Linear Significance In University Ranking

Christopher Jung, Jiayi Wang, Suhyun Park, Haoya Wang

Department of Statistics, University of Toronto

STA302: Methods of Data Analysis I

Professor Katherine Daignault

December 12, 2023

Contributions

Chris: Model Assumptions, additional conditions in the methods and results sections, discussion section. Led the group through the entire process and assigned each part

Jane: Inference and Decomposition in the methods section and model assessments in results section

Sue: Diagnostics and Model selection in methods section and diagnostics in results section

Haoya: Introduction section and Ethics section

Introduction

World university rankings have become increasingly popular as an indicative measure of a university's performance. We wanted to research "How do the number of students, teaching score, research score, industry income score, and proportion of international students (categorical variable) influence the overall score of a university, response variable? What is the significance of each predictor variable to a change in the response variable?" The purpose of this report is to examine the factors affecting the world rankings of a university through analyzing the relationship between the overall score of a university (which is used to determine its ranking relative to other universities) and 5 predictors of performance mentioned in the research question. Our hypothesis is that these 5 predictors affect the overall score significantly and the research score affects the overall score the most. In this report, the proportion of international students is treated as a categorical predictor divided into four levels. Although previous literature has examined the effect that some of these variables have on a university's overall score (and hence its performance in world university rankings), they have not considered as many factors that contribute to overall university performance nor analyzed the possibility of interrelation between predictor variables (Lukman et al., 2010; Moed, 2016; Tan et al. 2014). Thus, this report will attempt to address these gaps by proposing a more complete regression model that examines the relationship between the overall score of a university and several other factors not previously considered in tandem before (such as the simultaneous effect of teaching score and international student proportion on overall university performance) to shed greater light on the variables that truly contribute to university performance and hence, their international ranking.

Methods

Data

The data we imported is a cleaned data version of World University Ranking 2023 (Tisha). International students, a numerical variable, representing the proportion of international students to a categorical variable of 4 levels.

Assessing Model Assumptions

First, we check the two additional conditions, the conditional mean response (Condition1) and the conditional mean predictor (Condition2). We study Condition1 through the Response Vs Fitted scatterplot and Condition2 through the piecewise scatterplots of the predictors. If there are any violations then the residual plots may lead to misleading conclusions. Then we study the residual plots for the four assumptions that we want the model to hold.

For linearity, constant variance, and uncorrelated errors, the Residual vs Fitted scatterplot is studied. For any fanning or too diverse spread, it is a violation of constant variance. To address this, we must use variance-stabilizing transformation on the response. For curved or non-linear patterns, it is a violation of linearity. To address this, we must use BoxCox transformation on each predictor until the violation disappears. If there are two or more distinct clusters, it is a violation of uncorrelated errors.

For normality, the Q-Q plot is studied. If there is a severe deviation from the line or discontinuous lines, it is a violation of normality. To address this, we must use BoxCox transformation on the response.

For this analysis, we will take iterations of checking the methods mentioned in this section. Once necessary transformations are performed, we pick the most desirable model. Finally, we must check again if the additional conditions and model assumptions of the model holds, specifically normality assumption, to be able to utilize any normal distribution properties in the Inference section.

Inference

In inference, we conduct an ANOVA test to assess the relationship between all the predictor variables and the response variable. The null hypothesis, H_o ,assumes that all predictor slopes are equal to zero. If the p-value of ANOVA is less than the significance level at α = 0.05, it will indicate H_A ,a statistically significant linear relationship for at least one predictor.

Subsequent to the ANOVA test, a significance test is performed for each coefficient estimate of the predictors. The null hypothesis, H_o, is that the corresponding coefficient is equal to zero. At α = 0.05, if a predictor variable exhibits p-values below α, we reject H_o, confirming the presence of significant linear relationships to the response variable in presence of other predictors.

