Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 13, Number 9 (2017), pp. 5013-5020
© Research India Publications
http://www.ripublication.com

Using stepwise regression to investigate customers' propensity to change cellular phone providers

Mohammad Aljarrah
Department of Mathematics, Tafila Technical University
Tafila, 66110, Jordan.

Yousef Al-Jarrah
Department of Mathematics, Tafila Technical University
Tafila, 66110, Jordan.

Abstract

In this paper, the stepwise regression procedure is used to build a regression model for describing and identifying the factors that influence the propensity to leave the service provided by cellular phone companies. With the use of a dataset found at SPSS (V.15), the regression model and its estimation for the parameters are proposed. We find that some interactions between the predictor variables are significant, and they can be considered within the model.

1. INTRODUCTION:

One goal of multiple linear regression is to describe and estimate the relationship between one or several independent variables or predictors and a dependent variable or criterion variable. Furthermore, it can be used to forecast the dependent variable changes and values. The multiple linear model with p independent variables and n data points has the form

$$y = X\beta + \varepsilon,$$

where y is a vector $n \times 1$, X is a matrix $n \times (p+1)$, β is a vector $p \times 1$, and ε is a vector $n \times 1$, and $\varepsilon_i \sim N(0, \sigma^2)$. Notably, the multiple linear regression model in above can be extended by adding interaction terms and transformed criterion variables. Adding interactions and transformed criterion variables is a useful way of understanding the relationships among the variables in the model.

A popular technique to estimate the previous parameters vector β is the least squares method (LSM). The LSM estimates for β are

$$\hat{\beta} = (X^t X)^{-1} X^t y \, .$$

In many cases, certain predictors are not significant or most effective predictors are unknown. A stepwise procedure can be used to select and identify a useful subset of the important predictors, as well as the appropriate model. Stepwise regression is a procedure to build a model in successive steps, and predictors can be added or deleted at each step. The selection criteria are common for linear regression. Efroymson (1960) and Beale (1970) discussed many stepwise methods. In this paper, the *F*-test and a test of significance of each variable are used on the variable that is added or deleted from the model during each step.

This paper uses stepwise regression to identify the major factors that influence the propensity to leave the service provided by a cellular phone company to allow us to know the key factors that are responsible to reduce churn. Furthermore, we want to be able to predict which customers would be looking to change providers. We will use a dataset to obtain the best regression model to conduct forecasting by using a dataset available in the cellular company. In this paper, we intend to explore the following questions: First, what are the factors that influence the propensity to leave? Second, what is the best model to predict the factors that customers will be looking for when selecting a provider?

2. DATA PROCESSING FOR DETERMINING OF THE PROPENSITY TO LEAVE THE SERVICE PROVIDED BY CELLULAR PHONE COMPANIES:

The dataset is found via SPSS (V.15) and focuses on a cellular phone company. The dataset consists of one dependent variable, five independent variables, and 250 observations. The independent and dependent variables are listed below:

$y =$ (Score) Churn propensity scores that range from 0 to 100 are applied to customer’s accounts; accounts with a score of 50 or above may be looking to change providers.

$X_1 =$ (Minutes) Average monthly minutes the customer used.

$X_2 =$ (Bill) Average monthly customer bill (US dollars).

$X_3 =$ (Business) Percentage of calls that the customer used for business (ranging from 0 to 100).

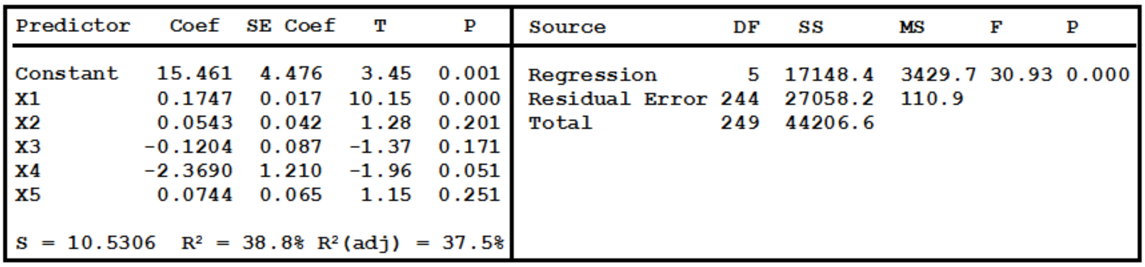
$X_4 =$ (Longserv) Years using the service.

$X_5 =$ (Income) The customer's household income in 1998 (US thousand dollars).

