Lab 4 – Spring 2020

19th March 2020

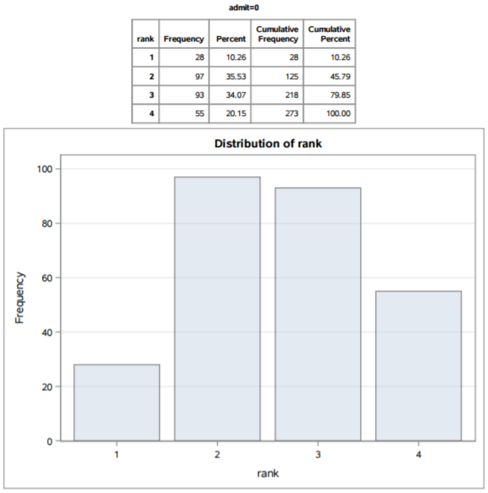
carlo.cavicchia@uniroma1.it

1 Introduction to Logit Model

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable. This data set has a binary response (outcome, dependent) variable called **admit**. There are three predictor variables: **gre**, **gpa** and **rank**. We will treat the variables gre and gpa as continuous. The variable rank takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest.
Use the data set Admit.

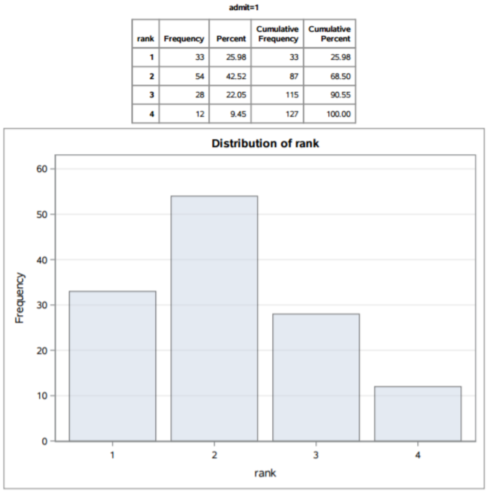
1.1 Descriptive Analysis

Let's start with the descriptive analysis.



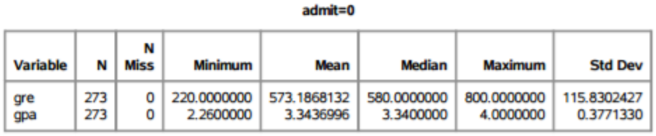
admit=0

rank	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	28	10.26	28	10.26
2	97	35.53	125	45.79
3	93	34.07	218	79.85
4	55	20.15	273	100.00



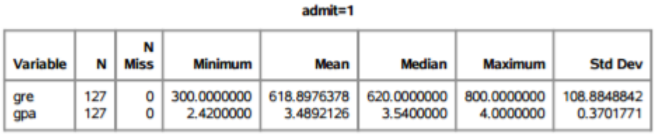
admit=1

rank	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	33	25.98	33	25.98
2	54	42.52	87	68.50
3	28	22.05	115	90.55
4	12	9.45	127	100.00