Next, we conduct a Partial F-test to assess whether a reduced model is as good as the full model, T-model. The null hypothesis states that the coefficients of the *k* predictors that are tested to be dropped are simultaneously equal to zero. If the p-value from the Partial F-test is less than the significance level (α = 0.05), the null hypothesis is rejected, indicating that out of the *k* predictors, there exists at least one predictor that is linearly significant with Y.

Diagnostics

In the diagnostic process, we look for potential issues. We first study the scatter plots to identify any problematic observations such as leverage points, outlying points, and influential observations in the residuals plots. Leverage points may possibly shift the regression line. We find leverage value, h_{ii}. Next, we assess outlying points that can affect constant variance. So, we find standardized residuals, $r_{i,}$ where residuals measure “outlying-ness” accounting for leverage, standardizing error variance.

Influential observations influence how the model is estimated. We measure Cook’s Distance to see if an observation affects the estimation of all fitted values. Other measurements such as $DFFITS_i$ and $DFBETAS_i$ are measured for other effects the observation has. For these measurements, we decide if it is substantial or not by their respective cutoffs.

Multicollinearity is studied with Variance Inflation Factor (VIF) to identify if more than 2 predictors are related. VIF values greater than 1 indicate the presence of multicollinearity, and those exceeding 5 suggest severe multicollinearity. A severe multicollinearity can lead to predictors in hypothesis tests of linear significance failing to reject the Null when it should reject. To address severe multicollinearity, we can change the form of the predictors. To address multicollinearity, we can drop predictors. However, we must try not to remove variables of interest. A model with the predictor with the highest VIF value dropped is introduced as a V-model.

Model Selection

We use manual selection to select our model. The tools for model comparison are Adjusted R-squared, providing insights into the goodness of fit while adjusting for the number of predictors, Akaike Information Criterion(AIC) to assess the trade-off between model fit and complexity, corrected AIC, BIC.

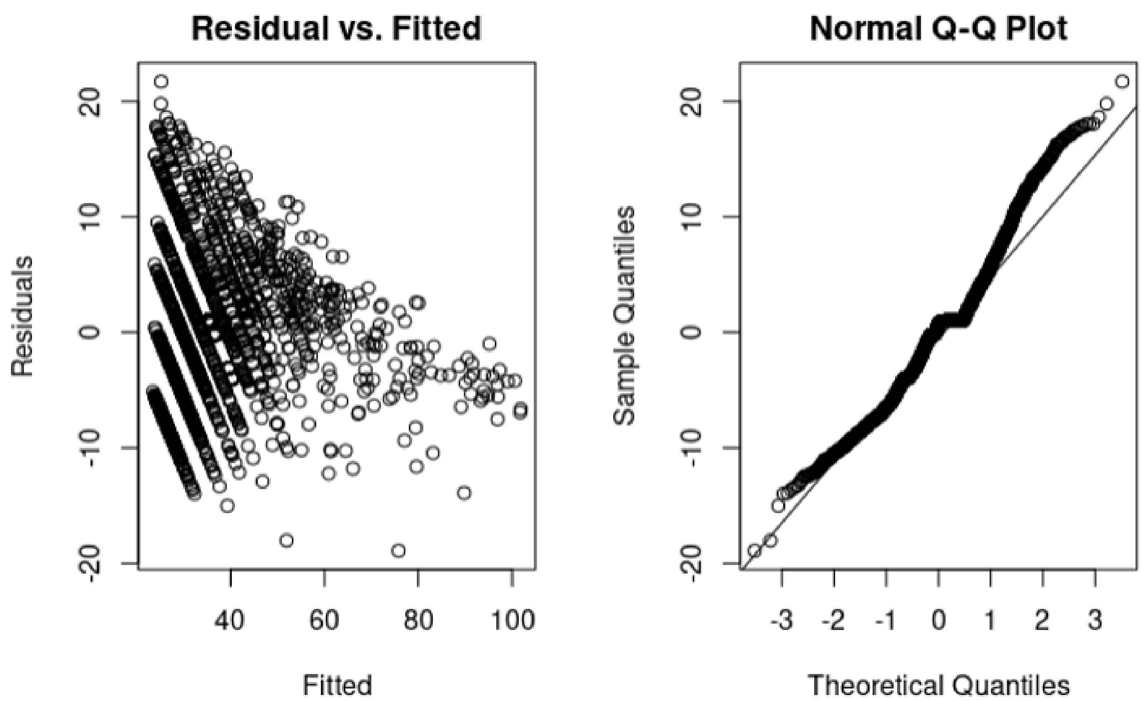
The assessment begins with evaluating the Adjusted R-squared values of the T-model and V-model to determine their respective explanatory powers, accounting for the number of predictors. Additionally, we compare VIF values. A lower AIC indicates a better balance between model fit and complexity. The adherence to model assumptions, including linearity and normality of residuals, is assessed for both models. The comprehensive evaluation includes considering the number of predictors and their significance in each model. The final model is selected by striking a balance between all factors. (940)

