Statistical Analysis of Student Exam Scores and the Influence of Different Factors

Final Report

Ibrahim Weaver

Department of Statistics

Florida State University

Tallahassee, FL 32306

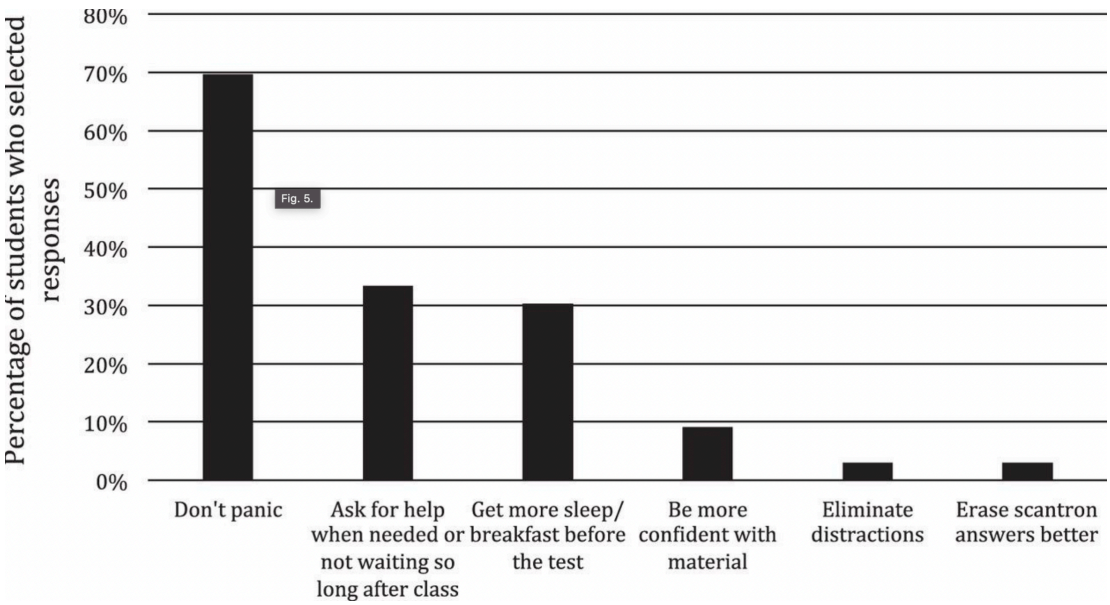
Table of Contents

Introduction

Our goal here is to determine whether or not the students exam scores are influenced by any of the factors in our dataset. These include gender, parents education, if they took a test prep course and so on and so forth. Of course some factors might be more influential than others, and we want to find that out as well. Factors that can affect a student's grades can include their level of engagement with the course material, their study habits, the support they receive from their teachers and parents, and any external factors that may be impacting their ability to focus and learn. By understanding the factors that are contributing to a student's grades and exam scores, educators and parents can work together to provide the necessary support and interventions to help improve their academic performance.

The importance of exam score analysis should not be overlooked, as it is crucial to the next generations future. Students should be given all the situational advantages they can get. We've all been in middle school and high school where we've had to take all different types of standardized tests, and so we should understand the effect these tests can have on students. So students should be able to take tests in an optimal testing environment, where there are so many things that could have an affect on how a student performs, such as the classroom temperature, the comfortableness of the tables or chairs, whether or not the tests are digital

or paper, and so many other factors that one may not initially think of! One important example that I mentioned in my presentation was of a study that advocates for the positive effect of student exam analysis that was done by the American Physiological Society*, and how it marks “don’t panic” as the top answer picked by students as to answer what would’ve helped them perform better on their exam.



