0^* x^{\beta_1} u \tag{4}$$

Where $\beta_0^* = e^{\beta_0}$ e $u = e^{\varepsilon}$ with a linear form (5):

$$\ln(y) = \beta_0 + \beta_1 \ln(x) + \ln(u) \tag{5}$$

In both cases u is the error term and it is a multiplicative error term which differs from the simple regression model. This model is fitted using it transformed version, in the log space, where the log error term is additive. When transforming it back to the unit space it is necessary to apply a correction as shown by Miller (1984). He showed that in the unit space the estimator provided a consistent estimator of median response, underestimating the mean response. So, when transforming back it is necessary to use a multiplicative correction factor where is the standard error estimate (SEE) of the linearized model.

Model selection is the next step. Always start with a minimal model choosing significant variables. This is known as parsimony principle. When comparing within the same model family, for example a linear model with another linear model, the coefficient of determination, R2, is a good proxy. When comparing models from different families the SEE in the unit space should be the criterion. For other important considerations about model selection refer to Applied Regression Analysis (Draper and Smith, 1998).

Regression diagnostics are another important issue when developing regression models. They are important part of model validation and verification. In this subject there a lot of good references, for this paper the books Regression Diagnostics: Identifying Influential Data and Sources of Collinearity (Belsley, Kuh and Welsch, 1980) and An R Companion to Applied Regression (Fox and Weisberg, 2011) are the main source of usual techniques.

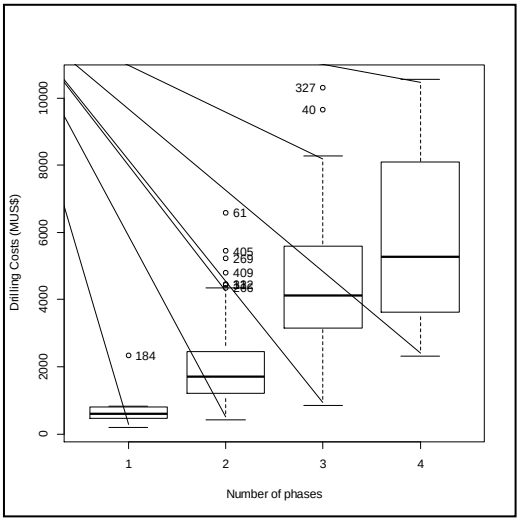
4. Case Study

In this case study data from 418 onshore wells were initially selected. The wells came from the same region and all of them are directional. This data set includes drilling costs, well depth and number of phases (casing strings). All cost data was normalized to the same data base. Considering the need to protect the proprietary nature of the data supplied, the dependent variable, drilling cost, was multiplied by a random factor.

An initial exploratory data analysis is always a good start. Figure 3 shows the box plot relating drilling costs with the number of phases or number of casing strings. The cost increases as the number of phases increases. It is also possible to identify some

outliers in the data set. These outliers should be analyzed because they could represent a data information problem or wells with some unusual operational problems. In this data set the reason for this different behavior was related to operational problems, for example, fishing problems and loss of circulation. In general, these anomalies should receive a different treatment for instance through a risk analysis, but these methods are beyond the scope of this paper.

Figure 3: Box Plot of Drilling Cost ~ Number of phases



Source: Elaborated by the author.

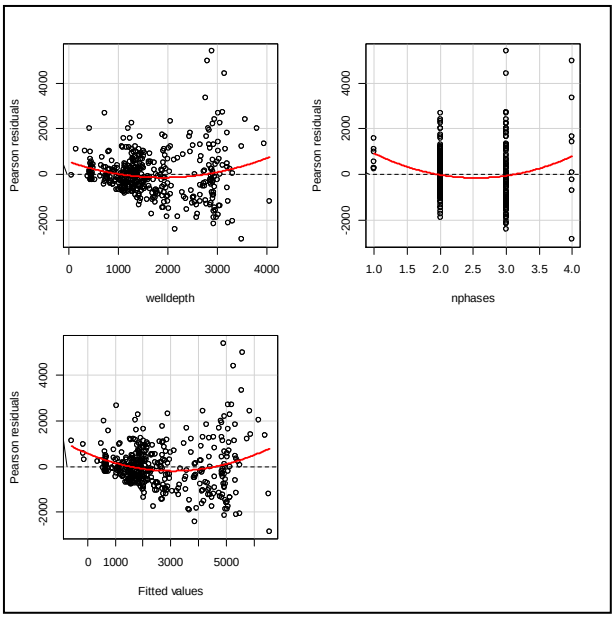
There other useful graphical techniques that could help this initial data screening please refer to An R Companion to Applied Regression (Fox and Weisberg, 2011). As pointed above this data set includes drilling costs, well depth and number of phases so the initial model will include all variables in a simple multiple regression model like bellow (6).

$$Drilling\ Cost = \beta_0 + \beta_1 Well\ Depth + \beta_2 \text{Number of phases} \quad (6)$$

This initial model wasn't selected because some models assumptions were violated, specifically the residuals behavior. In Figure 4 we see the Pearson residuals plotted against each independent variable and the fitted model. These plots show that we have some indication of a nonlinear relationship and a strong indication that the residual variance increases as the independent variable increases (heteroskedastic). The previous presented nonlinear relationships could solve these problems: exponential growth model and power model. The latter is a more common choice in CER modeling but both were

evaluated. For other modeling options, you should refer to Engineering Cost Estimating (Ostwald, 1992).