Results

From exploratory data analysis, we found normality or linearity violations may occur due to right skew in the response variable. Linearity violation may have occurred due to right skew in the predictor variable.

Figure 1: Model Assumptions plots for the originally fitted model

This figure shows the Residual Vs. Fitted model on the left and the Normal Q-Q plot on the right. It was used to study the improved model assumptions of the original model. There appears to be many violated assumptions.

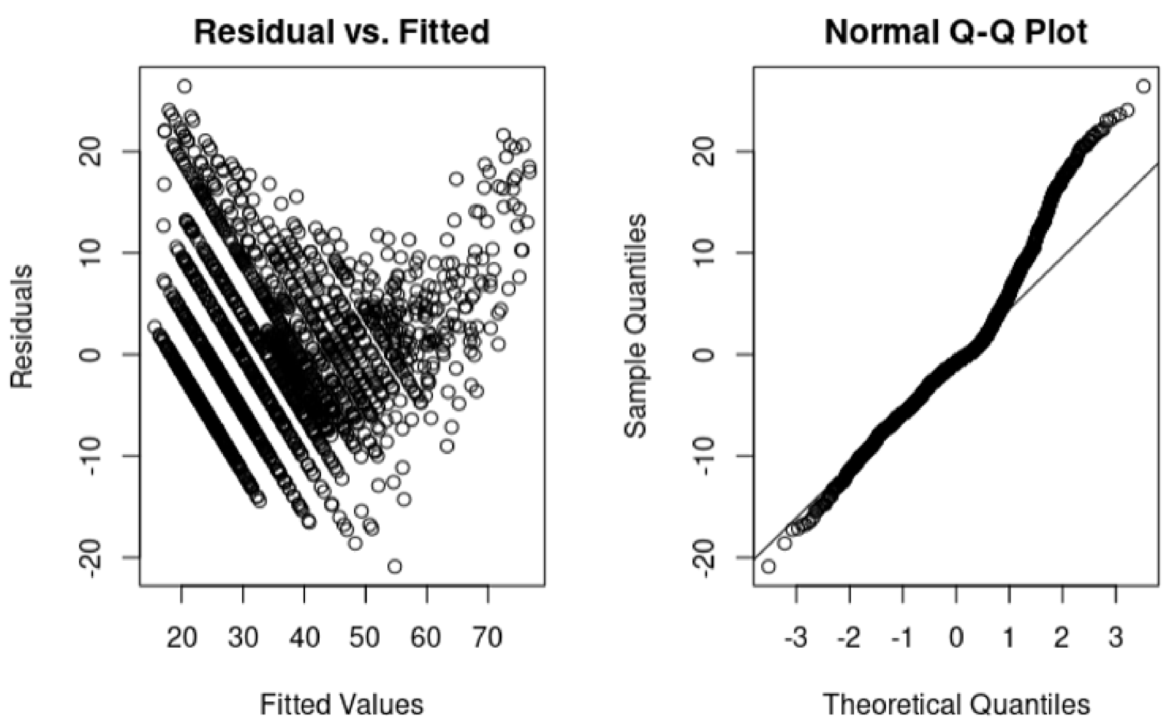


Condition1 was not violated because there were no obvious non-linear patterns. Condition2 was not violated because the predictors were at most linear with another (see Appendix A). In Figure 1, the Residuals Vs. Fitted Values scatterplot showed that the model may violate the constant variance. The Q-Q plot showed a possible normality assumption violation, but not as severe. The Q-Q plot displayed right-skew which matched EDA's right skewness.