As you can see panic beats out other answers by a landslide, coming in over other more obvious answers. Now even panic itself is something that can be affected by all these different factors, just to understand how many levels there can be when dealing with a study such as this, and recognize its importance. Exams are often used to evaluate a student's knowledge and understanding of a subject. By doing well on exams, students are able to demonstrate their mastery of the material and can earn good grades,

which can be important for their future academic and professional success. Additionally, doing well on exams can boost a student's confidence and motivation, and can help them to develop a positive attitude towards learning. Furthermore, helping students do better on their exams can also benefit teachers and schools by improving overall academic performance and contributing to a positive learning environment. Overall, there are many reasons why it is important to help students do better on their exams, and to study and analyze all the different factors that play a part in said exams.

For my project I will be using logistic regression, as my output is binary (Pass/fail). Logistic regression is a widely used and effective tool for predicting a binary response. This is because it is a simple and flexible approach that can be easily applied to many different types of data. Unlike linear regression, which is used for predicting continuous numeric values, logistic regression is specifically designed to predict a binary outcome, such as whether an email is spam or not spam, or whether a customer will buy or not buy, or in this case whether or not a student gets a passing grade on his exams.

One of the key advantages of logistic regression is that it provides a probabilistic framework for making predictions. This means that, rather than simply predicting a binary outcome, logistic regression can also estimate the probability that a given input belongs to each of the two possible classes. This can be useful for making decisions and assessing the uncertainty of a prediction.

The data set I am using contains 1000 observations, and 8 or so variables. I will be testing the variables to see the influence they each have on a student passing, and whether they are statistically significant or not. Using the Logistic Model, we should be able to predict the probability of a student passing based on the values of the different independent variables. One benefit we have to using logistic regression is that we make a lot less assumptions than we would with linear regression. So we don't need to run tests to check if our errors are normally distributed, or for a linear relationship between the dependent and independent variables, along with many other assumptions. By applying logistic regression to this dataset, we hope to gain insights that can be used to improve decision-making and to better understand the factors that affect the response variable (A student passing/failing).

Summary

The data set chosen is public data from a random selection of 1000 students, which I got from Kaggle datasets. I will be running logistic regression using SAS code, and using the results to determine the influence of the different factors on the response. My model will be explained along with the code I used and the important results I get from running it.

The response variable I'm using is Status, which Is either a 1 or a 0, 1 indicating a pass, and 0 indicating a fail. This response variable didn't come with the original data set, it was added by me. However, it is based on the average of other variables within the data set. The code used in SAS for the logistics regression is proc logistic, and the logistic model is used to determine the probability based on the intercept and coefficients we find in the results after running the program. I will use my output from SAS to determine the effect each factor has on student exam scores, and whether they are statistically significant.

Data Description

For the data itself, it is from Kaggle, under the title “Students Performance in Exams”. It contains 1000 observations of random selected students information such as gender, ethnicity, and parental level of education, along with grade scores in three different categories; math score, reading score, and writing score. The website provides a clear and concise overview of the data, allowing researchers and other interested parties to understand the content and structure of the dataset, as well as to identify any potential issues or concerns.

A gender		A race/ethnicity	
Gender		Race/Ethnicity	
male	52%	group C	32%
female	48%	group D	26%
		Other (415)	42%

parental level of e...	
Parental Level	
some college	22%
associate's degree	20%
Other (575)	57%

lunch	
Lunch	
standard	65%
free/reduced	35%

test preparation c...	
Test Preparation	
none	67%
completed	34%

math score	
Math Score	
13	100

reading score	
Reading Score	
27	100

writing score	
Writing Score	
23	100

The previous page shows how the variables and their distribution are represented on Kaggle. You can see how for the score variables the data is normally distributed for the most part. Gender seems to be evenly divided with a 52% 48% split. Race/Ethnicity in this data set are not actually named their real names, but instead called “group A” or “group B” all the way to “Group E”. For parental level of education it tells u whether the parent graduated high school or with an associates degree, along with some other levels of education. The variable lunch tells us whether the student is on a lunch plan at their school - if they are on the standard plan or if they are on the free/reduced plan. Then we have the test preparation course variable, which tells us if the student has completed a test prep course. As we can see almost 70 percent of the students have not taken the course. Our last 3 variables are the test results, they show us how a student has done on the three different subjects by giving us a grade out of 100 for each one.

