Predicting Admissions

Chinmay Sheth

November 10, 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.

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. SOP
6. LOR
7. CGPA
8. Research
9. Chance of Admit

The dataset contains 400 entries with no data missing.

Diagnostics

Through the PROC GLM function, SAS is easily able to build a basic linear model which we can further investigate to determine that it is satisfactory. PROC GLM is also able to differentiate between categorical and continuous regressors. Here we see the different possible categorical values that can be held in these independent variables.

Multiple Linear Regression -- Admission Prediction Data

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
SOP	9	1 2 3 4 5 1.5 2.5 3.5 4.5
LOR	9	1 2 3 4 5 1.5 2.5 3.5 4.5
University_Rating	5	1 2 3 4 5
Research	2	0 1

Number of Observations Read	400
Number of Observations Used	400

Here we see the different possible categorical values that can be held in these independent variables. The current basic linear regression model contains 26 regressor variables where one is an intercept. Here are the respective estimates for each continuous variable regressor and categorical variable regerssor.

Estimates

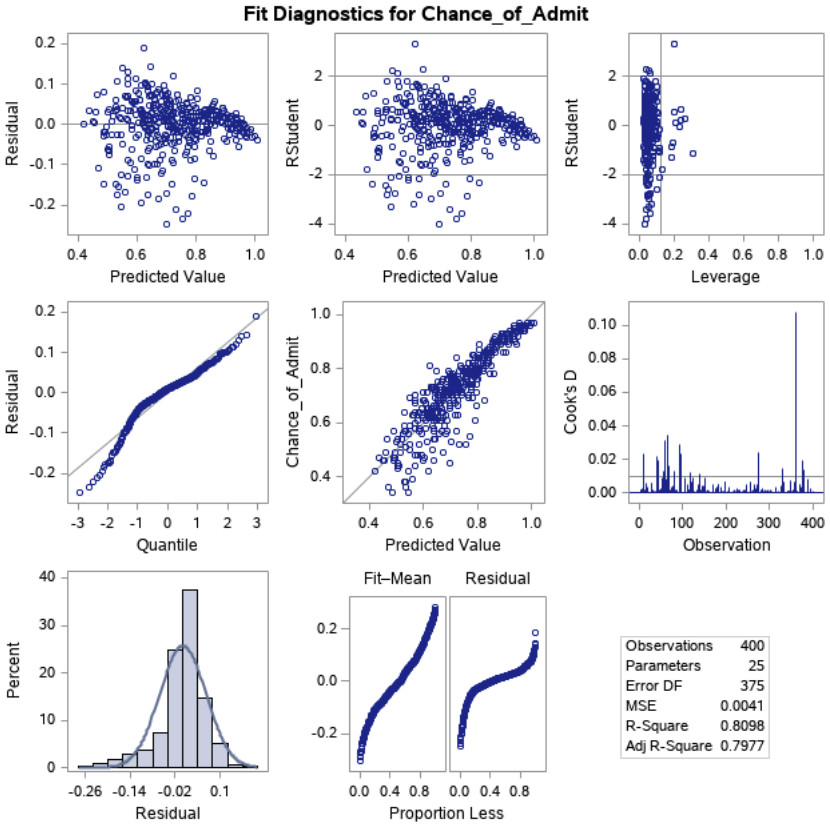
Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1.107001042	B	0.14359964	-7.71	<.0001
GRE_Score	0.001647188		0.00060772	2.71	0.0070
TOEFL_Score	0.003231499		0.00111501	2.90	0.0040
SOP 1	-0.001495195	B	0.03506060	-0.04	0.9660
SOP 2	0.000829510	B	0.01878692	0.04	0.9648
SOP 3	0.009948823	B	0.01485052	0.67	0.5033
SOP 4	-0.006124603	B	0.01229913	-0.50	0.6188
SOP 5	0.001593734	B	0.01499420	0.11	0.9154
SOP 1.5	-0.013286205	B	0.02264320	-0.59	0.5577
SOP 2.5	0.023408226	B	0.01688774	1.39	0.1665
SOP 3.5	-0.004154230	B	0.01430650	-0.29	0.7717
SOP 4.5	0.000000000	B	.	.	.
LOR 1	-0.105685860	B	0.07270737	-1.45	0.1469
LOR 2	-0.049470364	B	0.01871473	-2.64	0.0086
LOR 3	-0.041614638	B	0.01428441	-2.91	0.0038
LOR 4	-0.008740534	B	0.01275542	-0.69	0.4936
LOR 5	0.010286967	B	0.01502952	0.68	0.4941
LOR 1.5	-0.086203904	B	0.02922270	-2.95	0.0034
LOR 2.5	-0.038699436	B	0.01727835	-2.24	0.0257
LOR 3.5	-0.024213357	B	0.01372637	-1.76	0.0785
LOR 4.5	0.000000000	B	.	.	.
CGPA	0.117788966		0.01242238	9.48	<.0001
University_Rating 1	-0.001163373	B	0.02259310	-0.05	0.9590
University_Rating 2	-0.029267239	B	0.01594390	-1.84	0.0672
University_Rating 3	-0.017091742	B	0.01390335	-1.23	0.2197
University_Rating 4	-0.016769065	B	0.01185280	-1.41	0.1580
University_Rating 5	0.000000000	B	.	.	.
Research 0	-0.025495595	B	0.00813488	-3.13	0.0019
Research 1	0.000000000	B	.	.	.

