Predicting Admissions

Chinmay Sheth

December 20, 2019

Introduction

In this exercise, I'm analyzing a university dataset which contains information about student admissions. I am hoping to build a generalized linear model in order to reliably predict a student's admission to the university using SAS.

Exploratory Data Analysis

Dataset

The dataset that I am using was obtained from Kaggle and it contains data on university admission data. The dataset contains the following columns:

1. Serial No.
2. GRE Score
3. TOEFL Score
4. University Rating
5. Statement of Purpose Rating
6. Letter of Recommendation Rating
7. Cumulative Grade Point Average
8. Research
9. Chance of Admit

There are nine variables with 400 records, where there are no missing values.

Data Exploration

There is one categorical variable, Research, which is already encoded as a 0 for no participation in research and 1 for participation in research. The predicted variable of interest is the ninth variable, Chance of Admittance, I am hoping to use a subset of the first eight variables in order to reliably fit a model which is able to explain the variability in the ninth variable well. Furthermore, it doesn't make sense to include the first variable, Serial No., in the basic model because it is random for each student and doesn't effectively convey any meaning.

Analysis

Diagnostics

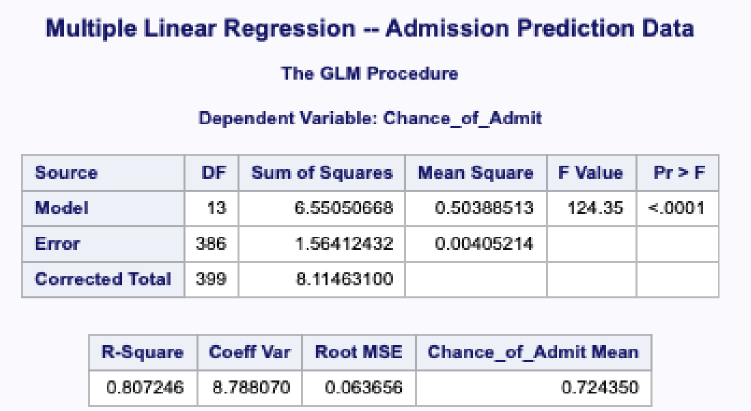
Through the PROC GLM function, SAS is easily able to build a basic linear model that considers all the variables with interactions, which can be further investigated to determine to see if the linear regression assumption are met. PROC GLM is also able to differentiate between categorical and continuous variables by indicating so in the CLASS definition of the function. As such, ”Research” is specified as a categorical variable in the CLASS method. The basic model with interaction terms is as follows:

$$\mathbf{y = B_0 + B_1x_1 + B_2x_2 + B_3x_3 + B_4x_4 + B_5x_5 + B_6x_6 + B_7x_7 + B_17x_1x_7 + B_27x_2x_7 + B_37x_3x_7 + B_47x_4x_7 + B_57x_5x_7 + B_67x_6x_7}$$

1. $\mathbf{y}$: Chance of Admittance
2. $\mathbf{x_1}$: GRE Score
3. $\mathbf{x_2}$: TOEFL Score
4. $\mathbf{x_3}$: University Rating
5. $\mathbf{x_4}$: Statement of Purpose Rating
6. $\mathbf{x_5}$: Letter of Recommendation Rating
7. $\mathbf{x_6}$: Cumulative Grade Point Average
8. $\mathbf{x_7}$: 0 if research completed otherwise 1

After performing the PROC GLM Procedure in SAS, the resulting estimates are displayed in Figure 2. The diagnostic plots are checked to see if the regression assumptions are met which are:

1. $\mathbf{E}(\epsilon_i) = 0 \; \forall i \in 1..N$
2. $Var(\epsilon_i) = \theta \; \forall i \in 1..N$



Multiple Linear Regression -- Admission Prediction Data

The GLM Procedure

Dependent Variable: Chance_of_Admit

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	6.55050668	0.50388513	124.35	<.0001
Error	386	1.56412432	0.00405214		
Corrected Total	399	8.11463100			

R-Square	Coeff Var	Root MSE	Chance_of_Admit Mean
0.807246	8.788070	0.063656	0.724350