To address constant variance violation, we performed a variance-stabilizing transformation by applying log onto the response variable to address the violated assumption, but it worsened the constant variance and developed linearity assumption violation. Next, the BoxCox transformation on the response variable was applied, but there were no improvements to the violated normality assumption.

Figure 2: Model Assumptions plots for the Transformed model T-model

This figure shows the Residual Vs. Fitted model on the left and the Normal Q-Q plot on the right. It was used to study the improved model assumptions of the model after it was transformed into T-model. There appears to be a slight right-skew according to the Q-Q plot, but it is due to the limitation of the original data.



Finally, we attempted BoxCox transformation on the predictors by using simple powers. The transformation satisfied the additional conditions (see Appendix B). It improved the model assumptions. Due to the limitation of the data being too heavily right-skewed, the Q-Q plot of the predictors-transformed model still displayed some right-skew. On the other hand, the transformation addressed the constant variance violation and made each predictor to be closer to being normal. The predictor-transformed model that was chosen during this step was named T-model.

Table 1: Summary Table for Transformed Model 3

Results of the summary of the transformed model 3, which contains the intercept and estimates (coefficients), standard errors, t-value and p-values for the T test for each predictor variable in the model, and residual standard error, R-squared value, F-statistic and p-value for the ANOVA test.

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	1.352e+01	3.279e+00	4.123	3.88e-05
fifth root of number of student	9.199e-01	1.073e-01	8.570	< 2e-16
negative cube root of teaching score	-9.857e+01	5.807e+00	-16.975	< 2e-16
logarithm of research score	1.530e+01	4.976e-01	30.742	< 2e-16
negative cubed of industry income score	3.142e+05	3.352e+04	9.373	< 2e-16
international student 11-20%	7.441e-01	3.596e-01	2.069	0.0387
international student 21-30%	5.146e+00	5.496e-01	9.362	< 2e-16
international student >30%	6.196e+00	5.522e-01	11.220	< 2e-16

Residual standard error: 6.63 on 2333 degrees of freedom
Multiple R-squared: 0.7487
Adjusted R-squared: 0.7479
F-statistic: 992.9 on 7 and 2333 DF
p-value: < 2.2e-16

We conducted an ANOVA test to check if all predictors have a 0 slope or if at least one predictor is exhibiting a linear relationship with the response variable. According to Table 1, the p-value < 2.2 e-16, is below the significance level ($\alpha = 0.05$), indicating that a statistically significant linear relationship exists for at least one predictor.

Further analysis involved testing the significance of each coefficient estimate for the transformed predictors using the T-model. Table 1 shows that all predictors exhibited p-values below the significance level of $\alpha = 0.05$, leading to the rejection of the null hypothesis. This confirmed the presence of significant linear relationships between these predictors and the overall score.

Given the absence of insignificant predictors and the statistical significance of all transformed predictors, a partial F-test was not conducted. Consequently, the best-fitting model for our analysis was identified as the T-model, suggesting that it adequately captures the linear relationships between the chosen predictors and the overall score of educational institutions.

Table 2: Diagnostic values of T model and V model

These two tables below show the value of Adjusted R^2, p-value for each predictor, VIF values, h_ii, r_i, D_i, DFFITS, DFBETAS, AIC, BIC and corrected AIC for T model and V model.

