Linear Regression Model in R

Final Report

**Towards an analysis of possible factors related to admission rate in universities & colleges in US**

Shu Wang(1004712965)

**Introduction**

Admission rate is the rate at which applicants are accepted. It is calculated by dividing the number of accepted students by total number of applicants. Admission rate vary from university to university and this may be the result of various reasons, including type of university, facilities and equipment of campus, tuition fee and even some other conditions important to applicants themselves. We need to filter factors that affect admission rate first, then describe and understand the relationship between admission rate and these factors.

The aim of this project is to come up with a linear regression model which describes the relationship between admission rate and predictors most precisely in our population of 1,508 observations of universities and colleges in US. The goal of this project is to help various stakeholders to understand the relationship by using the model. Also, we should be able to make predictions of admission rate if certain factors change or we have a new observation by using our model.

**Methods**

**1. Dataset**

**1.1 Observations**

We have 1,508 observations of universities and colleges in United States. The dataset is derived from a larger collection of measures on schools in the United States (<https://collegescorecard.ed.gov/data/>).

**1.2 Variables**

The dataset has 30 variables, with one response variable (ADM\_RATE) and 29 possible predictor variables. The variables are divided into 3 categories: school identifiers, school characteristics and applicant characteristics.

**2. Statistical Analyses**

**2.1 Training set and testing set**

The dataset is divided equally (50-50 ratio) into a training set and testing set. We perform the regression model on our training set and use testing set to test fitness of the model.

**2.2 Beforehand selection**

As we see from the dataset, we have 31 columns (possible predictors). X is a numeric number with no actual meaning described in the dataset, “UNITED” and “INSTNM” are identification of each university/college, so we remove these 3 variables from our predictor variable list.

**2.3 Full model**

We first perform a full model to test overall significance of the predictor variables. By checking “NA” value in coefficients and p-value for F-statistics, we detect singularity problems and test our null hypothesis of all coefficients.

**2.4 Stepwise selection method using AIC & BIC**

We use stepwise selection method to select our model. Since we have 28 possible predictors, it is not realistic to perform “all possible subset” method. In order to obtain more possible candidate models, we use forward & backward & both direction selection method based on both AIC and BIC score. We then calculate the value, sum of squared regression value, AIC, AIC (corrected), BIC and number of predictors in each model to make comparisons.

**2.5 Model assumptions**

To test whether our model is valid, we check model assumptions. We plot actual response vs. fitted value to see whether condition 1 is satisfied, plot pairwise scatterplot of predictor variables to see whether condition 2 is satisfied. We use residuals vs. fitted values and residual vs. each predictor variable to see whether there is fanning pattern/curvature in the plot. We use QQplot to check whether normality of errors in held.

**2.6 BoxCox method and PowerTransformation**

We perform PowerTransform function (if applicable) on our model to correct model violations.

**2.7 Leverage points, Outliers and Influential points**

To deal with problematic observations, we find leverage points based on hat-values and outliers based on the value of standardized residuals. We find influential points based on Cook’s distance, DFFITS value and DFBETA value. For extreme observations in the dataset, we consider removing some of them for better fitness of the dataset.

**2.8 Multicollinearity**

After we finish editing our model, we check multicollinearity of the model. By calculating VIF value of all predictors in the model, we can conclude that there is/is not multicollinearity of predictors.

**2.9 Fitting the regression model into testing set**

Our final step is to fit the regression model into testing set to see fitness of the data. By summarize the fitted model, we can check whether the model is overfitting or underfitting.

**3. Software**

All analyses were done using the software RStudio Cloud (R 4.0.0).

**Results:**

**1. Full model**

The full model (with 28 predictor variables) indicates that 1 coefficient is not defined because of singularities. From the dataset, we see that “REGION” has a 1-1 correspondence to the “STABBR” variable. Since variable “REGION” doesn’t add extra information to the regression model, we remove it from the predictor list.

The full model has a p-value of , so we reject the null hypothesis that all coefficients are zero.

**2. Stepwise selection method**

From table 1, we see that we have 4 possible candidate models. They have similar value of around 0.22-0.24. Their 4-criteria value are similar with each other. In this case, we prefer the model with fewer predictor variables for simplicity and better interpretation. Model 3 has a higher value compared to model 4, which means the residual sum of squares of model 3 is smaller than model 4(since they both have 7 predictors.) We choose model 3 as our final candidate model.

**3. Model assumptions**

As we plot actual response vs. fitted values of model3(see Figure 1), most of the data fits in the centre of the plot. The regression line we made by the plot is linear, so we can conclude that condition 1 is held.

For scatterplots of 7 predictors, in this case, the predictors have no linear relationship, but also no non-linear relationship, so condition 2 is satisfied.