Figure 1: Basic model output

3. ϵ_i are normally distributed $\forall i \in 1..N$
4. ϵ_i are independent $\forall i \in 1..N$

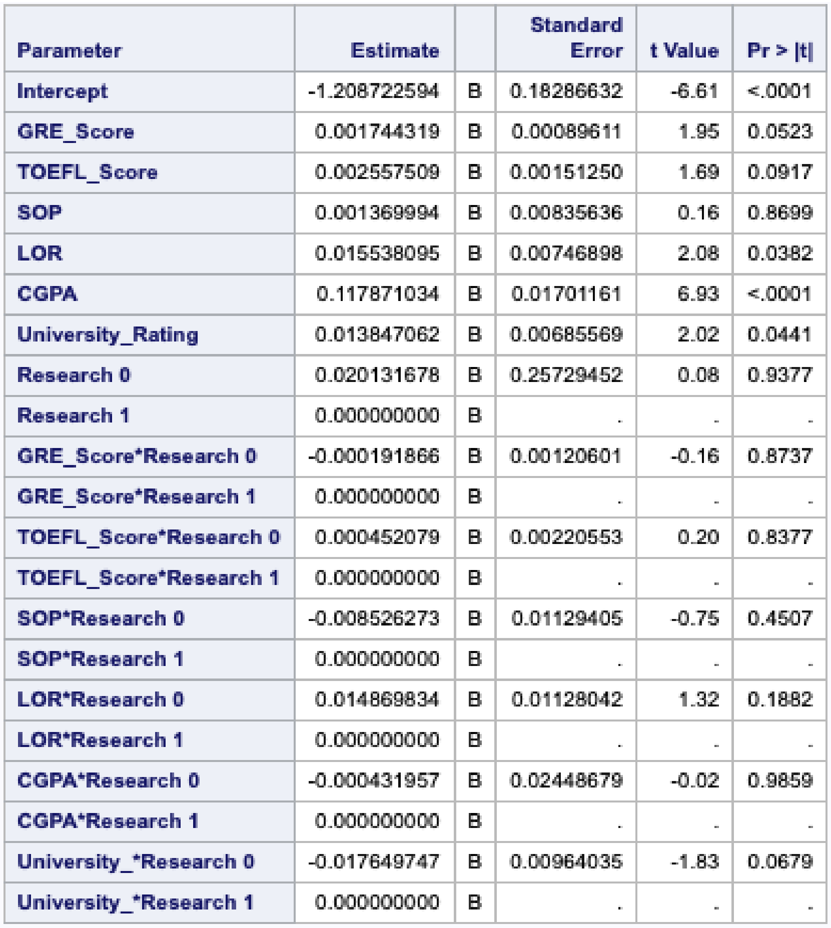
In the diagnostic plots from Figure 3 there is a departure from the homoskedasticity assumption, that is of constant variance throughout the data, because of the decreasing variance in the data in the Residual vs Predicted Value plot and RStudent vs Predicted Value plot. There also appears to be a departure from the assumption of the expected value of the residuals to equal zero as the data in the two aforementioned plots is not centred around the horizontal zero axis. Finally, the data does not appear to be normally distributed as the Residual vs Quantile Plot shows that the points are not very closely hugging the reference line.

Transformations

Using the data-driven Box-Cox technique, showed in Figure 4, we can estimate the best transformation for the basic linear model which should result in diagnostic plots which have less of a departure from the linear regression assumptions. The R-Squared value has increased from 0.807246 (Figure 1) to 0.860190 in the transformed model, indicating that the transformed model is able to explain more variability in the predicted values than the initial model was able to. Furthermore, there appears to be a significantly less departure from the linear regression assumptions as is seen in the diagnostic plots.