Now as you might’ve noticed the response variable I mentioned previously is not included in these variables, and that is because I have added it after downloading the dataset. The original data and the new data are as follows on the next page:

Original data set

exams

gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
male	group A	high school	standard	completed	67	67	63
female	group D	some high school	free/reduced	none	40	59	55
male	group E	some college	free/reduced	none	59	60	50
male	group B	high school	standard	none	77	78	68
male	group E	associate's degree	standard	completed	78	73	68
female	group D	high school	standard	none	63	77	76
female	group A	bachelor's degree	standard	none	62	59	63
male	group E	some college	standard	completed	93	88	84
male	group D	high school	standard	none	63	56	65
male	group C	some college	free/reduced	none	47	42	45

Updated data set

examsupdated2

Sex	Race	ParentsEdu	lunch	TestPreptext	mathScore	readingScore	writingScore	total	status	gender	testprep
male	group A	high school	standard	completed	67	67	63	197	1	1	1
female	group D	some high school	free/reduced	none	40	59	55	154	0	0	0
male	group E	some college	free/reduced	none	59	60	50	169	0	1	0
male	group B	high school	standard	none	77	78	68	223	1	1	0
male	group E	associate's degree	standard	completed	78	73	68	219	1	1	1
female	group D	high school	standard	none	63	77	76	216	1	0	0
female	group A	bachelor's degree	standard	none	62	59	63	184	1	0	0
male	group E	some college	standard	completed	93	88	84	265	1	1	1
male	group D	high school	standard	none	63	56	65	184	1	1	0

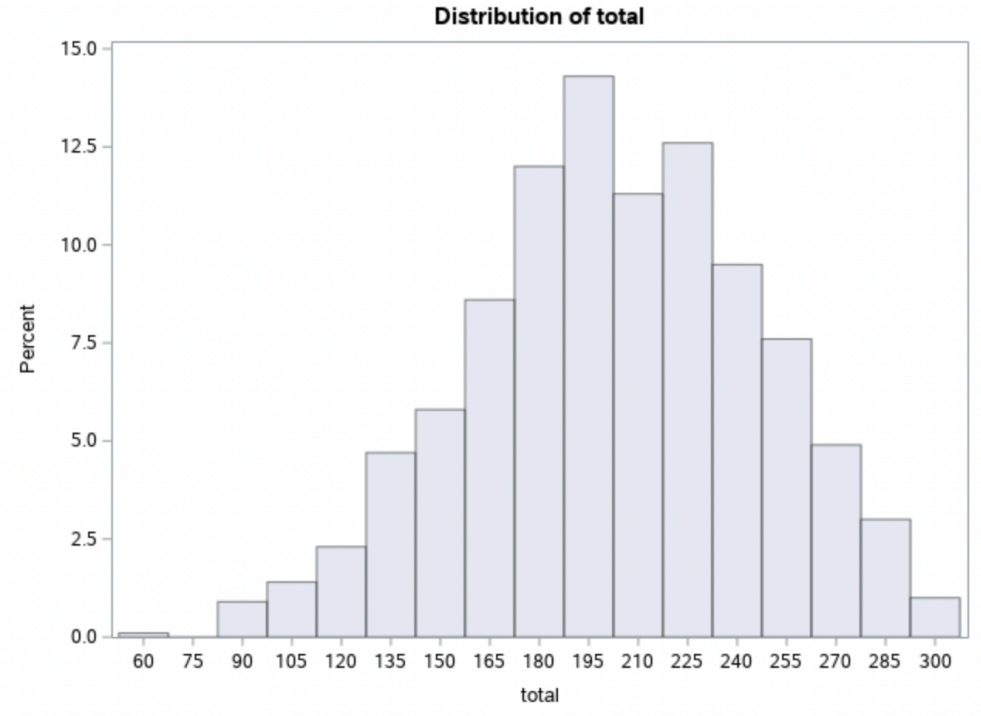
On the previous page, I showed a screenshot of the original data set and then one after I updated it. To update it and add a response variable I used Apples “Numbers” program (similar to excel), and created a new variable “total". This variable was just the total of the 3 scores, math score + writing score + reading score. This however was not the response variable! After adding total I added another variable “status”, which was a conditional binary variable. If total for that observation was over 180, then status would equal 1, indicating the student has passed. If it was under 180 then status would be 0, indicating the student has failed. This was based on 180 being the total if the student averaged a 60 on all three of his exams.