3. REGRESSION MODEL BUILDING:

The first attempt model is obtained by fitting a multiple linear regression of Y on all predictors, X_1, X_2, X_3, X_4, X_5. Table 1 shows the results of the regression analysis of the first attempt model. The results in Table 1 indicate that the p-value for the F-test is extremely low, thereby indicating that the model is reasonable. By examining individual t-tests for coefficient, we see that only one predictor is significant (using $\alpha = 0.05$). We use stepwise regression to find the best subset regression model that is appropriate for these data. The best subset regression that has the higher R^2 contains all predictors except X_5. Furthermore, the corresponding R^2 of this subset is equal to 37.5%.

Table 1: Multiple regression and variance analysis (first attempt model)

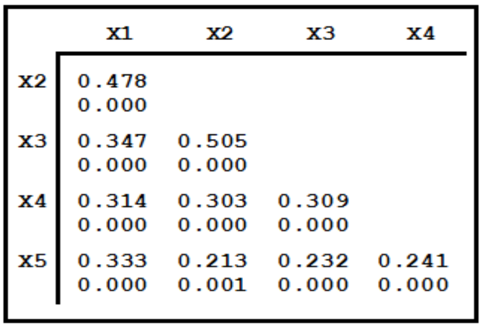


```
Predictor     Coef   SE Coef    T       P   | Source          DF    SS       MS      F     P
                                            |
Constant    15.461   4.476     3.45   0.001 | Regression       5  17148.4   3429.7 30.93 0.000
X1           0.1747  0.017    10.15   0.000 | Residual Error 244  27058.2   110.9
X2           0.0543  0.042     1.28   0.201 | Total          249  44206.6
X3          -0.1204  0.087    -1.37   0.171 |
X4          -2.3690  1.210    -1.96   0.051 |
X5           0.0744  0.065     1.15   0.251 |
                                            |
S = 10.5306   R² = 38.8% R²(adj)  = 37.5%   |
```

4. MULTICOLLINEARITY:

Multicollinearity is a statistical phenomenon in which a strong correlation occurs between some predictor variables. Multicollinearity can increase the variances of the parameter estimates, thereby possibly causing insignificant predictors even though the overall model is significant (Joshi, 2012). Multicollinearity can be detected by examining the correlation matrix (Belsley et al.,1980). Given that we have one significant predictor X_1, we will calculate the correlation matrix to determine if multicollinearity exists. The results in Table 2 indicate that no high correlations exist between the predictors. This finding indicates that multicollinearity does not exist.

Table 2: Correlation matrix between predictor variables



	x1	x2	x3	x4
x2	0.478 0.000			
x3	0.347 0.000	0.505 0.000		
x4	0.314 0.000	0.303 0.000	0.309 0.000	
x5	0.333 0.000	0.213 0.001	0.232 0.000	0.241 0.000

5. INTERACTION AND TRANSFORMATION OF VARIABLES:

Given that we have insignificant predictors and no multicollinearity is detected among predictor variables, transformation of the dependent variable Y may be useful. Nordberg (1982) used data transformation to overcome the variable selection problem for a general linear model. To determine which transformation is needed, Box-Cox (1964) transformations are used to find potentially nonlinear transformations of a dependent variable. The figures shows the Box-Cox plot of Y. Figure 1 shows that the Box-Cox plot suggests lambda 1, which indicates no transformation for the dependent variable. We will attempt to transform the independent variable, but before doing so, we will observe the relationship between the dependent variable Y and the predictors. Useful observations may be found.

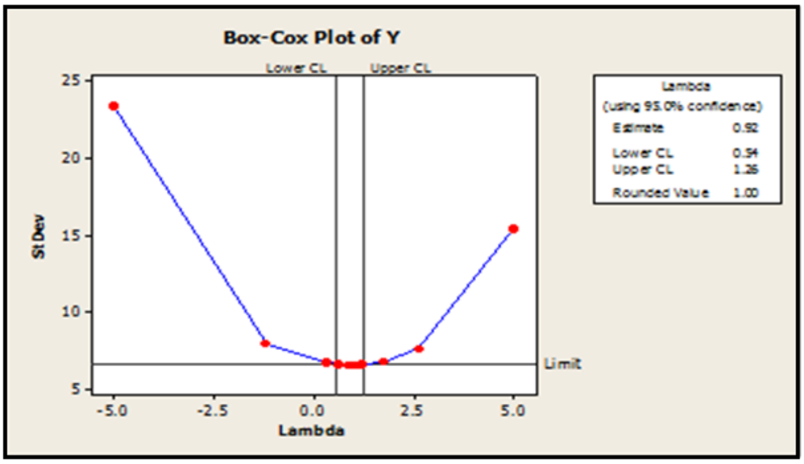


Figure 1. Box–Cox plot of independent variable Y.