Model Selection

Now that the regression assumptions have been satisfied, model selection can be performed in order to determine whether all the covariates are truly necessary. According to the Principle of Parsimony, we want to be able to explain the



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1.208722594	B	0.18286632	-6.61	<.0001
GRE_Score	0.001744319	B	0.00089611	1.95	0.0523
TOEFL_Score	0.002557509	B	0.00151250	1.69	0.0917
SOP	0.001369994	B	0.00835636	0.16	0.8699
LOR	0.015538095	B	0.00746898	2.08	0.0382
CGPA	0.117871034	B	0.01701161	6.93	<.0001
University_Rating	0.013847062	B	0.00685569	2.02	0.0441
Research 0	0.020131678	B	0.25729452	0.08	0.9377
Research 1	0.000000000	B	.	.	.
GRE_Score*Research 0	-0.000191866	B	0.00120601	-0.16	0.8737
GRE_Score*Research 1	0.000000000	B	.	.	.
TOEFL_Score*Research 0	0.000452079	B	0.00220553	0.20	0.8377
TOEFL_Score*Research 1	0.000000000	B	.	.	.
SOP*Research 0	-0.008526273	B	0.01129405	-0.75	0.4507
SOP*Research 1	0.000000000	B	.	.	.
LOR*Research 0	0.014869834	B	0.01128042	1.32	0.1882
LOR*Research 1	0.000000000	B	.	.	.
CGPA*Research 0	-0.000431957	B	0.02448679	-0.02	0.9859
CGPA*Research 1	0.000000000	B	.	.	.
University_*Research 0	-0.017649747	B	0.00964035	-1.83	0.0679
University_*Research 1	0.000000000	B	.	.	.