Now this was chosen by me, there can be many other ways to set the response variable, maybe by conditioning none of the scores to be under 60 at all so if even one was under 60 it would fail, or some other condition along those lines. There are two more variables I added on the end of the dataset, and these are not too important. “Gender” and “testprep” both were just variables I converted to 0s and 1s instead of Male/female or taken test prep/not taken, just to check whether having the variables as numbers helped me out in an error I was getting. The error turned out to be something else relating to the spaces between variables so having gender or test prep as 0 or 1 as opposed to its string value was redundant.

One difference you might have noticed between the data sets are the variable names. The original data had variable names with spaces in them. SAS (and many other programming languages) do not allow variable names with spaces in them. Instead, variable names are typically written in all lowercase letters, with underscores (_) used to separate words within the variable name. For example, instead of using a variable name like "favorite color", you might use a name like "favorite_color" or "favoriteColor". This makes the code easier to read and understand, and reduces the likelihood of errors.
I originally tried to use and work on the data set without changing the names of variables, but this proved unsuccessful, as SAS would just give me an error anytime I used a variable with a space in it. One thing that might be interesting was our distribution of total, which we based our response variable on.

Code used to get histogram:

```
/* Exploring Data */
proc univariate data=WORK.IMPORT2;
    ods select Histogram;
    var total;
    histogram total;
run;
```



This histogram tells us that our data is pretty much normal distributed, and we should have even proportions of students that have passed vs students that have failed, which is good for our tests and analysis. There is an outlier on the left side where one student got a 60, and even though for logistic regression we make the assumption that there is no influential outliers, it works because although its an outlier the response variable is binary and based on being over/under 180 which makes the outlier not influential.

The last thing I did to prepare the data was to split it into training set and testing set, with 70% training and 30% testing. When performing logistic regression, it is common to split the dataset into a training set and a test set. The training set is used to train the logistic regression model, while the test set is used to evaluate the performance of the trained model. There are several benefits to this approach. First, it allows you to assess the performance of your model on data that it has not seen before, which is important for determining how well the model will generalize to new, unseen data. Second, it allows you to optimize the model's performance by adjusting the model's parameters based on the results on the training set, before evaluating its performance on the test set.

Code used for splitting data

```
data temp;
set work.import2;
n=ranuni(8);
proc sort data=temp;
  by n;
  data training testing;
   set temp nobs=nobs;
   if _n_<=.7*nobs then output training;
    else output testing;
   run;

PROC CONTENTS DATA=WORK.TESTING; RUN;
PROC CONTENTS DATA=WORK.TRAINING; RUN;
```

SAS has a pre built splitting data option under “tasks and utilities” in SAS studio, however for some reason it could not recognize my data set, so I had to just create a split. After running this code and then the proc contents lines, if everything went well the “proc contents” for the test set should show 300 observations, and the same function for the train set should show 700 (a 70/30 split from 1000 observations):

