RESEARCH PROJECT
UNIVERSITY ADMISSIONS

Casiraghi Matteo Giacomello Tommaso

January 2023 - 30549 Mathematical Statistics

Contents

1 Introduction and purpose of the project

The admission to the graduate university is one of the most crucial moment of a person's life, since it will influence the future and the education of a student. It is the beginning of the specialization, therefore a turning point. The student's choice of one university over another can influence remarkably his chances to fulfill his career aspirations and to achieve personal goals. There are people who dream and desire to enter in a specific university, aware of the opportunities that only some schools can offer. In most of the cases, the university system creates limited enrollment degree courses, therefore not everyone is granted access in. So only students with certain merits are admitted. With this in mind it becomes crucial for every student to understand what these merits are. These are the reasons why we have chosen to investigate what requirements affect considerably the chances of admissions in universities.

In this paper we decided to consider the country that boasts the best degree courses in STEM and so the most popular destination to find the perfect college, hence we'll analyze the criteria that American universities use to admit someone and after that we'll observe which one of the selected variables is more influential.

2 Explanation of variables

In America the graduate university system consider a wide span of factors to admit to each degree course only the most deserving students:

The undergraduate curriculum (GRE or GMAT, TEOFL or IELTS, CGPA), the prestige of the undergraduate university (University Rating) and the behavior, the areas of interest and the career goals and beliefs of each student (SOP, LOR, Research Experience).

$\boldsymbol{Y}$ **- Chance of Admit** (ranging from 0 to 1).

$\boldsymbol{x_1}$ **- GRE Scores** (out of 340): the student score attained in a standardized test intended to measure verbal and quantitative skills and abstract thinking of the student, not considering his fields of specialization.

$\boldsymbol{x_2}$ **- TEOFL Scores** (out of 120): the student score attained in a standardized test to measure his English language ability.

$\boldsymbol{x_3}$ **- University Rating** (out of 5): the score attained by the undergraduate university previously attended by the student on the base of the rating system of the QS International trade rankings.

$\boldsymbol{x_4}$ **- Statement of Purpose (SOP)** (out of 5): the score assigned to the motivational letter written by the student.

$\boldsymbol{x_5}$ **- Letter of Recommendation (LOR)** (out of 5): the score assigned to the letter written by some student's professor.

$\boldsymbol{x_6}$ **- CGPA** (out of 10): the cumulative grade point average, which is the score obtained by a student on the base of his performances during his undergraduate career.

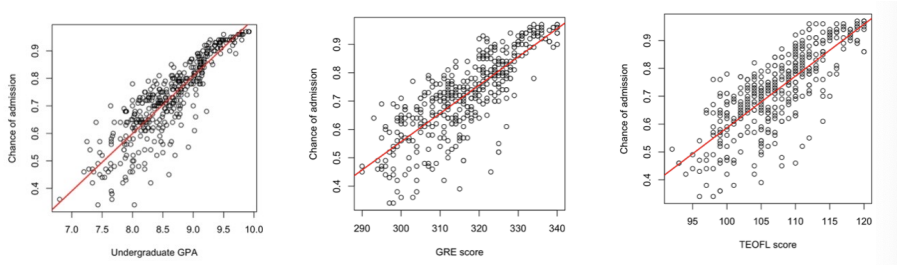
$\boldsymbol{x_7}$ **- Research Experience** (either 0 or 1): variable that specify if the students did external researches not connected with his plan of study.

3 Data Cleaning

We found the dataset, we formatted it and ran a careful check to analyze if there were missing elements and if all of them were manageable to complete the regression and the other operations in R. We did not detect anything of unusual and we proceeded renaming the labels of some columns to have only short names and to make the dataset cleaner.

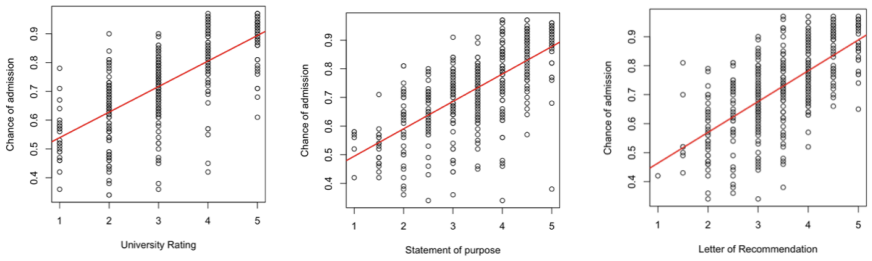
4 Exploring the dataset

We have considered a lot of predictor variables so it is impossible to have a visual representation of the correlation between the **Chance of Admission** and all the factors of merits, this because we are working in high dimensions. Thus we have plotted graphs of the dependent variable $\boldsymbol{Y}$ against each $\boldsymbol{x_i}$ $\forall i \in \{1, ..., 6\}$ and we have colored in red the line that best fits the data in every case. We didn't report the graph taking into account $\boldsymbol{x_7}$ (the **Research Experience**) because the dummy variable didn't show relevant information in the graphical visualization.