For residual plot vs. fitted values and all predictors, there is no clear pattern of residuals vs. fitted value in the plot (see Appendix 1). For most of the predictors, the residuals are randomly located in the graph, except for residuals vs. POVERTY\_RATE (see Figure 2). There is a separation of the residuals in the plot, which shows group of residuals might be related to each other, this violates the law of independence errors.

From the QQplot, we can conclude that normality assumption is held in this model. (see Figure 3)

**4. BoxCox method and PowerTransformation**

We perform a PowerTransformation on our model. We conclude that the power of NUMBRANCH is significantly different from 1, so we add this power to our model. For all other power transformation results, the estimated result is not far away from 0 so we leave other 6 predictor variables and our response variable with their original power.

**5. Leverage points, outliers and influential points**

It is clear that our for all models are around 0.23, which means that 23% of the variation in our response variable can be explained by the variation of predictor variables. We use leverage points, outliers and influential points to find certain observations which are highly influential to the regression model. We have 83 leverage points, 1 outlier, 0 influential point (based on Cook’s distance), 53 influential points (based on DFFITS value) and a bunch of influential points which has an impact on estimated coefficients. Since the number of influential points is quite large, we select those who has a significance impact on more than 4 of 7 coefficients and on their own predicted value.

It turns out that 14 observations have high DFFITS value and more than 4 extreme values in DFBETAS. (e.g. Observation 746 and 207 are influential to their own predicted value and all 7 estimated coefficients). Since they are considered highly influential to the regression model, we remove these observations for a better fitness of the model.

After we remove observations, our model has a value of 0.29, which shows that these observations are problematic to our regression model.

**6. Multicollinearity**

Results (see Appendix 2) shows that there’s no significant evidence of multicollinearity in this model.

**7. Testing set**

By fitting the model into testing set, we can see that the testing set has a value of 0.2119.(See Appendix3) This shows a sign of overfitting, but this is because we remove extreme observations in the training set(which is selected randomly). If we compare original value for 2 sets, the results are similar, which indicates that overall the model is a good model. Final summary for selected model on training set is in Figure4.

**Discussion**

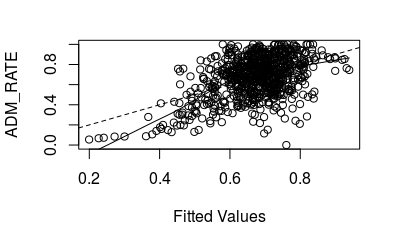
The final model is below:

This model shows that ADM\_RATE is most explained by the variation of these 7 predictor variables. For individual coefficient, all of them have a p-value of less than 0.05 in the model, which indicates that the coefficients are all different from 0. NUMBRANCH is not directly related to the linear model. Around 29% of the variation in Y can be explained by variation in X(i), which shows a moderate linear relationship between the predictor variables and response variables. This model is the best possible model because it has a relatively simple expression in 4 candidate models which have lowest AIC/BIC score. value for 4 models are similar, so the one with fewer predictor variables is preferred since it’s easier to understand and interpret this model. Model assumptions are satisfied and no multicollinearity problem is found. This model is also good for making predictions since we have done PowerTransform on the original model. Limitations of this model is that the residuals seem to stay near left of the whole plot, which shows that we have most of our data on the left-hand side of the whole range. Also, there’s too many influential points in the model, suggesting the predicted value may have a wide range. In all, this model is the best possible model which is good for both interpretation and prediction.

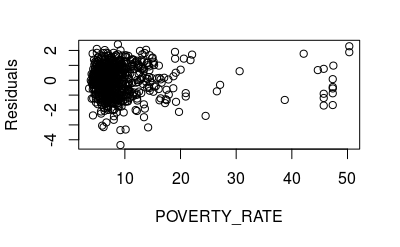
**Tables and Figures**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model name | SS res |  | AIC | AIC(c) | BIC | # of predictors |
| Model 1 | 22.63 | 0.2399 | -2617.585 | -2616.926 | -2544.194 | 13 |
| Model 2 | 22.68 | 0.2392 | -2617.825 | -2617.258 | -2549.069 | 12 |
| Model 3 | 23.19 | 0.2275 | -2611.268 | -2611.027 | -2565.640 | 7 |
| Model 4 | 23.28 | 0.2245 | -2608.269 | -2608.028 | -2562.640 | 7 |

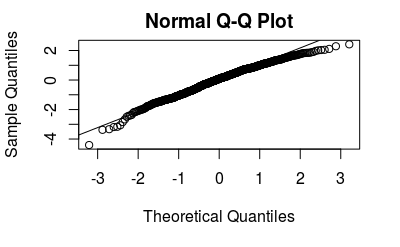
**Table1: R^2-adj,SS res, AIC/AIC(c) /BIC and number of predictors for 4 models selected by stepwise selection method**.