<T model>

predictor	Adjusted R^2	p-value	VIF	h_ii	r_i	D_i	DFFITS	DFBETAS	AIC	BIC	corrected AIC
	0.7479			266	0	0	175		8864.462	8910.529	8864.539
(intercept)		3.88E-05						183			
fifth root of number of students		< 2e-16	1.05025					124			
negative cube root of teaching score		< 2e-16	2.776339					208			
logarithm of research score		< 2e-16	3.795208					193			
negative cubed of industry income score		< 2e-16	1.81583					150			
international student 11-20%		0.0387						138			
international student 21-30%		<2e-16	1.266267					111			
international student >30%		<2e-16						126			

<V model>

predictor	Adjusted R^2	p-value	VIF	h_ii	r_i	D_i	DFFITS	DFBETAS	AIC	BIC	corrected AIC
	0.7479			327	1	0	146		9658.62	9698.93	9658.684
(intercept)		< 2e-16						168			
fifth root of number of students		< 2e-16	1.03342					122			
negative cube root of teaching score		< 2e-16	1.52184					196			
negative cubed of industry income score		1.03E-06	1.41232					171			
international student 11-20%		6.82E-07						156			
international student 21-30%		< 2e-16	1.16203					116			
international student >30%		< 2e-16						121			

To compare T-model and V-model for predicting university rankings, various key factors were considered to determine the superior model.

Firstly, examining the coefficients of determination revealed that T-model provides a better fit to the data and explains a larger proportion of the variance in the response variable. T-model achieved a higher Adjusted R-squared compared to V-model.

Regarding the number of predictors and their significance, T-model excels with a comprehensive set of predictors. Each predictor in T-model demonstrates statistical significance ($p < 0.005$). On the other hand, V-model comprises a reduced set of predictors, excluding the research score variable. Despite its simplicity, V-model still maintains statistical significance. However, the absence of a transformed research score might result in a partial representation of factors influencing the overall score. Therefore, considering both the breadth and significance of predictors, T-model emerges as the more comprehensive model.

With VIF values below 5, both models effectively managed multicollinearity. However, V-model generally shows lower VIFs, suggesting a better control over multicollinearity.

Looking at the Table 2, we observe that V-model has more problematic observations than T-model. T model surpasses V-model with lower AIC, BIC and corrected AIC values.

In conclusion, while both models demonstrated competence in predicting university rankings, T-model stood out as the preferred choice. It exhibited superior performance in terms of explaining variance, maintained low VIFs, and upheld model integrity with more predictors than V-model. Thus, T-model achieves a favorable balance between complexity and explanatory power, making it the model of choice for this analysis.

Discussion

The final model finds the Overall Score where we have predictors as fifth root of Number of Student, inverse cube root of Teaching score, log Research Score, inverse cubed Industry Income Score. The original hypothesis was that the predictor variables of interests are all significant and industry income score is the most significant. After the transformation to address model assumptions, it is shown that the industry income score is the most significant in a transformed setting.

To answer the research question, each predictors affected the overall score conditionally. After transformation, the predictors that increase the mean overall score per one unit increase when other predictors are fixed were numbers of students, research score, industry income score.
We expected these predictors to affect the overall score significantly based on the three literature and the result came out to justify this expectation.

There were few limitations. The response variable and the predictor variables were all heavily right skewed. So, despite assuming the population was normal and the final model to be close to being normal, there still remains some deviation in the Q-Q plot. Since the final model is transformed from the preliminary model that was used to answer the research question, what the model answers is different from the original hypothesis by condition.