(a) (b) (c)

The undergraduate curriculum seems to be one of the main factors to select deserving students, indeed $\boldsymbol{x_2}$ and $\boldsymbol{x_1}$ show stronger correlation (0.79, 0.8) than all the data concerning the status of the universities and the personal features of the students. Moreover, the most notable correlation is obtained by the first element of the curriculum, $\boldsymbol{x_6}$ which achieves a correlation coefficient of 0.87. This is easily verifiable thanks to the first plot.



(d) (e) (f)

However, also in these three graphs we report high correlation coefficients around 0.7, but which are smaller than the first ones, in fact, the points in the graphs are subject to a large variance. Therefore, these elements have a minor impact, probably due to the choice of universities to give more importance to the objective scores obtained during the undergraduate career. Nevertheless, the entire group of variables play an important role to affect $\boldsymbol{Y}$. In addition, $\boldsymbol{x_7}$ shows the lowest correlation coefficient (0.5), probably because colleges give less value to the **Research Experience** taking into account that the majority of the students begins these kind of analysis after the graduation.

5 Multilinear Regression

Now we have all the tools to proceed with the multivariate regression using $\boldsymbol{Y}$ as Response variable and the $\boldsymbol{x_i}$**s** as explanatory variables. We begin with the full model to have a wider view of all variables' significance. In the next image we observe that all the elements except $\boldsymbol{x_3}$ and $\boldsymbol{x_4}$ have a p-value less than 0.05. This is something we expected since, first of all, the p-value of $\boldsymbol{x_3}$ can be explained by the fact that colleges give less weight to the undergraduate **Universities Rating** because not everybody has economical resources to attend

high valued universities.

Secondly, $\boldsymbol{x_4}$ and $\boldsymbol{x_5}$ are variables strictly connected and frequently they show analogous results (the average difference between them is around 0.5), indeed is usual that a professor highlights the same skills emphasized by his student and taking in consideration both variables make us face the risk of overfitting if we doublecount this contribution.