In all linear models we hope that the following four assumptions will be held:

1. $\mathbb{E}(\varepsilon) = 0$
2. $Var(\varepsilon) = \sigma^2$
3. $Cov(\varepsilon_i, \varepsilon_j) = 0$
4. $\varepsilon_i \ N(\sigma, (0, 1))$

In order to determine if these four assumption are held, we can investigate the diagnostic plots that SAS also generates with the PROC GLM function.

Diagnostic Plots

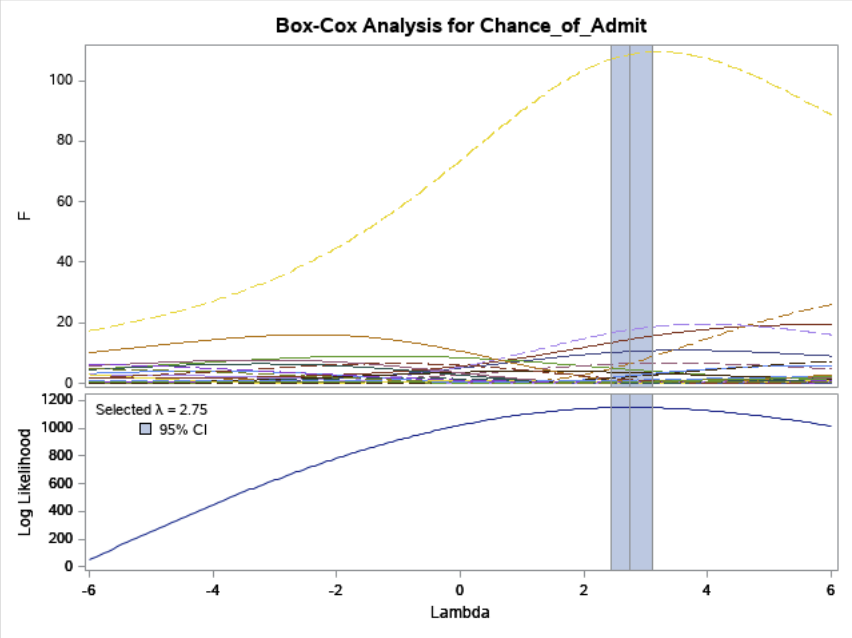


In the Residual vs. Predicted Value plot (top left) we see a fanning pattern in the residuals. While the first assumption of the mean approximately being zero may hold, there seems to be a departure from the second assumption of constant variance; that is, there is hetereoskedasticity present in the dataset. A transform may be required in order to reinstate the constant variance assumption. In the