Observations	300

- 300 from the test set confirmed

Observations	700

- 700 from training set confirmed

Code

```
FILENAME REFFILE '/home/u59231693/sasuser.v94/examsupdated2.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=WORK.IMPORT2;
    GETNAMES=YES;
RUN;


/*status 1 = pass 0 = fail, gender 1= male 0=female,
testprep 1 = completed 0 = none*/

/*testing correlation between independant variables*/

proc freq data=work.import2 order=formatted;

tables testprep*lunch / chisq;

run;

/* Exploring Data */
proc univariate data=WORK.IMPORT2;
    ods select Histogram;
    var total;
    histogram  total;
run;


/*Splitting data into training and test data */
```

```
ods graphics / imagemap=on;
data temp;
set work.import2;
n=ranuni(8);
proc sort data=temp;
  by n;
  data training testing;
   set temp nobs=nobs;
   if _n_<=.7*nobs then output training;
    else output testing;
   run;

PROC CONTENTS DATA=WORK.TESTING; RUN;
PROC CONTENTS DATA=WORK.TRAINING; RUN;



proc logistic data=WORK.TRAINING;
    class Race ParentsEdu lunch / param=glm;
    model status(event='1')=Race ParentsEdu gender testprep lunch / link=logit
        selection=backward slstay=0.05 hierarchy=single details technique=fisher;
run;


proc logistic data=WORK.TRAINING;
    class Race ParentsEdu lunch testprep / param=glm;
    model status(event='1')=Race ParentsEdu testprep lunch ;
run;

```

Results

Now that my data is all set up and my variables fixed, I'm able to start working on results. The first thing I do after making sure my test and train set are working properly is run my data set through a backwards selection elimination process. Backwards selection is a method for selecting the predictors (also known as independent variables or features or variables or even factors in this case) to include in a logistic regression model. It is a form of model selection, which is the process of choosing the appropriate set of predictors to use in a model.

In backwards selection, the process starts by fitting a logistic regression model with all the available predictors. Then, the least significant predictor is removed from the model, and the model is refit. This process is repeated until all the remaining predictors are significant at a predetermined level (e.g. $p < 0.05$).

```
proc logistic data=WORK.TRAINING;
    class Race ParentsEdu lunch / param=glm;
    model status(event='1')=Race ParentsEdu gender testprep lunch / link=logit
        selection=backward slstay=0.05 hierarchy=single details technique=fisher;
run;
```

This is the code we use to run the backwards selection, we've added all the variables as should be done in this case (so you can start with the full model), and as u could see the significance level we've chosen is 0.05, which is what is normally chosen in most cases. However this can change depending on the study.

Analysis of Effects Eligible for Removal			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Race	4	18.6104	0.0009
ParentsEdu	5	19.3420	0.0017
gender	1	2.4041	0.1210
testprep	1	25.5135	<.0001
lunch	1	45.9702	<.0001

Note: No (additional) effects met the 0.05 significance level for removal from the model.

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	gender	1	4	2.4041	0.1210

The output you get from running the backwards selection proc logistic actually outputs a lot of information that can be a little confusing at times. However, what is shown above is pretty much the only thing that is important to us at this step.

In the first table titled “Analysis of Effects Eligible for Removal” we are given all our variables along with p values. The only P-value above 0.05 is for “gender”. This tells us that gender is not statistically significant on the response variable (does not really effect a student passing or failing) based on our significance level of 0.05. However it is the only factor with a p-value above that number, so gender is the only variable removed. Now we want to re run it with our completed model, so we remove gender from the code and run it again.

```
55 proc logistic data=WORK.TRAINING;
56     class Race ParentsEdu lunch testprep / param=glm;
57     model status(event='1')=Race ParentsEdu testprep lunch ;
58 run;
```

We get the following from the results:

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Race	4	18.0807	0.0012
ParentsEdu	5	18.5870	0.0023
testprep	1	25.7921	<.0001
lunch	1	45.9573	<.0001

This tells us the model works and that all the variables are statistically significant, with testprep and lunch being more significant than race and parentsEdu.

