Predictive Analytics
Course Project – Cracking the Customs Officer Admissions Test
Rodrigo Henrique Correa

Summary

Analyzing a test that grants you a USD 384,000/year salary (converted with purchase parity prices, OECD, 2016 values) as a Customs Officer for the Brazilian Internal Revenue Service. The objective is to predict topics that will be present in the next test using cross section data. Also, to determine score ranges that indicate that the endeavor will be successful.

Public careers in Brazil is are very coveted. Salaries are more than 30% above average and offer a steady life.

When a friend of mine comes and tells me that he is studying for the position of Customs Official, one with the fiercest competition among public sector openings, I tell him it is a life commitment for a couple of years. People study obsessively for the topics, hours, weekends, years on end. Some people save money for a couple of years just to stay unemployed and study.

The selection process for all public careers in Brazil is through objective testing, using a constitutional principle of impersonality. For this selection, all one needs are to detain a Bachelor's Degree in any area and pass the test.

The possible topics list is gargantuan, and many people get lost in the quantity of covered topics.

The test

The test comprises in 3 parts:

Part 1: basic knowledge tests, with Portuguese and Foreign Language tests (either English or Spanish), Constitutional Law (Brazilian Constitution is a 100+ pages book), Math and Statistics, Logic and specific legislation.

Part 2: specific knowledge tests, with Financial Accounting, International Trade and Customs Regulation, Tax Law, Tax Regulation and Auditing.

Part 3: written test on legislation and international trade.

Preparatory Courses for tests such as these is a USD 20 Billion industry in Brazil, so it caught my attention. Could I find out something they didn't?

Target of Analysis

Admission Tests for Customs Officer position in Brazil

Description of the context

Provide insights that simplify the path to approval. Identify the observed betas to determine which knowledge areas count the most for approval and which subjects and topics are key.

In this case, what is considered to be empirical is the possible points in the test: 270 overall, while part 1 totals 70 points, part 2 140 points and part 3 totals 60 possible points.

Characteristics

- Data Type: Cross-section data
- Estimation (expected): Least Squares (Gaussian)

Dependent variable: Admissions Test Final Score (Observed)

Data Sets:

Approved in the last selection process, 2014, with individual scores.

The following sample visualizations show the behavior of the data set for the Part 1 of the test for selected score ranges, with possible overall score of 210.

There is also the written test, granting possible 60 points. However, there had been only 3 exams with the written test. In this case, it will not enter in the analysis at this time.

Independent Variable		
Summary of independent variable	**Categorical or quantitative?**	**Argument for / description of the associates with the dependent variable**
Test 1 Score (Observed)	Quantitative	Basic Knowledge Test, contributes empirically for 26% of approval
Test 2 Score (Observed)	Quantitative	Specific Knowledge Test, contributes empirically for 52% of approval
Test 3 Score (Observed)	Quantitative	Written Test, contributes empirically for 22% of approval

Data Analysis Part 1: Passing Scores

In order to verify "what it takes" to pass, the starting point is to verify the characteristics of scores of the ones who did pass.

KNOWLEDGE AREA	POSSIBLE POINTS	% TOTAL	QUESTIONS
ACT - ACCOUNTING	40	19,05%	20
CMI - INTL COMMERCE	30	14,29%	15
TAL - TAX LAW	30	14,29%	15
AUD - AUDITING	20	9,52%	10
TRG - TAX REGULATION	20	9,52%	10
POR - PORTUGESE	20	9,52%	20
AGE - GEN. ADMINISTRATION	10	4,76%	10
ADL - ADMIN LAW	10	4,76%	10
COL - CONSTITUTIONAL LAW	10	4,76%	10
ENG - ENGLISH	10	4,76%	10
RLM - MATH AND LOGIC	10	4,76%	10
TOTAL	210		

Table 1: Knowledge Areas and Possible Points. Source: Receita Federal do Brasil, frequencies made by the author.

Verifying the exam characteristics, it becomes quite clear that Accounting, International Commerce and Tax Law are the most important areas to acquire knowledge in order to pass. From this starting point, it is important to check how the passing applicants scored in each of the knowledge areas.