Figure 2: Basic model estimates

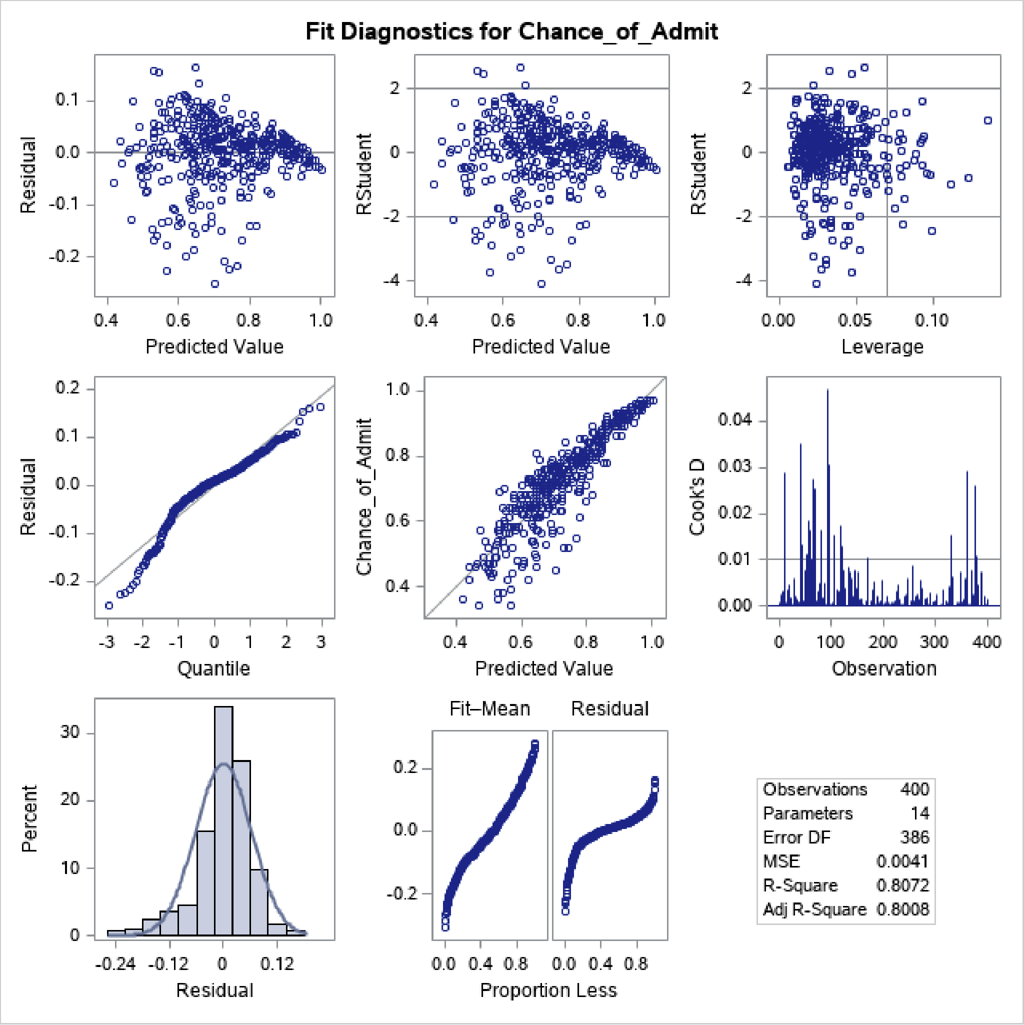


Figure 3: Model Diagnostics for Basic Model with Interactions

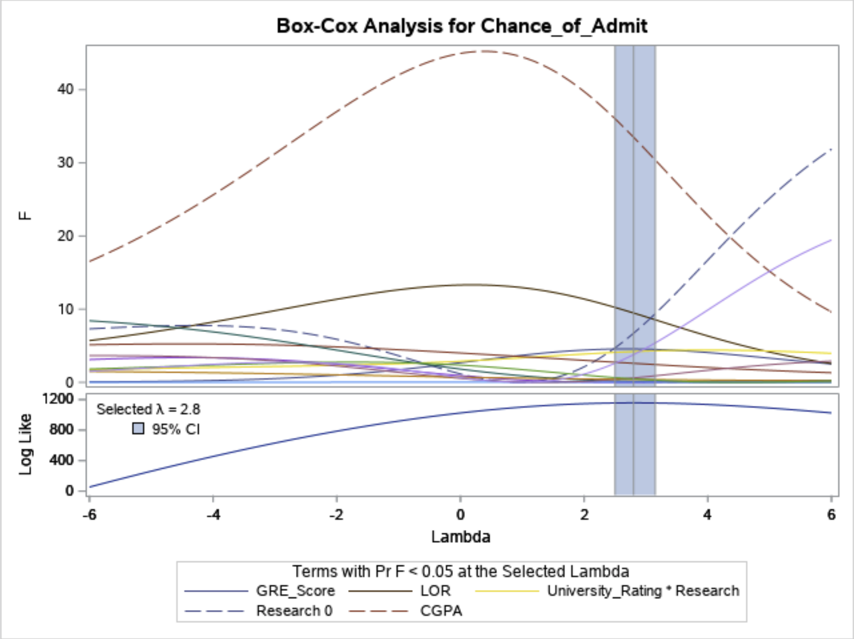
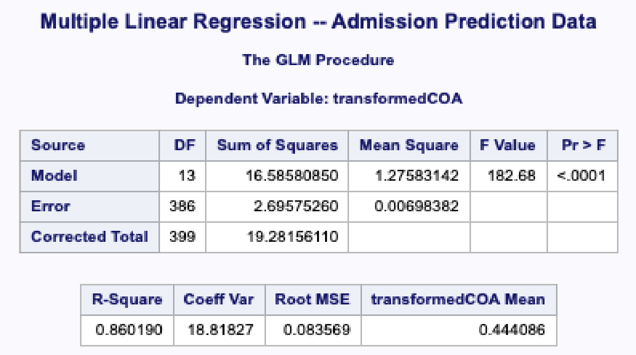