After that runs and works, we use our test set to run it again and evaluate our model. Starting with our backwards elimination we get the same result and remove gender as a variable. Now running the full model on the test set we get:

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	370.197	321.080
SC	373.901	365.526
-2 Log L	368.197	297.080

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	71.1171	11	<.0001
Score	65.6137	11	<.0001
Wald	52.1508	11	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Race	4	10.5958	0.0315
ParentsEdu	5	21.7850	0.0006
testprep	1	9.1740	0.0025
lunch	1	23.4137	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.3835	0.5333	6.7292	0.0095
Race	group A	1	0.5096	0.9137	0.3110	0.5770
Race	group B	1	-0.6168	0.5112	1.4562	0.2275
Race	group C	1	-0.8378	0.4626	3.2794	0.0702
Race	group D	1	0.2102	0.5009	0.1761	0.6747
Race	group E	0	0	.	.	.
ParentsEdu	associate's degree	1	1.6207	0.4696	11.9102	0.0006
ParentsEdu	bachelor's degree	1	1.7777	0.6092	8.5142	0.0035
ParentsEdu	high school	1	1.5419	0.4475	11.8702	0.0006
ParentsEdu	master's degree	1	2.0502	0.6713	9.3268	0.0023
ParentsEdu	some college	1	1.0739	0.4074	6.9496	0.0084
ParentsEdu	some high school	0	0	.	.	.
testprep	0	1	-1.0409	0.3437	9.1740	0.0025
testprep	1	0	0	.	.	.
lunch	free/reduced	1	-1.4487	0.2994	23.4137	<.0001
lunch	standard	0	0	.	.	.

From the Type 3 Analysis of Effects table we can see that again all the variables that we kept after backwards elimination are statistically significant. However, we do see some difference from our training set, as in this one “lunch” is showing as the most statistically significant, followed by parents education, then testprep, and finally race. Race being very close to 0.05 (our significance level) and seeming to be a lot less significant than our other variables, although still statistically significant.

It was surprising to me at first that lunch plan status seemed to be our biggest variable effecting whether or not the student graduated. However, after doing some research, I realized that socioeconomic status of a students family can have a significant effect statistically on how students do on standardized tests, as mentioned in this study from West Kentucky University: https://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=1032&context=csa_fac_pub.

After the analysis as we see what can be considered as important factors in students exam performances, it should be noted that this model can be used for any data set regarding exam scores, with any factors to check which can be significant, which in turn can help us increase the average scores of students by using our data to make changes.

Model

Logistic model: ln(P/1-P) = B0+B1X1+B2X2
As you can see its very similar to a linear regression model

Odds: P/1-P = e^(B0+B1x1+B2X2)

Probability: P=e^(B0+B1X1+B2X2)/(1+e^(B0+B1X1+B2X2)

These values are given to us by SAS:
In the table to the right we have all the values to fit our model, the intercept which is our B0, and our coefficients for our response variables
B1, and B2.

Analysis of Maximum Likeliho			
Parameter		DF	Estimate
Intercept		1	1.3835
Race	group A	1	0.5096
Race	group B	1	-0.6168
Race	group C	1	-0.8378
Race	group D	1	0.2102
Race	group E	0	0
ParentsEdu	associate's degree	1	1.6207
ParentsEdu	bachelor's degree	1	1.7777
ParentsEdu	high school	1	1.5419
ParentsEdu	master's degree	1	2.0502
ParentsEdu	some college	1	1.0739
ParentsEdu	some high school	0	0
testprep	0	1	-1.0409
testprep	1	0	0
lunch	free/reduced	1	-1.4487
lunch	standard	0	0

So our Odds formula becomes:
P/1-P=e^(0.6967-0.2149*gender+0.9622*testprep)

All the different variables can be added to the formula and using that we can predict the probability of a student passing based on any of the features we have.

References

https://journals.physiology.org/doi/full/10.1152/advan.00060.2016

https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams

https://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=1032&context=csa_fac_pub