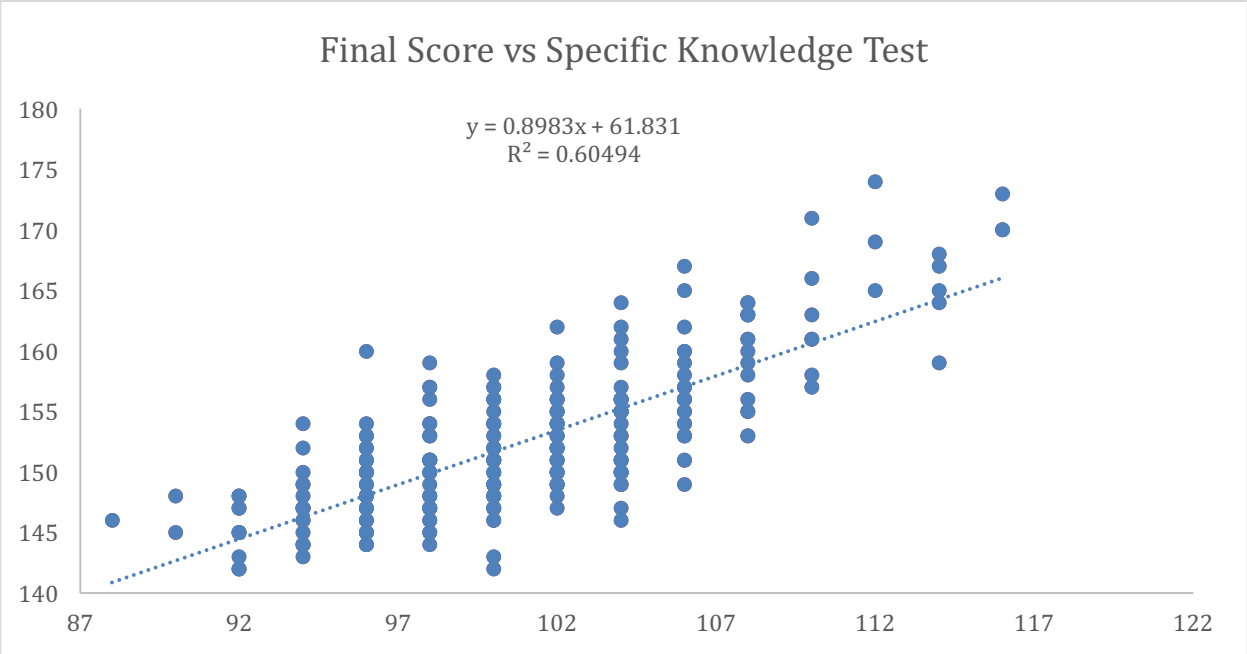
KA	AVG	STD DEVIATION	ACCURACY
ACT	32,26	3,34	80,6%
CMI	21,10	3,07	70,3%
TAL	20,41	2,77	68,0%
AUD	16,74	2,18	83,7%
TRG	10,44	1,87	52,2%
POR	16,76	1,77	83,8%
AGE	8,70	0,97	87,0%
ADL	5,69	1,23	56,9%
COL	6,95	1,22	69,5%
ENG	6,94	1,50	69,4%
RLM	6,53	1,42	65,3%
TOTAL	152,53	AVG ACC	71,5%

Table 2: Knowledge Areas and Actual Scores, 2014. Source: Receita Federal do Brasil, frequencies made by the author.

As another relevant information, the highest score was 174 out of 210; the lowest, 142. From this, it is possible to ascertain that winning strategies focus on high scores in the key knowledge areas, such as Accounting and Auditing. Lower value subjects, such as Math and Logic, English and Constitutional Law show significantly lower accuracy. This

shows that winners, aware of the lower relevance may have focused their efforts in the key areas.

The test, as mentioned before, is divided in two: Basic and Specific Knowledge.



Visualization 1: Specific Knowledge Test Scores vs Final Score, Scatterplot. Source: Receita Federal do Brasil, plot made by author.