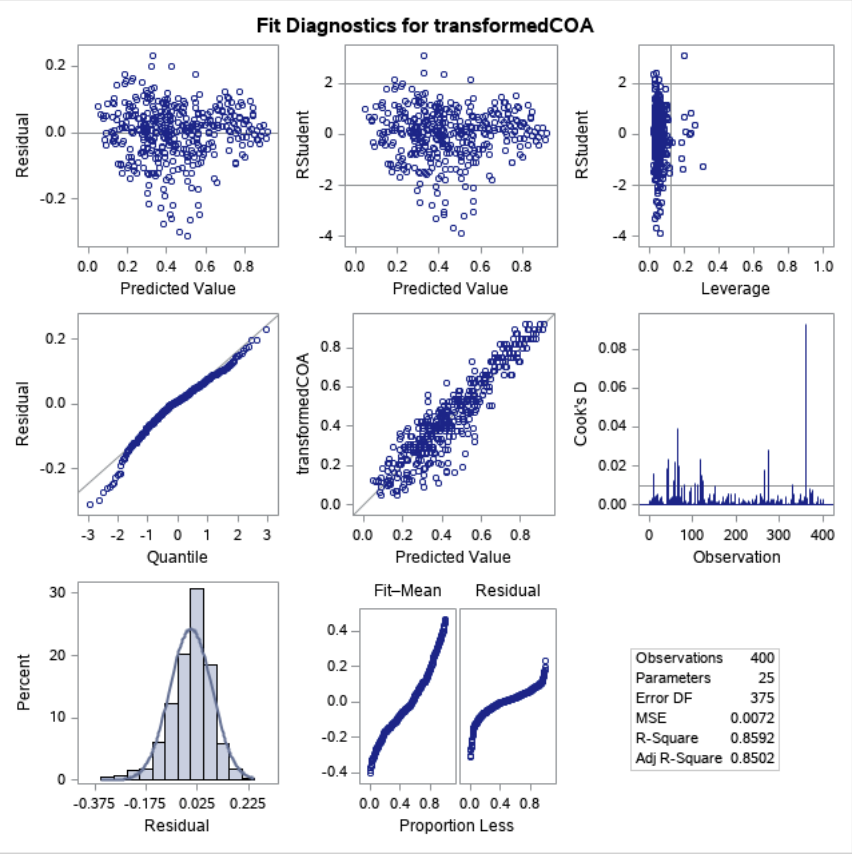
Quantile-Quantile plot (middle left) we see that the points are not hugging the y=x line very strongly. As such, it seems that there is also a departure from the fourth assumption, which is that the residuals are normally distributed. If the residuals were normally distributed, the points would hug the y=x line a lot more tightly. We also know that there are more than 20 observations in our dataset, so we expect the Quantile-Quantile plot to be normal.

In regards with outliers, we see a strongly influential point in the Cook's Distance plot (middle riht) just above approximately 350 on the horizontal axis and it is clearly past the reference line. In fact, we also have a point in the RStudent vs. Leverage (top right) plot which is and outlier and a leverage point since it is past the horizontal and vertical reference lines. Furthermore from the RSdtudent vs Leverage plot there also seems to be a cluster of points that seem to have a lot of influence in the x-space between [0.2, 0.4] on the horizontal axis.

While violating the second and fourth assumptions, our R^2 value sits at 0.809. Perhaps the model can be transformed so that this R^2 value can be increased so that there is a better fit. The Box-Cox method is a good fit for this type of problem because there is no clear solution as to what the transformation should be. If the $Var(e)$ $E(y)$, it would make sense to apply a square root transformation or if $Var(e)$ $E(y)^2$ then it would make sense to apply a log transformation. However, neither of these are the case and a Box-Cox transformation seems like the best option.



From our Box-Cox analysis, the ideal $\lambda = 2.75$. After applying the appropriate transformations to the dependent variable, here is what the diagnostics look like:



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	24	16.45592841	0.68566368	95.35	<.0001
Error	375	2.69674155	0.00719131		
Corrected Total	399	19.15266996			

R-Square	Coeff Var	Root MSE	transformedCOA Mean
0.859198	18.86312	0.084802	0.449563

Here the scatter in the Residual vs Predicted Value plot is a lot more random, and the fanning pattern seems to have disappeared. The residuals have also adhered a lot more to the constant variance assumption. The R^2 value has also increased from 0.809 to 0.859 indicating that the new model is able to explain more of the variance in the dataset. Furthermore, the points on the Quantile-Quantile plot also seem to hug the y=x line a lot more than they did in the previous model's Quantile-Quantile plot. Overall the new model is definitely

better in fit than the previous model.