Figure 2 contains scatter plots between Y and each predictor; these plots can be used to examine the nature of the relationship between the variables so that we can analyze the probable transformations on predictors before building an appropriate model.

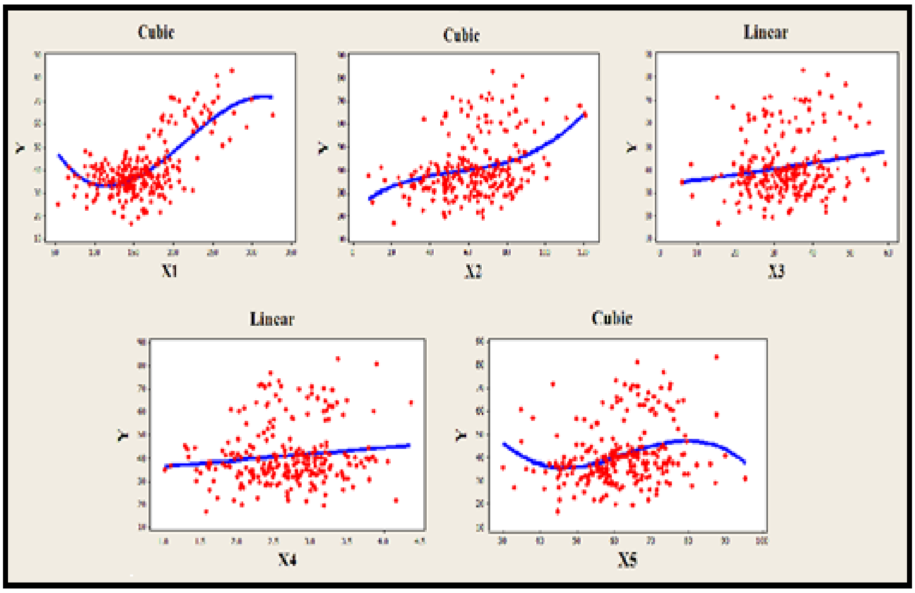


Figure 2. Pairwise scatter plots between the predictor variables and the dependent *Y*.

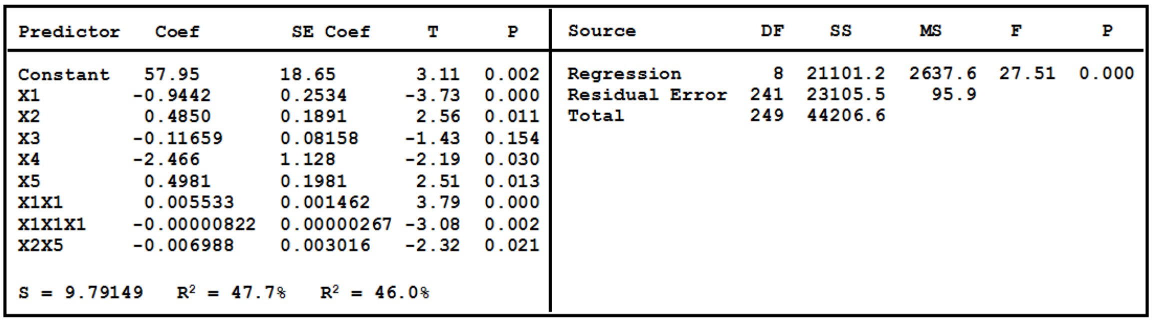
The scatter plots indicate that using cubic transformations for the predictors and adding an interaction between the variables will be most effective. Cubic transformation and some interaction among the predictors are needed. Therefore, we will include $X_1^2, X_1^3, X_2^2, X_2^3, X_3^2, X_3^3$, X_4^2, X_4^3, X_5^2, and X_5^3, as well as quadratic and cubic terms in the initial model. The quadratic term is added because we cannot consider the cubic transformation without lower-degree terms. Furthermore, interaction terms can be constructed by using a multiplication scheme that multiplies the two explanatory variables. Therefore, we added the following new terms to the initial model:

$$X_1X_2,\ X_1X_3,\ X_1X_4,\ X_1X_5,\ X_2X_3,\ X_2X_4,\ X_2X_5,\ X_3X_4,\ X_3X_5,\ X_4X_5.$$

The new regression model will then be obtained by using stepwise regression to find the best subset regression model. However, the final model should contain only the significant interaction terms in the initial model. The results obtained by using stepwise regression to select the best model from the predictors and their quadratic, cubic, and interaction terms are given in Table 3. The subset of independent variables that best predicts the dependent variable contains $X_1, X_2, X_3, X_4, X_5, X_1^2, X_1^3$, and X_2X_5

Furthermore, seven significant predictor variables have *p*-values that are less than 0.05, thereby indicating that they are significant. The value of R^2 is 47.7 % (increased from the first attempt model) and the *F*-test statistic is 27.51 with *p*-value reported as 0.000. In addition, the variable X_3 remains insignificant. A probable reason for this result is the weak relationship indicated by the scatter plot in Figure 2. Dropping X_3 from the model may be necessary.