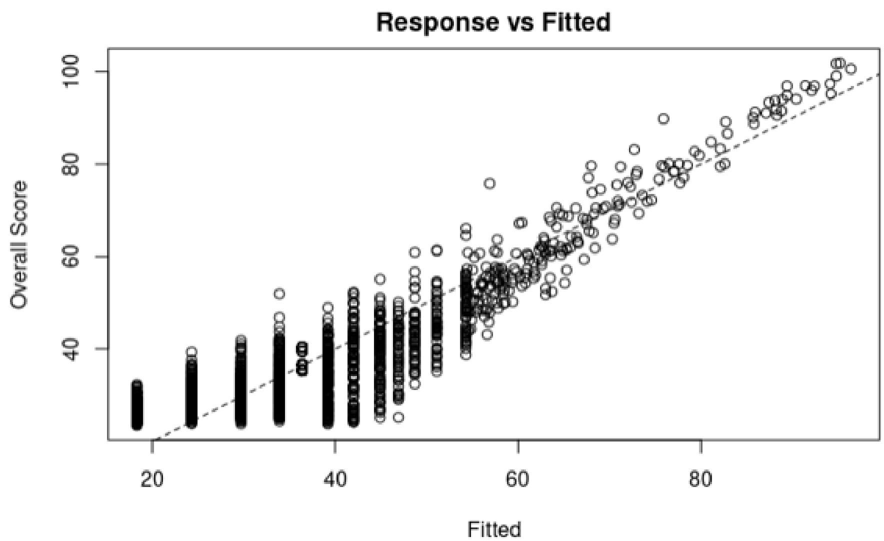
Ethics Discussion

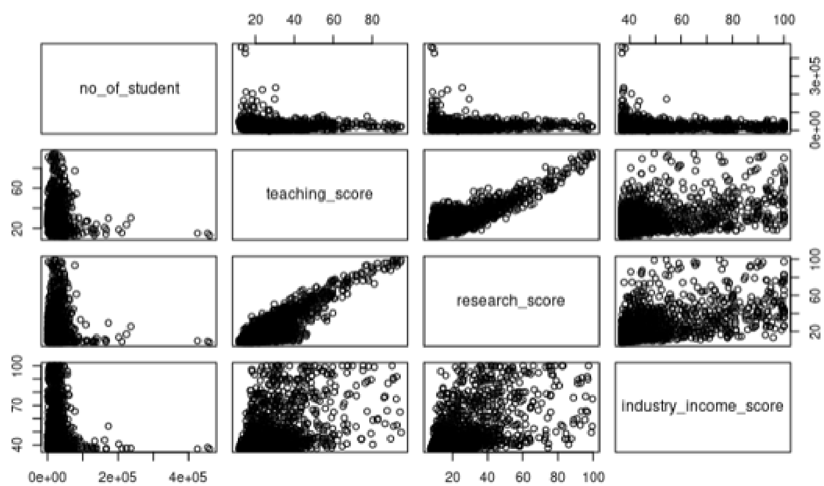
In this report, the manual selection method was chosen over the automated selection method when deciding the best regression model to describe the data collected. Particularly, the adjusted R-squared values of the T-model (0.7479) and the V-model (0.646) were compared to find the largest value, which indicates a better model fit for the data. In addition, the AIC values of both models were also compared, and the final model (T-model) determined to be the most representative of the data collected was the one which had the largest adjusted R-squared value and the smallest AIC value. While the automated selection method may have been more efficient in arriving at the final regression model, the manual selection method was preferred in our deduction because it accounted for factors which may have been neglected if solely relying on automated selection, making it the better choice both practically and ethically. For example, in the report, the manual selection method enabled us to consider both the impact of the adjusted R-squared value and the AIC value when determining the best regression model, while the automated selection model would have limited us to only analyzing the effects of the AIC values on the models. Furthermore, the manual selection method enabled us to pay greater attention to predictors known to have some relation to overall university performance from previous literature, such as the proportion of international students, which importance may have been neglected if simply relying on automated selection (Tan et al. 2014).

