Regression Analysis Projects

Jaechan Park

figo721@hotmail.com

1 Introduction

The objective of this study is analyzing the dataset which is a random sample of 113 hospitals across the US. by using statistical models for searching an appropriate model to interpret the possible association between the dependent and the independent variables. Also, a formulated regression model also needs to test with randomly selected 20 observations in a validation set. Lastly, a regularized, ridge, method uses to analyze the dataset and then compare to the outputs of formulated ordinary least estimate method in the previous section.

2 Linear Regression Model; Ordinary Least Estimate

In this data set, there are 8 predictor variables which is as shown in table 1

Table 1: The list of predictors

1. lstay	2. age	3. cultratio	4. xrayratio	5. nbeds	6. census	7. nnurse	8. facil

In a multivariate regression model, identifying significant regressor(s) is a crucial and the first step. A major concern in a multivariate regression analysis is multicollinearity. Variance Inflation Factor, *VIF*, is a method to find out its multicollinearity among predictor variables. From the result in table 2, some variables are highly correlated, for instance, predictor 5 and 6 variable (nbeds and census) have a significantly high value of its Variance Inflation Factor.

Table 2: Variance Inflation Factors *VIF*

VIF 1	*VIF 2*	*VIF 3*	*VIF 4*	*VIF 5*	*VIF 6*	*VIF 7*	*VIF 8*
2.269703	1.166625	1.494686	1.457862	37.686621	39.757258	6.827657	2.925686

2.1 AIC-based Stepwise Regression

For a proper regression model with significant regressors, stepwise method is a common procedure in selecting variable method. Three different methods; backward elimination, forward selection and a combination of backward and forward of stepwise regression have been conducted in this study. Based on the outputs of the stepwise regression methods, 4 predictors, "lstay," "cultratio," "xrayratio," and "facil" are significant variables of the regression model.

2.2 F-test-based Manual Method

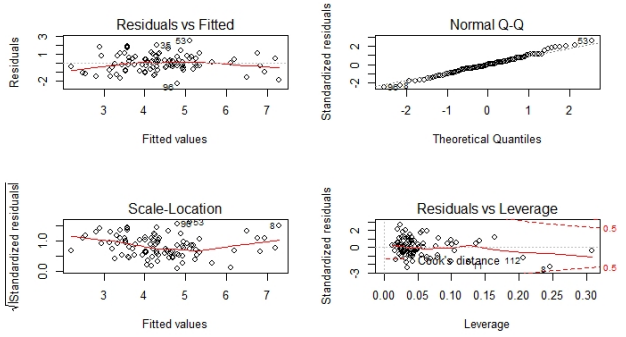
Apart AIC-based stepwise regression method, two F-test based manual methods; backward elimination and forward selection use to find significant predict variable(s). According to the outputs from F-test based manual methods, the selection of significant regressors is same as the AIC-based stepwise regression model as we have found in the previous section.

Hence, we develop a linear regression model with identified significant predictor variables and the equation of the linear regression is shown in the equation (1) as below:

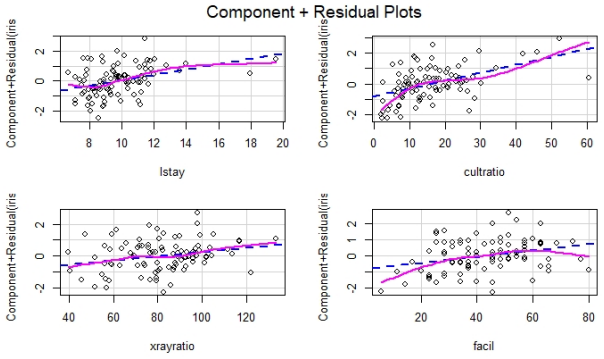
$$irisk = -0.05648 + 0.18220 lstay + 0.05065 cultratio + 0.01293 xrayratio + 0.01888 facil \quad (1)$$

2.2.1 Model assumptions

Model assumptions have to be satisfied with the validation of a linear regression model. The satisfied conditions are homoscedasticity, independence of error terms, normality, and controlled outliers. In figure 1, it is clearly seen the linear model is violated to meet the condition for normality based on the two upper plots. In addition, we could conclude that the linear regression model of the equation (1) has evidence of having heteroscedasticity. The reason is based on the Scale-Location (bottom-left) plot represents that the residuals are not randomly spread out. Furthermore, component residual plots in figure 2 reveal that two predictor variables, cultratio and facil, do not satisfy the linear relationship with the response variable which is also violated appropriateness of a regression model assumption.