Table 3: Multiple regression analysis results that use stepwise regression from the predictors and their quadratic, cubic, and interaction terms



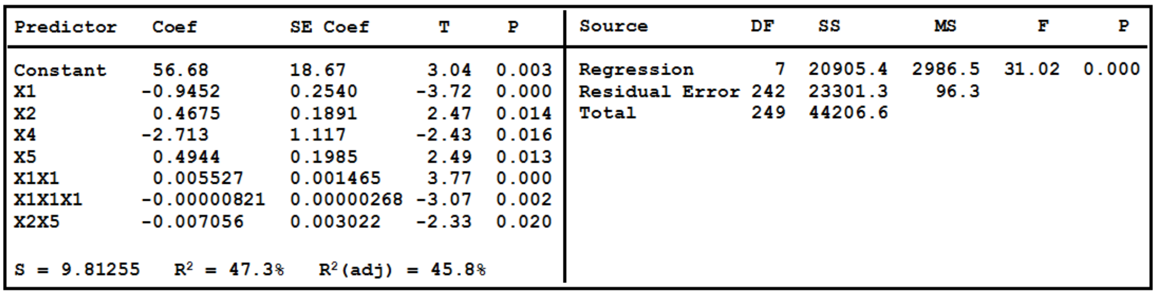
```
Predictor    Coef          SE Coef       T      P      Source            DF    SS       MS      F      P

Constant     57.95         18.65        3.11  0.002    Regression         8  21101.2  2637.6  27.51  0.000
X1          -0.9442        0.2534      -3.73  0.000    Residual Error   241  23105.5    95.9
X2           0.4850        0.1891       2.56  0.011    Total            249  44206.6
X3          -0.11659       0.08158     -1.43  0.154
X4          -2.466         1.128       -2.19  0.030
X5           0.4981        0.1981       2.51  0.013
X1X1         0.005533      0.001462     3.79  0.000
X1X1X1      -0.00000822    0.00000267  -3.08  0.002
X2X5        -0.006988      0.003016    -2.32  0.021

S = 9.79149   R² = 47.7%   R² = 46.0%
```

After X_3 is dropped from the model, the fitting results of multiple linear regression of Y on the predictors, $X_1, X_2, X_4, X_5, X_1^2, X_1^3$, and X_2X_5 are provided in Table 4. The results in Table 4 indicate that all the predictors are significant; the value of R^2 is 47.3%, and the F-test statistic is 31.02 with p-value reported as 0.000.

Table 4: Regression analysis: Y versus $X_1, X_2, X_4, X_5, X_1^2, X_1^3$, and X_2X_5



```
Predictor    Coef          SE Coef       T      P      Source            DF    SS       MS      F      P

Constant     56.68         18.67        3.04  0.003    Regression         7  20905.4  2986.5  31.02  0.000
X1          -0.9452        0.2540      -3.72  0.000    Residual Error   242  23301.3    96.3
X2           0.4675        0.1891       2.47  0.014    Total            249  44206.6
X4          -2.713         1.117       -2.43  0.016
X5           0.4944        0.1985       2.49  0.013
X1X1         0.005527      0.001465     3.77  0.000
X1X1X1      -0.00000821    0.00000268  -3.07  0.002
X2X5        -0.007056      0.003022    -2.33  0.020

S = 9.81255   R² = 47.3%   R²(adj) = 45.8%
```

6. DIAGNOSING THE OUTLIERS:

Figure 3 shows the box plot of residual for the model in Table 4. The box plot indicates five outlier values; the outliers are 11, 18, 28, 40, and 69. The previous outliers may produce large residuals but do not influence the regression coefficients. In a critical condition, the residuals rather than existing outliers should be normally distributed.