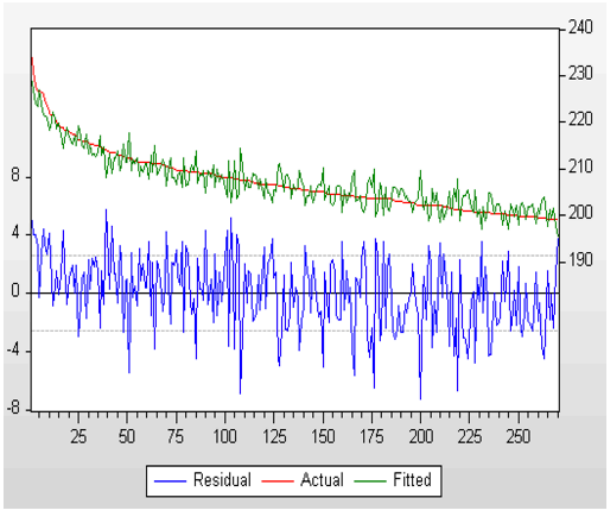
The scatterplot above shows the relationship with the Specific Knowledge test scores with Final Test Scores. There is a significant dispersion between equal final scores. This means that there are a number of winning combinations that are possible in order to obtain the same score.

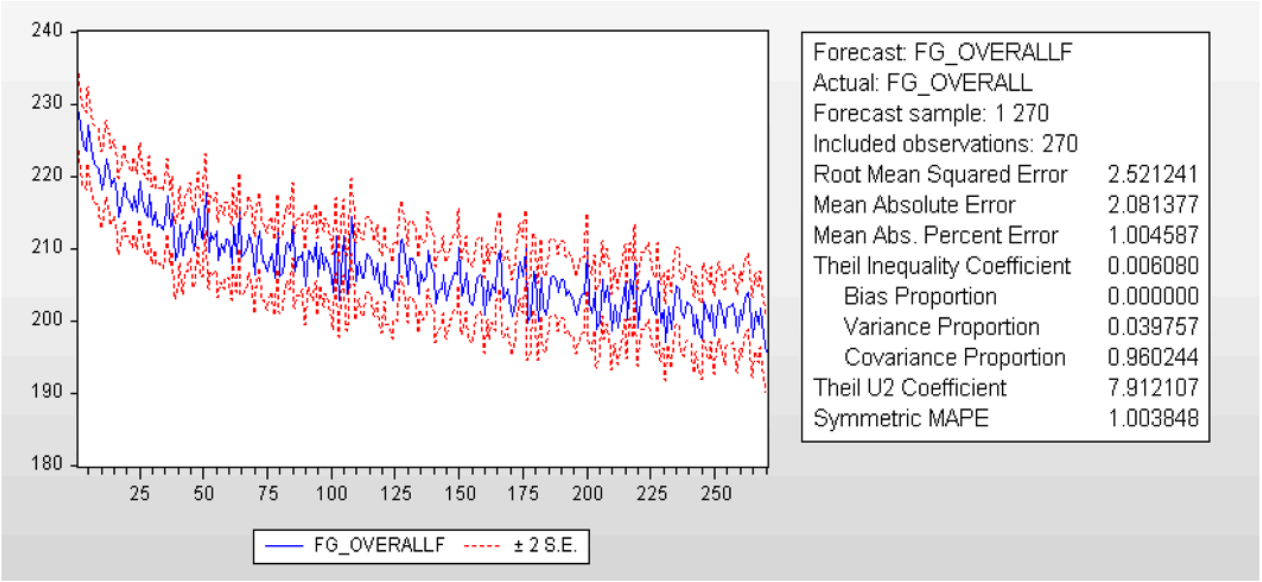
Data Analysis Part 1: Multiple Regression Analysis for Passing Scores.

Dependent Variable: FG_OVERALL
Method: Least Squares
Date: 08/19/17 Time: 16:14
Sample: 1 270
Included observations: 270

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	56.50211	4.032793	14.01067	0.0000
D1_POR	1.118117	0.091015	12.28496	0.0000
D2_ENG	1.156153	0.111752	10.34572	0.0000
D3_RLM	1.119093	0.116486	9.607069	0.0000
D4_AGE	0.837833	0.167028	5.016109	0.0000
D5_COL	1.052082	0.131189	8.019607	0.0000
D6_ADL	1.223628	0.132419	9.240577	0.0000
D7_TAL	0.934420	0.058537	15.96294	0.0000
D8_AUD	0.984724	0.073378	13.41997	0.0000
D9_ACT	0.916941	0.049767	18.42459	0.0000
D10_TRG	1.094305	0.085777	12.75754	0.0000
D11_CIM	0.878141	0.053838	16.31079	0.0000

R-squared	0.852903	Mean dependent var	207.2356
Adjusted R-squared	0.846632	S.D. dependent var	6.585950
S.E. of regression	2.579208	Akaike info criterion	4.776268
Sum squared resid	1716.297	Schwarz criterion	4.936198
Log likelihood	-632.7962	Hannan-Quinn criter.	4.840489
F-statistic	135.9951	Durbin-Watson stat	1.689312
Prob(F-statistic)	0.000000		





Visualization 2: Multiple Regression, Passing Scores, Overall. Residue Analysis. Forecasting.
Made with eviews 9.5, Student Version.