Figure 4: Box Cox Analysis for Basic Model with Interactions



Multiple Linear Regression -- Admission Prediction Data

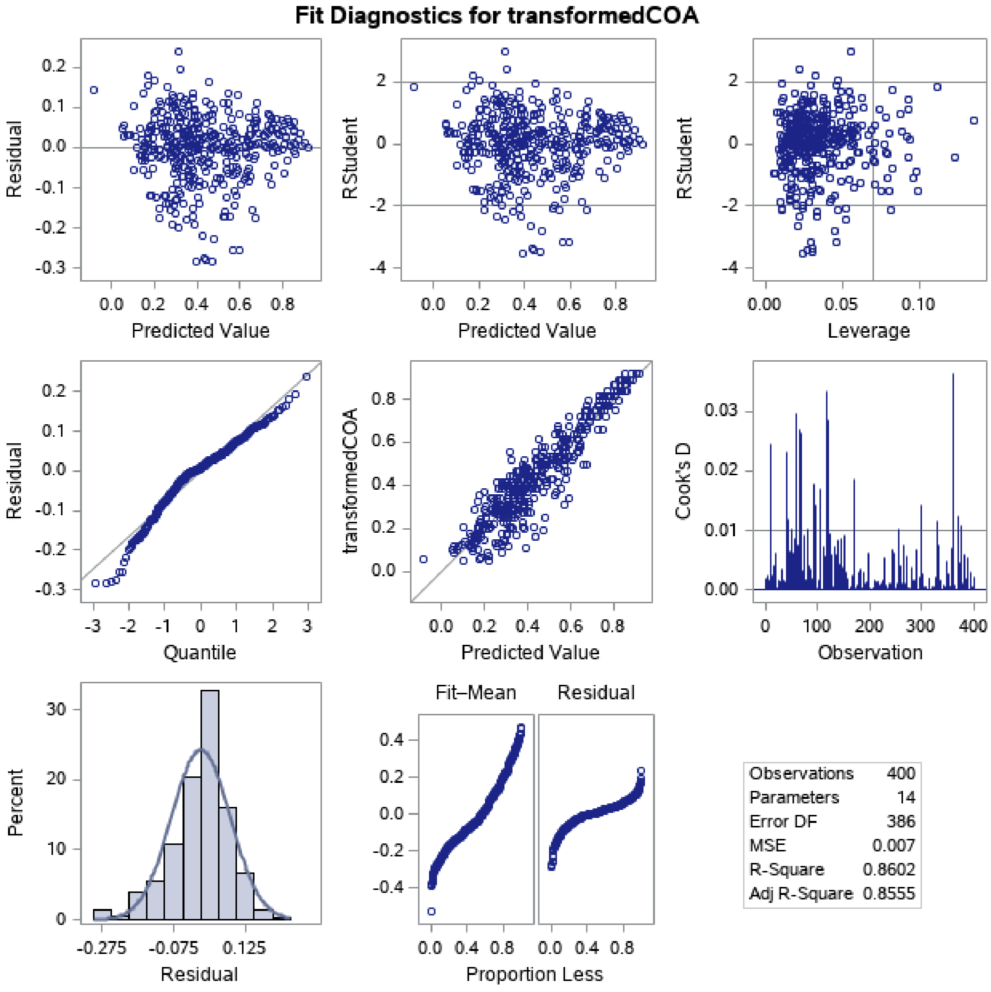
The GLM Procedure

Dependent Variable: transformedCOA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	16.58580850	1.27583142	182.68	<.0001
Error	386	2.69575260	0.00698382		
Corrected Total	399	19.28156110			

R-Square	Coeff Var	Root MSE	transformedCOA Mean
0.860190	18.81827	0.083569	0.444086

Figure 5: Transformed Model Output



Observations	400
Parameters	14
Error DF	386
MSE	0.007
R-Square	0.8602
Adj R-Square	0.8555