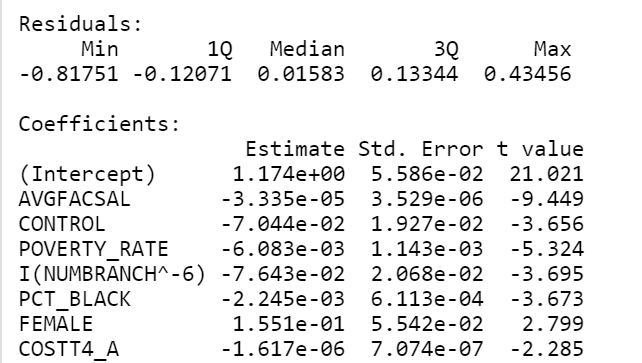
**Figure1:Actual response vs. Fitted values for model 3(used for condition 1 checking)**

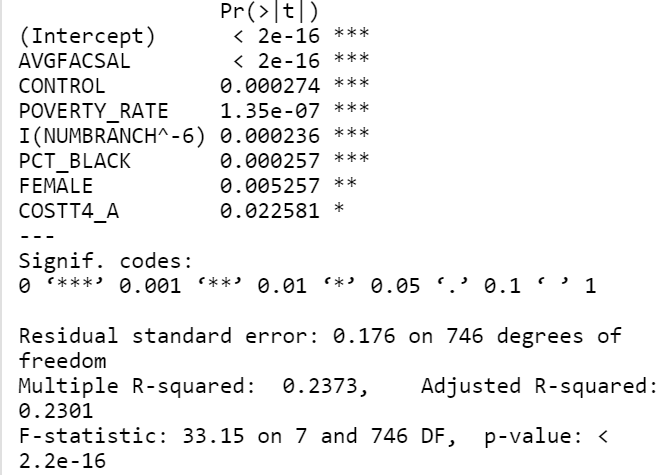


**Figure2: Residuals vs. predictor variable(POVERTY\_RATE) for model 3**



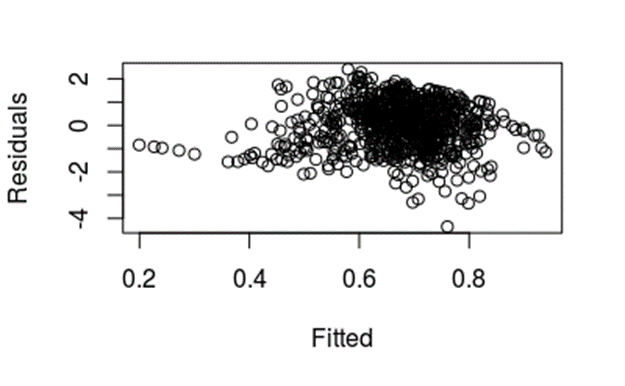
**Figure3: Normal QQplot of standardized residuals for model 3**



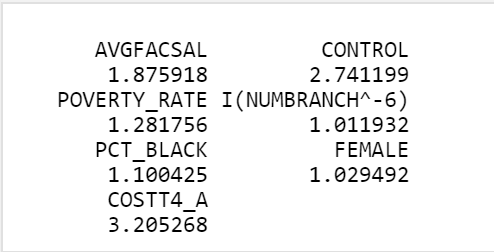


**Figure4: summary of corrected model3 on training set**

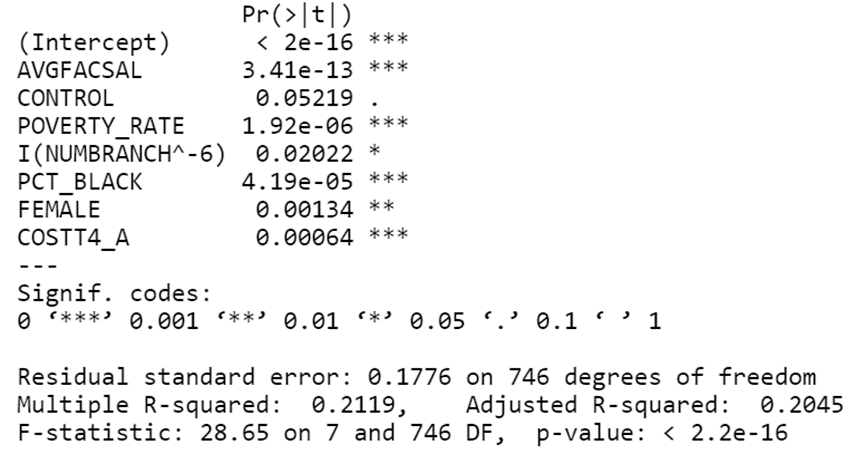
**APPENDIX**



**Appendix 1: Residuals vs. Fitted value for model3**



**Appendix2: Multicollinearity test result for corrected model3**

****

**Appendix3: part of summary of corrected model3 on testing set**