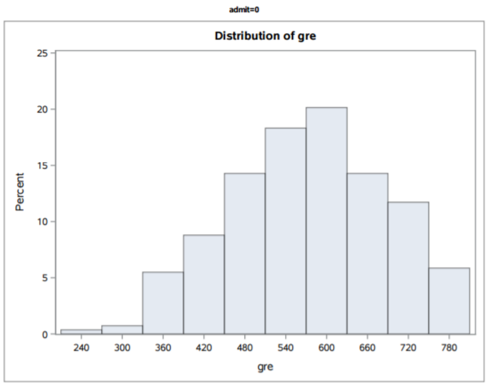
admit=0

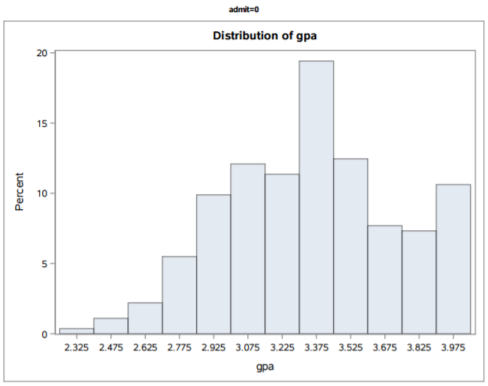
Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
gre	273	0	220.0000000	573.1868132	580.0000000	800.0000000	115.8302427
gpa	273	0	2.2600000	3.3436996	3.3400000	4.0000000	0.3771330

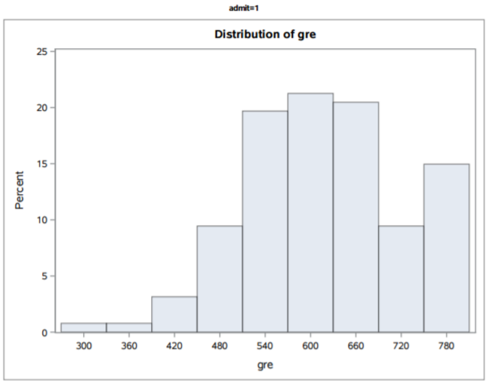


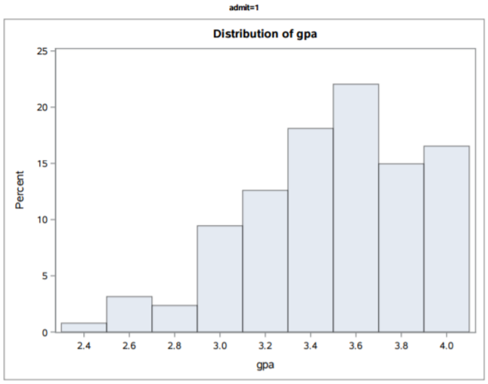
admit=1

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
gre	127	0	300.0000000	618.8976378	620.0000000	800.0000000	108.8848842
gpa	127	0	2.4200000	3.4892126	3.5400000	4.0000000	0.3701771







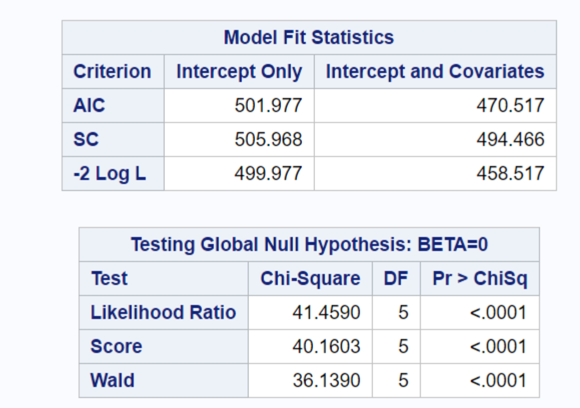


The distributions Rank over the admit values are quite different. When admit takes value 0, most institutions have rank 2 or 4; on the other hand when admit takes value 1, institutions with rank 4 decrease and those with rank 1 increase. The distributions of gpa over the two groups (admit=0 and admit=1) is quite similar. The distribution of gre has higher mean and median when admit takes value 0, compared to the other group (admit=1). Looking at the histograms, the only distribution that result to be more symmetric is gpa when admit takes value 0.

1.2 Fitting a Logistic Model

In this section we fit a logistic model using all predictors. The output is given below.