(a) Figure 1: Model Assumptions Plots



(b) Figure 2: Component Residual Plots

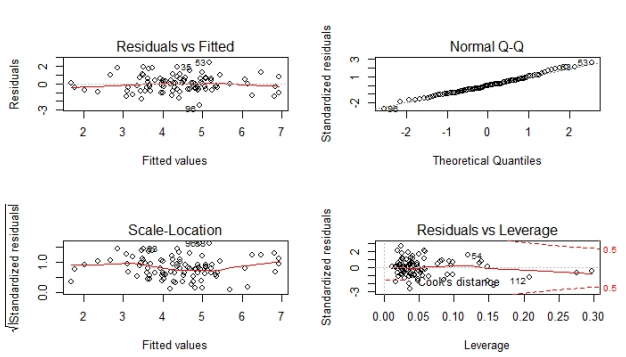
2.3 Transformation

Variable transformation considers as a handy method to address any violation of the model assumptions. Among several transformation options, log and square root transformation commonly use for statistical analysis for the model assumptions. After several trials, a proper transformed linear regression model has been obtained, and the equation of transformation model is in the equation (2).

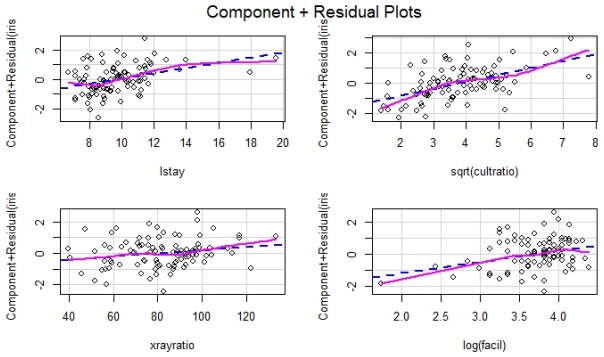
$$irisk = -2.581655 + 0.181987 lstay + 0.460877\sqrt{cultratio} + 0.715803 \ln(facil) \tag{2}$$

2.3.1 Model Assumptions

Diagnostic plots (figure 3) and Component Residual plots (figure 4) provide evidence to conclude that the transformation model (equation 2) is satisfied the model assumptions for validation of the regression model.



(c) Figure 3: Model Assumptions Plots



(d) Figure 4: Component Residual Plots

2.4 Model Comparison and Validation

Value of predicted sum of squares (PRESS) of a regression model is to estimate the mean squared error of predictors and PRESS widely uses for comparison of well-fitness or measurement of good candidate model when we have more than one linear regression model. A linear regression model with smaller PRESS value is considered a better model than others. In this study, transformation linear model, equation (2), has a smaller predicted sum of squares values than equation (1). Therefore, we can conclude that equation (2) linear regression model is a more suitable/appropriate model than the model of equation (1), according to the values of the PRESS in table 3.

The result of the linear regression model, equation 2, by using 20 observations in the validation set is in table 4. According to the result of the regression model with the validation set, none of the predictor variables are significant. A possible reason for the outcome is that the sample size of the validation set is not sufficiently large for the test.

Table 3: PRESS

Model	PRESS
Equation (1)	94.39128
Equation (2)	84.85802

Table 4: Regression with validation set

	Estimate Std.	t-value	p-value
Intercept	-2.9256	-1.336	0.2002
lstay	0.2191	1.128	0.2760
sqrt(cultratio)	0.2484	1.290	0.2154
ln(facil)	1.1250	2.107	0.0513

3 Ridge Regression

As Ordinary Least Squares (OLS) estimation often has improper statistical properties and challenging for interpretation accurately, regularized regression method needs to overcome shortcomings of OLS estimation. By using the ridge regression, the best lambda values from the cross-validation result are 0.389443, and the lambda value with $|oneSE|$ is 2.280976. Based on the results of λ, it can be computed the coefficient estimates for each of these λ values, which are given in table 5. As well, table 6 represents the coefficient estimate of the ridge regression model with selected predictor variables were identified in the previous section. Since the ridge regression model does not perform variable selection itself when λ is sufficiently large enough, the ridge regression model with selected predictor variables may provide a relatively efficient to interpret and predict model than a ridge model with the full predictor variables where in table 5.

Table 5: Coefficient estimates

	$Ridge_{\lambda=0.389443}$	$Ridge_{\lambda=2.280976}$
Intercept	$-1.512579e^{-01}$	1.5849719035
lstay	$1.422485e^{-01}$	0.0882050716
age	$1.447646e^{-02}$	0.0057190687
cultratio	$4.128600e^{-02}$	0.0223083986
xrayratio	$1.336024e^{-02}$	0.0093272810
nbeds	$5.418344e^{-05}$	0.0002933593
census	$1.537956e^{-04}$	0.0004177194
nnurse	$9.814465e^{-04}$	0.0006154636
facil	$9.938008e^{-03}$	0.0062372868

Table 6: Selected variables

	$Ridge_{\lambda=0.389443}$	$Ridge_{\lambda=2.280976}$
Intercept	0.41473541	1.851747061
lstay	0.16001930	0.101077811
cultratio	0.04100923	0.022843475
xrayratio	0.01298400	0.009299995
facil	0.01642009	0.010074562