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.2594325  0.1247307 -10.097  < 2e-16 ***
GRE                0.0017374  0.0005979   2.906  0.00387 **
TOEFL              0.0029196  0.0010895   2.680  0.00768 **
University.Rating  0.0057167  0.0047704   1.198  0.23150
SOP               -0.0033052  0.0055616  -0.594  0.55267
LOR                0.0223531  0.0055415   4.034  6.6e-05 ***
CGPA               0.1189395  0.0122194   9.734  < 2e-16 ***
Research           0.0245251  0.0079598   3.081  0.00221 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06378 on 392 degrees of freedom
Multiple R-squared:  0.8035,    Adjusted R-squared:    0.8
F-statistic: 228.9 on 7 and 392 DF,  p-value: < 2.2e-16
```

(g)

Knowing what previously said, in the next chapter we'll search for the best model which describes better the data and will check the assumption of the Regression regarding residuals only in that model.

6 Model selection

6.1 Step-down

In order to perform the Step-down method we begin the multivariate regression taking into account all variables and we discard the covariate with larger p-value (always > 0.05) and we iterate this process until we get all p-values below 0.05.

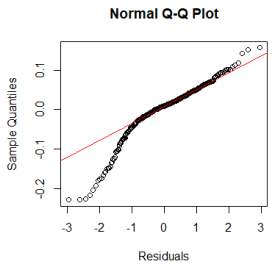
In this case we omit from our analysis $\boldsymbol{x_4}$ during the first step and $\boldsymbol{x_3}$ in the second one. Therefore we get a model as efficient as the full one, since we obtain two similar $\boldsymbol{R^2}$ and in order to seek for the best combination of covariates and response variable we use another method of model selection.

6.2 Cross Validation

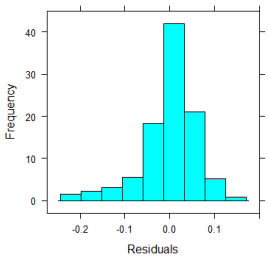
We execute a Cross Validation by randomly divide our data into two parts: *Training Data* to estimate the $\hat{\beta}_d$ $\forall d \in \{1, ..., 7\}$ and *Testing Data* used to compute the residuals sum of squares on the new data. Lastly we find d, the minimizer of the residuals sum of squares.

We perform the Cross Validation on four models: the full one, one without $\boldsymbol{x_4}$, one without $\boldsymbol{x_3}$ and one without both. This because the two selected variables were the discarded ones from the Step-Down Method. The result was exactly what we expected on the base of what we have discussed in *chapter 4*. The best model is the second one where we have neglected $\boldsymbol{x_4}$, so the analysis shows the evidence that **SOP** causes overfitting, while $\boldsymbol{x_3}$ differently from what we found in the Step-Down is considered by the model as a relevant variable. Probably, universities take into account this rating, despite giving it less weight.

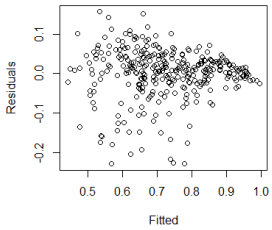
6.3 Checking Assumption



(h)



(i)



(j)

Finally, we conclude our analysis by checking *normality* and *homoscedasticity* of the residuals in the best model. As we can see graphically the constant variance is not respected since the data are spread following a fluctuating trend. Regarding *normality* we perform Shapiro and K-S tests, which give us statistical evidence to reject normality assumption, confirming what is visible in the plots above. In the light of this, we will discuss this problem in the next chapter.

7 Conclusion and limitation of our study

To conclude, we can talk about what problems we have encountered during the tasks and what outcomes we have found. We understand that some factors are considered more than others because personal skills that could come out from the **Letter of Recommendation** or the **Statement of Purpose** are not tested and certificated as **TOEFL** and **GRE** scores. In our first analysis, the full model seems to offer a nice correlation, but at the end a simpler model with less predictor variables showed up as the best one for us. We are aware that the assumptions on the residuals are not satisfied, consequently we cannot look at the model as the perfect one to explain reality but it is close to do it. We must always remember that we are working on a small-scale sample.

Moreover, we faced-off other issues. During the collection of the data we were looking for a ready-made dataset related to our topic, this due to the fact we thought that we would encounter more errors if we collected data by ourselves, hence we found this dataset with the variables reported above and we did a cross-check on different sites to make sure selected data were reliable information. We found that what we had recovered was excellent material, but we detected a bias that could have distorted the accuracy of our analysis. As previously mentioned, we used **GRE** and **TEOFL** scores despite there exist also GMAT and IELTS tests. Nobody can send an application containing both **GRE** and GMAT marks and the same reasoning works for **TEOFL** and IELTS, because universities require only one of each pair. Therefore, it is important to remember we restricted our study to the students that took only the above reported tests.

Now, to summarize we would like to emphasize the most fulfilling result of the project. Among all variables, the one who affect mostly the **Chance of Admission** is the **CGPA** (i.e. the grade points earned during our studies), so we can be glad that applying for universities we will be rewarded for our efforts.

APPENDIX :

Sitography and Bibliografy

1. https://www.kaggle.com/datasets/akshaydattatraykhare/data-for-admission-in-the-university
2. Fetsje Bijma, Marianne Jonker, Aad van der Vaart, An Introduction to Mathematical Statistics, Amsterdam University Press, 2016
3. https://www.wikipedia.org/
4. https://www.braingainmag.com/the-most-important-elements-of-your-grad-school-application.htm
5. https://www.investopedia.com/articles/personal-finance/020315/applying-grad-school-gpa-vs-work-experience.asp
6. https://gradschool.duke.edu/admissions/how-choose-right-graduate-school/how-prepare-strong-graduate-school-application/

Missing image

	R2	RMSE
1	0.8076527	0.06624360
2	0.8109050	0.06558769
3	0.8079016	0.06626461
4	0.8102797	0.06576415

(k)

Here we see the values of the cross validation analysis: at the left $\boldsymbol{R^2}$ and at the right the square root of the MSE. We note that the second one (model without $\boldsymbol{x_4}$) is the best one since it has the biggest $\boldsymbol{R^2}$ and lowest RSME.