- Overall significance test:

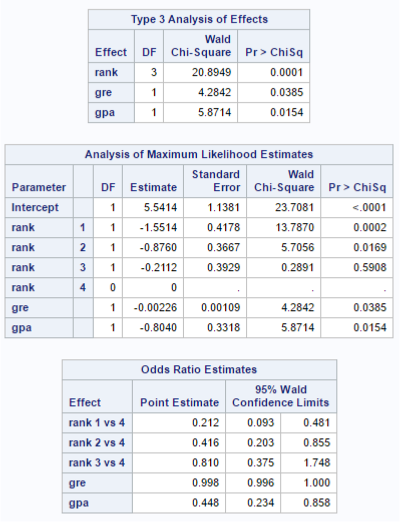


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	501.977	470.517
SC	505.968	494.466
-2 Log L	499.977	458.517

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	41.4590	5	<.0001
Score	40.1603	5	<.0001
Wald	36.1390	5	<.0001

First, we note that the p-values are less than $\alpha = 0.05$, so we reject the null hypothesis. There is at least one predictor that is significant.

- Interpretation of coefficients



Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
rank	3	20.8949	0.0001
gre	1	4.2842	0.0385
gpa	1	5.8714	0.0154

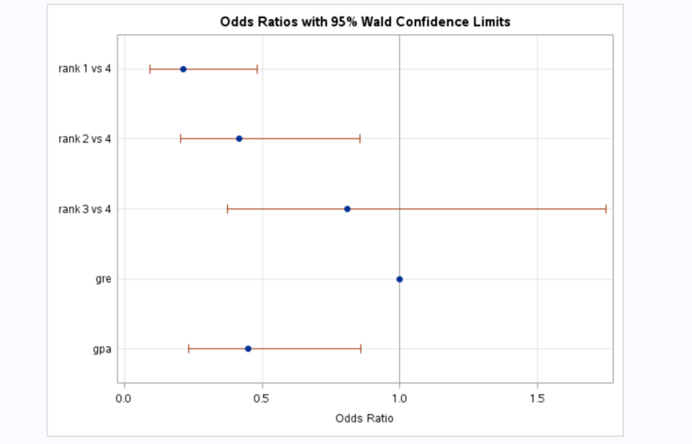
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	5.5414	1.1381	23.7081	<.0001
rank	1	1	-1.5514	0.4178	13.7870	0.0002
rank	2	1	-0.8760	0.3667	5.7056	0.0169
rank	3	1	-0.2112	0.3929	0.2891	0.5908
rank	4	0	0	.	.	.
gre		1	-0.00226	0.00109	4.2842	0.0385
gpa		1	-0.8040	0.3318	5.8714	0.0154

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
rank 1 vs 4	0.212	0.093	0.481
rank 2 vs 4	0.416	0.203	0.855
rank 3 vs 4	0.810	0.375	1.748
gre	0.998	0.996	1.000
gpa	0.448	0.234	0.858

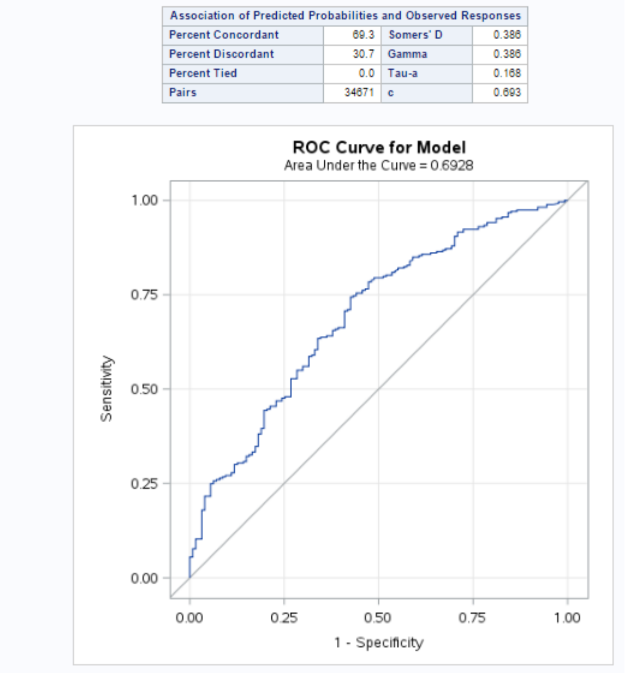
Both gre and gpa are statistically significant, as are the three terms for rank. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable. For every one unit change in gre, the log odds of non-admission (versus admission) decreases by 0.002. For a one unit increase in gpa, the log odds of not being admitted to graduate school decreases by 0.804. The indicator variables for rank have a slightly different interpretation. For example, having attended an undergraduate institution with rank of 1, versus an institution with a rank of 4, changes the log odds of admission by -1.5514.