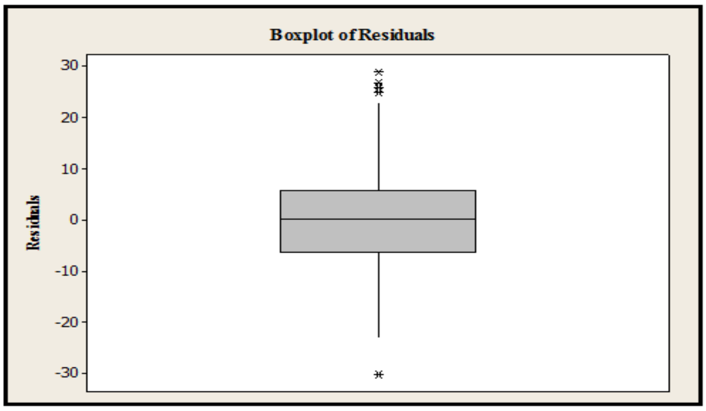


Figure 3. Box plot of residual for fitting: Y on $X_1, X_2, X_4, X_5, X_1^2, X_1^3$, and $X_2 X_5$.

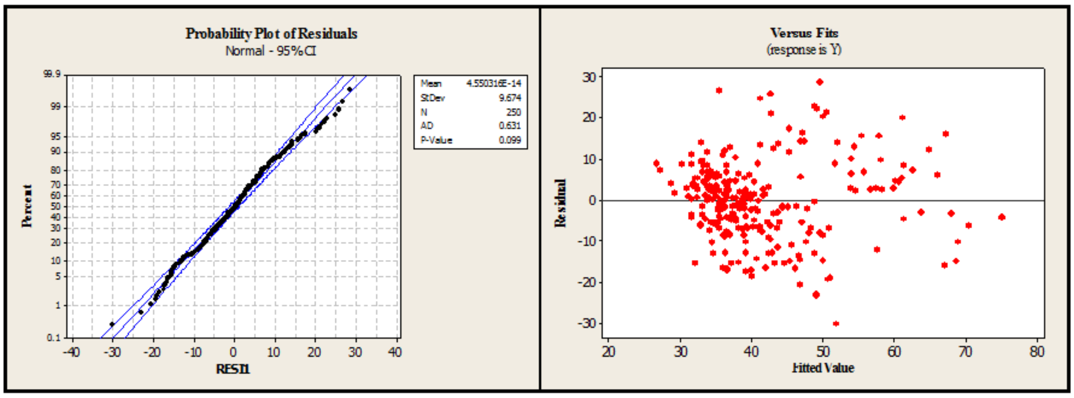


Figure 4. The residuals plot and the normal probability plot of residuals.

The residuals plot in Figure 4 does not show any discernible pattern. The normal probability plot in Figure 4 shows that the points follow a linear pattern; p-value = 0.099 > 0.05. We conclude that the residuals are normally distributed, the residuals plot versus the fitted value does not indicate any problems in the model, and we do not need to be concerned about a cumulative fitted value between 30 – 40 because this result indicates that most customers are not thinking about leaving. Thus, the model is accepted.

7. Summary and Conclusions:

The final model is as follows:

Score = 56.7 – 0.945 Minutes + 0.468 Bill – 2.71 Longserv + 0.494 income + 0.00553 Minutes^2 – 0.000008 Minutes^3 – 0.00706 Bill * Income.

No significant relationship is observed between the propensity to leave the cellular service and using the service for business or personal use. Furthermore, a significant relationship is found between propensity to leave the cellular service and with the following variables: minutes, bill, and income. Years using the service is a significant predictor of score propensity to leave because the cellphone number will be a reference in many places for the person and is difficult to change. (Increased years will reduce the score). We also notice a significant interaction between the income and the bill. The above model can help reduce churn by notifying the cellphone company about the significant factors that influence churn.

REFERENCES

[1] Beale, E. M. L. (1970). Note on procedures for variable selection in multiple regression, Technometrics, 12: 909-914.

[2] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). Regression Diagnostics - Identifying Influential Data and Sources of Collinearity, John Wiley and Sons, New York.

[3] Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations, Journal of the Royal Statistics Society. Series B, 26: 211-234.

[4] Efroymson, M. A. (1960). Multiple regression analysis, Mathematical Methods for Digital Computers. Eds. A. Ralston & H.S.Wilf, New York: John Wiley & Sons, Inc.

[5] Joshi, H., (2012). Multicollinearity Diagnostics in. Statistical Modeling and Remedies to deal with it using. SAS. Session SP07 - PhUSE 2012.

[6] Nordberg, L. (1982). On variable selection in generalized linear and related regression models, Communications in Statistics Theory and Method, 11: 2427-2449.

[7] SPSS Inc. Released 2008. SPSS Statistics for Windows, Version 17.0. Chicago: SPSS Inc.