Figure 4: Residual analysis in simple multiple regression



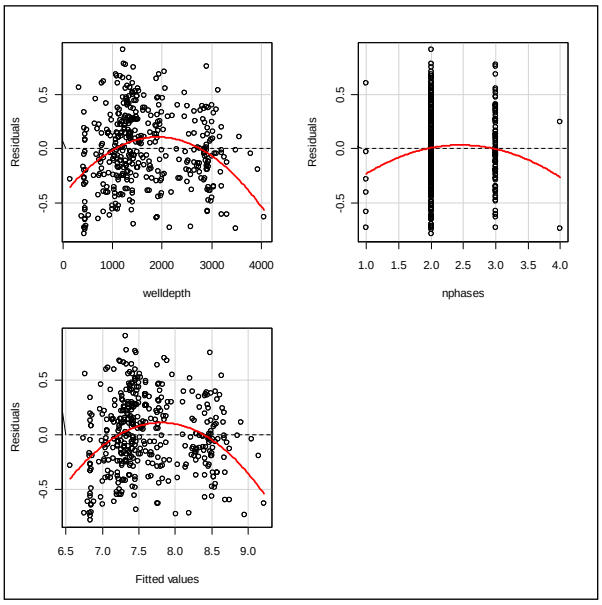
Source: Elaborated by the author.

The exponential growth model was fitted through this expression (7):

$$Drilling\ Cost = \beta_0^* e^{(\beta_1 Well\ Depth + \beta_2 Number\ of\ phases)} \quad (7)$$

After fitting the exponential growth model residual diagnostics indicated that the model wasn't good, heteroskedastic errors disappeared but the nonlinear behavior has increased, so the functional form of the model isn't adequate. Figure 5 shows the residual analysis in the exponential model.

Figure 5: Residual analysis in the exponential model



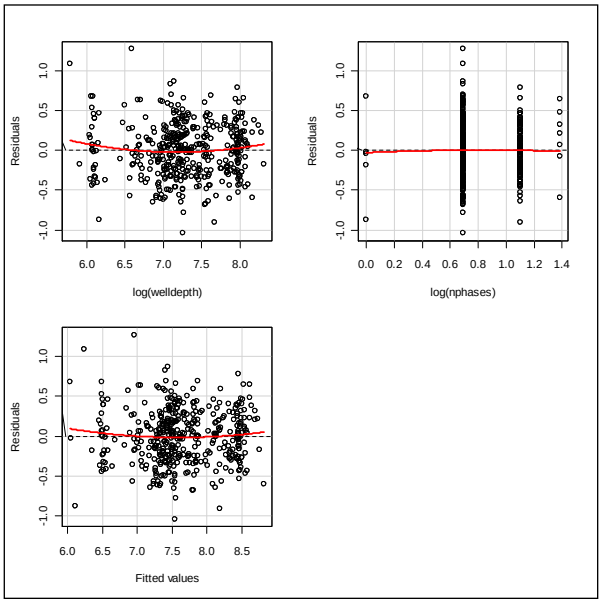
Source: Elaborated by the author.

The power model was fitted through this expression (8):

$$Drilling\ Cost = \beta_0^* Well\ Depth^{\beta_1} Number\ of\ phases^{\beta_2} \quad (8)$$

Residual diagnostics indicated the model was promising (Figure 6), but it is necessary to check if there are influential data points.

Figure 6: Residual analysis in the power model



Source: Elaborated by the author.

Some influential measures could be used to do this job. Hat values are a useful statistic obtained from the hat matrix. They are the diagonal element hi from (9):

$$H = X(X^tX)^{-1}X^t \quad (9)$$

Each hi reflects the influence of a particular data point on the estimated regression coefficients. Higher values of hi indicates an influential data point. A recommended cutoff value is 2(p+1)/n, where p is the number of estimated coefficients excluding the constant and n the sample size (Fox and Weisberg, 2011).

Studentized residuals for observation i is standardized using the hat matrix element hi and the estimated error variance without observation i, s(i) (10).

$$e_{Si} = \frac{e_i}{s(i)\sqrt{1-h_i}} \quad (10)$$

Usually the magnitude 2.0 for e_{Si}will be a screening criterion. Observations with higher values will be considered outliers (Belsley et all, 1980).