Considering that the final result is completely known in terms of final score (we know what produces a high score in empirical terms), the idea was to analyze the actual results from the successful applicants and compare them with the final grade overall (Objective testing and Written Test).

All p-values shown in eviews as Prod. are zeroed. This means all the variables belong to the model. The coefficients, β, mean the degree of impact of each variable in the model. Shockingly, scores in Accounting did not domain the betas of the regression. One hypothesis is that this subject shows particularly hard questions. Another is that not all scores were maximized to their full potential in order to study and maximize grades,

meaning it would be possible to strategize and save efforts in other peripheral subjects; this makes sense given the fact that a great number of people spend years chasing this goal, studying thoroughly everything that comes in front of them. In any case, this insight will be noted for future reference.

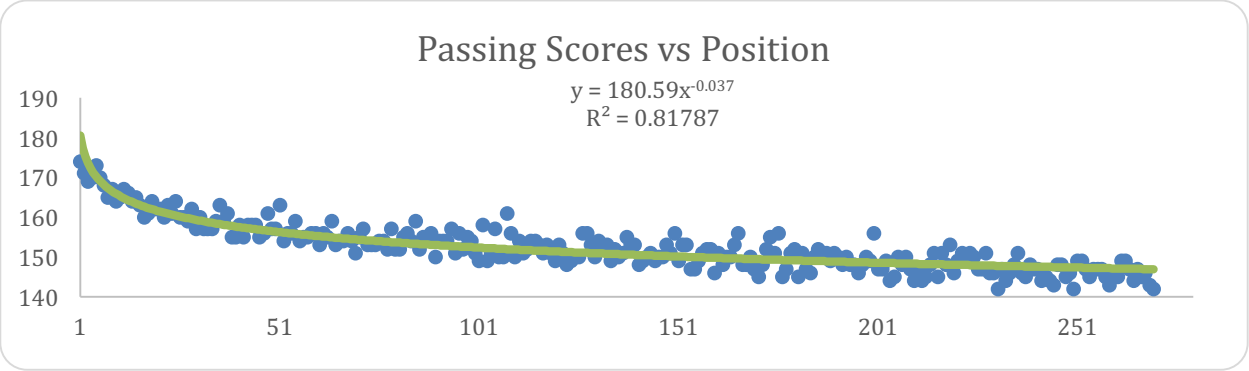
The R-squared of the regression shows a quite acceptable level, which indicates a high degree of comfort making assumptions over the model.

Finally, C or residual error of the regression: since it took into account the final score (Test 1 + Test 2 + Test 3), C represents similar values to possible test 3 scores, creating some prediction space.

For this analysis, since it is a closed model that all variables are actually known, there seems to have no need for testing for omitted variables. This will be done, however, for the second part of the analysis, when historical frequencies of subjects come into play.

Therefore, at this point there had been determined that:

1. There are a number of passing strategies regarding test 1 and 2. Accuracies and influence in scores, however, tend to favor Accounting.
2. The multiple regression shows, however, that high scores in lower value areas, such as English, Portuguese and Math show incremental value and relationship with passing for higher scores. It is probably the “bare minimum” necessary or a good predictor for overall performance.



Visualization 3: Passing Scores vs Position, Scatterplot. Made with Excel for Mac.

Just as a bonus, let’s look at the scatterplot that shows passing scores versus position. This presents the results in Tests 1 and 2 and final score. This answer the question: could someone who did not go so well in the objective tests save their approval through the written test? This question, if answered “yes” would bring little validity to the studied model. However, the answer is hardly ever, since there is a very high correlation with T1+T2 scores with position. There is no data on the dual of it, meaning, someone that maybe had a good T1+T2 grade but didn’t go well on the written test; that would belong to the failed scores, but there is no way to know that.

Data Analysis Part 2: Subjects and Frequencies.