Figure 6: Transformed Model Diagnostics

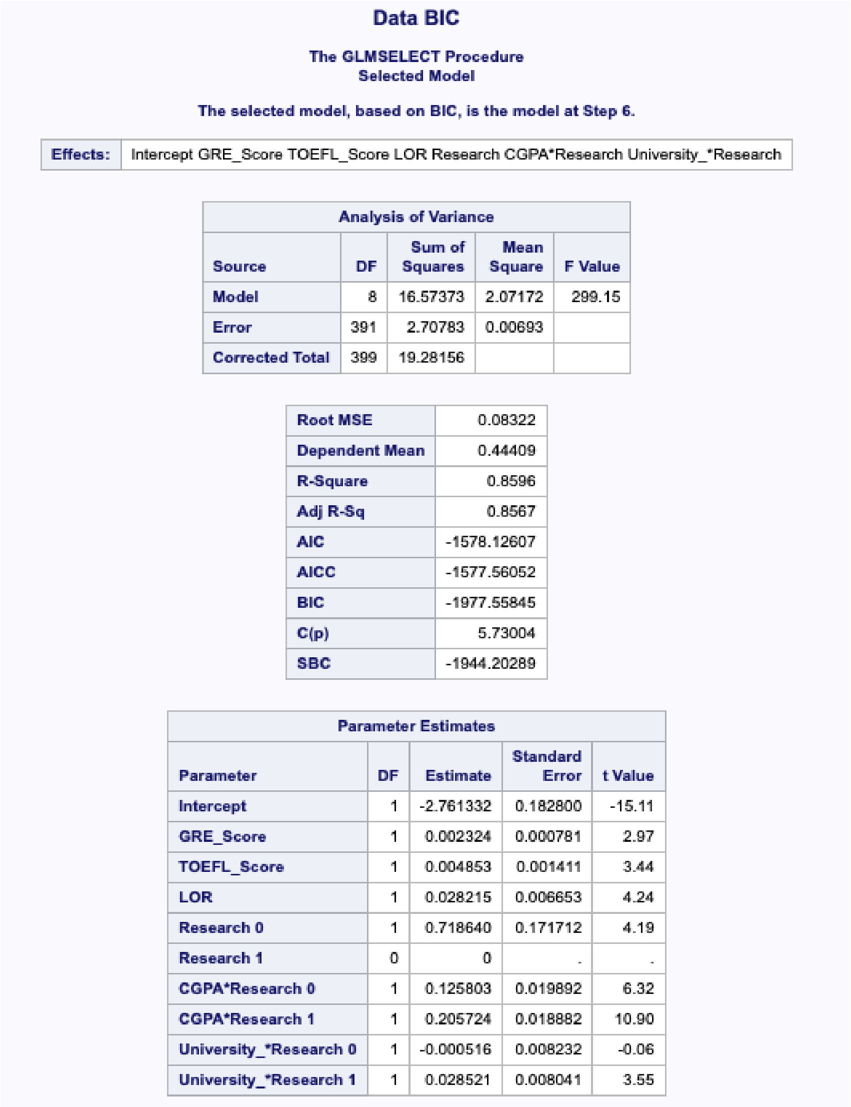
variability in the model with the fewest covariates possible, that is, we want to maximize the Adjusted R-squared value while having the fewest number of covariate. In SAS, the GLMSELECT procedure with BIC can be used in order to select the best model.

The results of this are presented in Figure 7 which shows that only the following covariates are meaningful to the model:

1. GRE Score
2. TOEFL Score
3. LOR
4. Research
5. CGPA * Research (Interaction Term)
6. University Rating * Research (Interaction Term)

Conclusion

Overall in this exercise, using SAS, we determined that not all of the covariates that were provided were important in model selection. Furthermore, 85.67% of the variability in the predicted values could be explained with the reduced model that was determined with BIC. The reduced model resulted in an increase in the Adjusted R-Squared value of 4.95%.



Data BIC

The GLMSELECT Procedure
Selected Model

The selected model, based on BIC, is the model at Step 6.

Effects:	Intercept GRE_Score TOEFL_Score LOR Research CGPA*Research University_*Research

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	8	16.57373	2.07172	299.15
Error	391	2.70783	0.00693	
Corrected Total	399	19.28156		

Root MSE	0.08322
Dependent Mean	0.44409
R-Square	0.8596
Adj R-Sq	0.8567
AIC	-1578.12607
AICC	-1577.56052
BIC	-1977.55845
C(p)	5.73004
SBC	-1944.20289

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-2.761332	0.182800	-15.11
GRE_Score	1	0.002324	0.000781	2.97
TOEFL_Score	1	0.004853	0.001411	3.44
LOR	1	0.028215	0.006653	4.24
Research 0	1	0.718640	0.171712	4.19
Research 1	0	0	.	.
CGPA*Research 0	1	0.125803	0.019892	6.32
CGPA*Research 1	1	0.205724	0.018882	10.90
University_*Research 0	1	-0.000516	0.008232	-0.06
University_*Research 1	1	0.028521	0.008041	3.55

Figure 7: BIC Model Selection