Appendix

Appendix A: Preliminary Model Additional Conditions plots

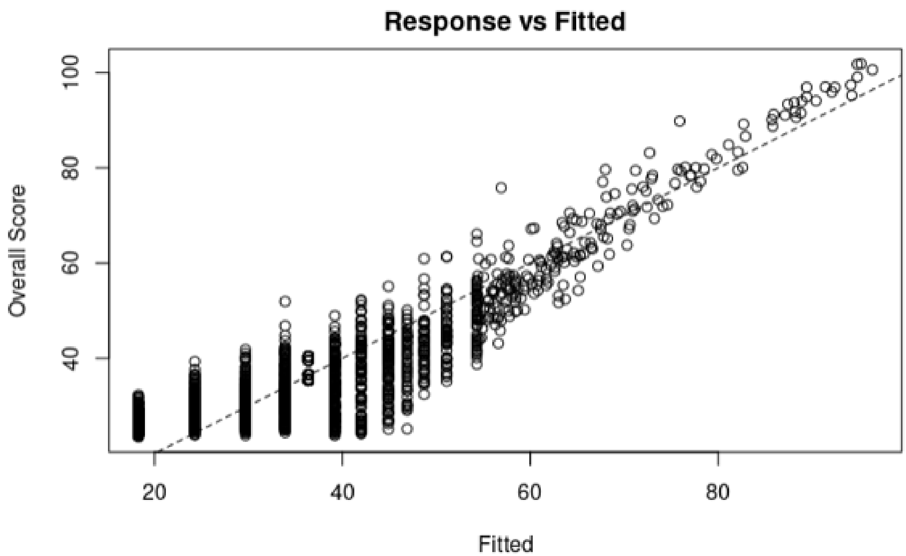
This figure shows the Response Vs. Fitted model on the top and the piecewise scatterplot on the bottom. The plots show that the Additional Conditions: condition1 and condition2 hold

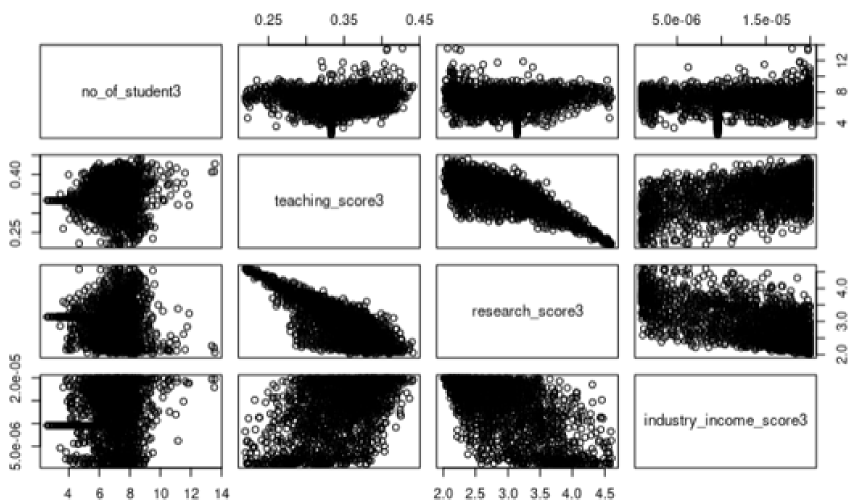




Appendix B: Transformed Model (T-model) Additional Conditions plots

This figure shows the Response Vs. Fitted model on the top and the piecewise scatterplot on the bottom. The plots show that the Additional Conditions: condition1 and condition2 hold





References

1. Lukman, Rebeka & Krajnc, Damjan & Glavič, Peter. (2010). University ranking using research, educational and environmental indicators. Journal of Cleaner Production. 18. 619-628. 10.1016/j.jclepro.2009.09.015.
2. Moed, H.F. A critical comparative analysis of five world university rankings. *Scientometrics* **110**, 967–990 (2017). https://doi.org/10.1007/s11192-016-2212-y
3. Tan, Y.S., Goh, S.K. International students, academic publications and world university rankings: the impact of globalisation and responses of a Malaysian public university. *High Educ* **68**, 489–502 (2014). https://doi.org/10.1007/s10734-014-9724-2
4. Tisha, Samia Haque. World University Rankings 2023 - Cleaned. https://www.kaggle.com/datasets/samiatisha/world-university-rankings-2023-clean-dataset/data