- all the CIs do not include 1 (note that the CI for gre has 1 as upper limit, indeed the p-value corresponding to its estimate is 0.0385, that is almost 0.05).
- Looking at the Odds Ratio Estimates, now we can say that for a one unit increase in gpa, the odds of not being admitted to graduate school (versus being admitted)

increase by a factor of 0.45. For a one unit increase in gre, the odds of not being admitted to graduate school (versus being admitted) increase by a factor of 0.998. The odds of not being admitted to graduate school (versus being admitted) having attended an undergraduate institution with rank of 1 (2 or 3), versus an institution with a rank of 4, increase by a factor of 0.212 (0.416 or 0.810). This is depicted in the plot below.

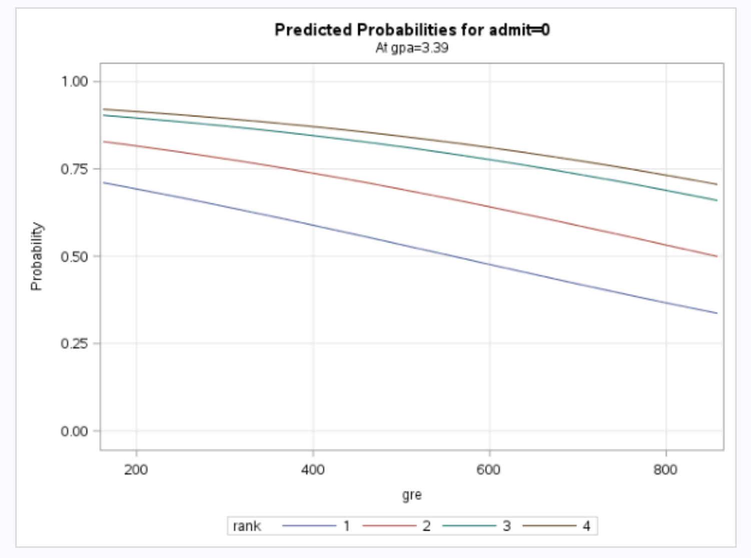


- Model Performance: Looking at the ROC curve and the AUC, 0.69, we deduce that the classification performance is quite poor. Further variables (and thus information), are needed to boost the classification performance.



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	69.3	Somers' D	0.386
Percent Discordant	30.7	Gamma	0.386
Percent Tied	0.0	Tau-a	0.168
Pairs	34671	c	0.693

- Prediction



Fixing the gpa at 3.39 is is possible to note how the probability of being not admitted decreases with gre for every rank (although the gap between curves depends on the specific gre value).

2 Assignment

A bank wants to build a credit scoring model for its customers by using the following variables:

- **Default**: specifies if the client is in default (1) or not (0);
- **BAccount**: a factor with levels no ,Äì good running ,Äì bad running, quality of the credit clients bank account;
- **Months**: duration of loan in months;
- **Past**: a factor with levels bad payer ,Äì good payer if the client previously have been a bad or good payer;
- **Use**: a factor with levels private - professional, the use to which the loan is made;
- **DM**: the size of loan in DM;
- **Gender**: a factor with levels M - F, sex of the client;
- **Status**: a factor with levels no single ,Äì single, status of the client.

Work on yourself on the following tasks using the data set Bank:

- Conduct an exploratory data analysis
- Fit a logistic model
- Draw some conclusions based on the output of your fitted model