Cook's distances are obtained using the previous statistics (11):

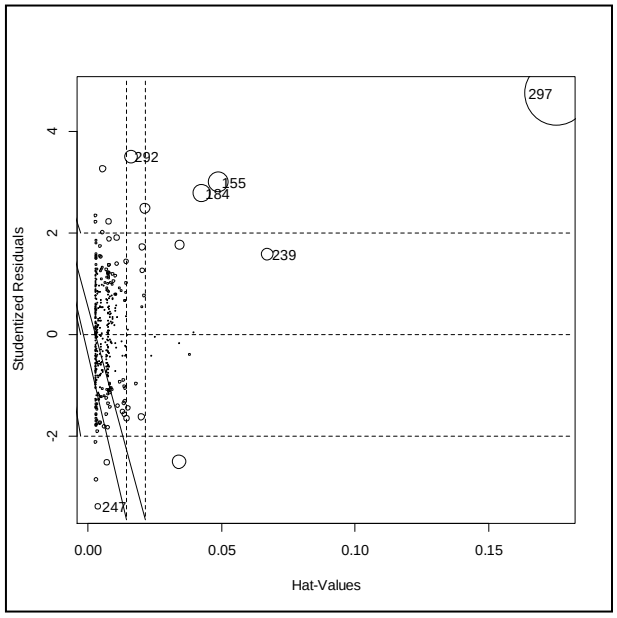
$$D_i = \frac{e_{Si}^2}{p+1} \times \frac{h_i}{1-h_i} \quad (11)$$

The first term is a measure of discrepancy and the second is a measure of leverage.

Figure 7 presents these three measures with the suggested cutoffs (Belsley et all, 1980 and Fox and Weisberg, 2011). The sizes of the circles are proportional to Cook's distance statistic (Fox and Weisberg, 2011).

All these diagnostics were drawn using the R library car that accompany the book An R Companion to Applied Regression (Fox and Weisberg, 2011).

Figure 7: Plot of hat-values, Studentized residuals, and Cook's distances



Source: Elaborated by the author.

After analyzing and removing the influential observations that really represented problematic data, the results were tabulated in Table 1.

Table 1: Regression Statistics

Regression Statistics			
R Square	0,77		
Adjusted R Squared	0,77		
Standard Error	0,311		
Observations	396		
Coefficients			
	Estimate	Std Error	t stat
Intercept	0,308	0,237	1,298
log(welldepth)	0,939	0,038	24,485
log(nphases)	0,613	0,108	5,652

Source: Elaborated by the author.

All coefficients are coherent with the expected behavior since an increase in well depth and in number of phases (casing strings) will result in higher costs.

5. Conclusions

This paper demonstrates how cost estimates relationships (CERs) could be developed using regression analysis. This type of methodology can be used even when there is little information, what makes it very useful. When developing these CERs it is important to check the models assumptions to be sure that it is a good representation of the data.

It is very important for companies to maintain databases with this kind of information because they represent their knowledge that could be transformed in such models or other metrics.

References

AUGUSTINE, C.; TESTER, J.W.; ANDERSON, B.; PETTY, S.; LIVESAY, B., A comparison of geothermal with oil and gas well drilling costs, Thirty-First **Workshop on Geothermal Reservoir Engineering**, Stanford University, 2006.

BELSLEY, D. A.; KUH, E.; WELSH, R. E. **Regression Diagnostics: Identifying Influential Data and Sources of Collinearity**, Wiley, New York, 1980.

BUREAU OF LABOR STATISTICS, Producer Price Indexes: PCU213111213111- Drilling oil and gas wells, Retrieved in 2016, from https://www.bls.gov/ppi/.

CHAMBERS, J.M.; CLEVELAND, W.S.; KLEINER, B.; TUKEY, P.A. **Graphical Methods for data Analysis**, Wadsworth, Belmont, 1983.

DRAPER, N. R.; SMITH, H., **Applied Regression Analysis**, 3ed., Wiley, New Jersey, 1998.

FOX, J.; WEISBERG, S., **An R Companion to Applied Regression**, 2ed., Sage, California, 2011.

KAISER, M.J.; PULSIPHER, A.G. **Generalized Functional Models for Drilling Cost Estimation**, SPE Drilling & Completion, June, 67-73. 2007.

MILLER, D.M. Reducing Transformation Bias in Curve Fitting, **The American Statistician**, 38, 124-126. 1984.

OSTWALD, P.F. **Engineering Cost Estimating**, 3ed., Prentice Hall, New Jersey, 1992.

PETROLEUM ONLINE. Module Drilling & Well Completions. Retrieved in 2016, from www.petroleumonline.com.