After verifying what is the "secret" to pass by analyzing the passing grades, it is time to analyze the subjects that actually are pertinent to the test.

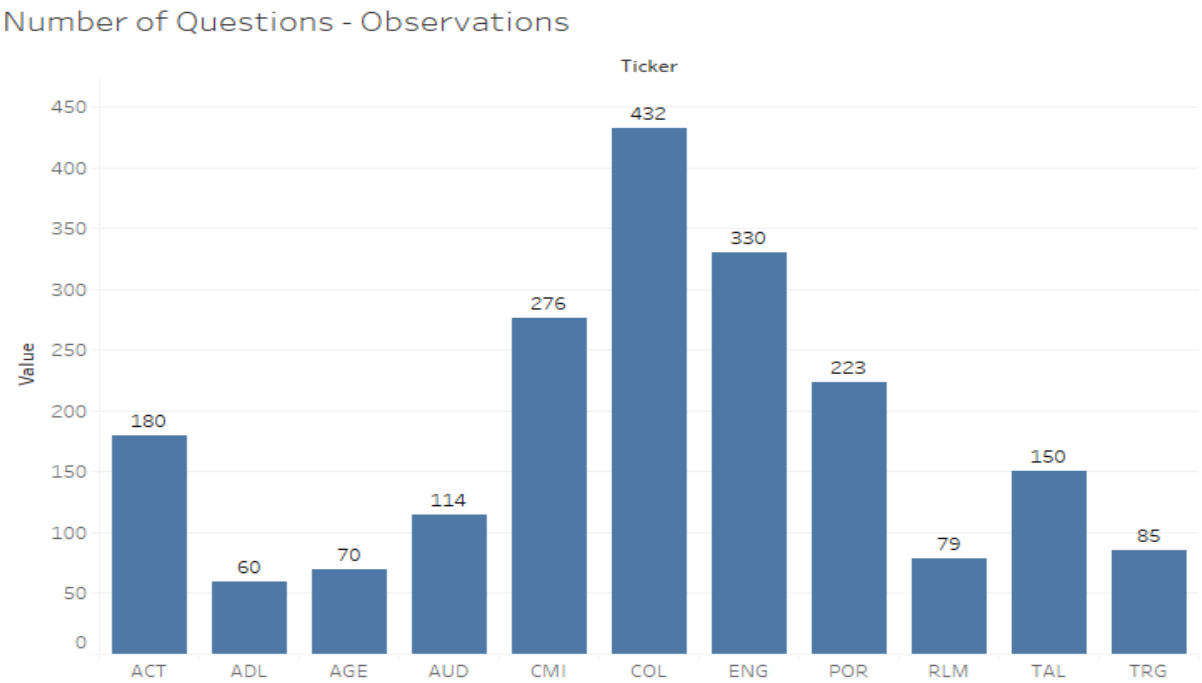
For this, it is necessary to analyze the Syllabus of the test, past tests, similar examinations and then, finally, assign frequencies to all the subjects.

The past tests for Customs Officer was conducted by ESAF (Escola Superior de Administração Fazendária). This is a School designed to provide training to public employees linked to the Brazilian Ministry of Finance.

ESAF conducts most of the public selection processes for auditing areas of the government and for tax officials.

The last public selection involved 268 topics distributed along the 11 Knowledge Areas.

The sample selected comprises of 1.999 questions, separated by Knowledge Area and then further separated by topic. These questions come from past tests made by ESAF from 2004 to 2014, including 3 selections for our target selection process.



Visualization 4: 1999 questions used in public selections by ESAF, 2004-2014. Made with Tableau.

The initial analysis revealed that almost 60 out of the 268 topics listed in the syllabus for study never actually appeared in a test. This is a strategy to give the examiner "carte blanche" to create a difficult question if he sees fit.

Clusterizing: Alphas, Pis and Omegas.

From the fact that many subjects never appeared on the actual selection tests, came the idea of analyzing frequencies in which each subject actually appeared. It means that not only it is possible to refine the strategy by focusing in specific knowledge areas but also specific topics within these knowledge areas.

For that, frequencies were assigned to each topic, according to its n. Then, the highest frequency topic within each Knowledge area was assigned as alpha, the most powerful, as an inspiration in the animal kingdom, the alpha male. The least powerful, omega.

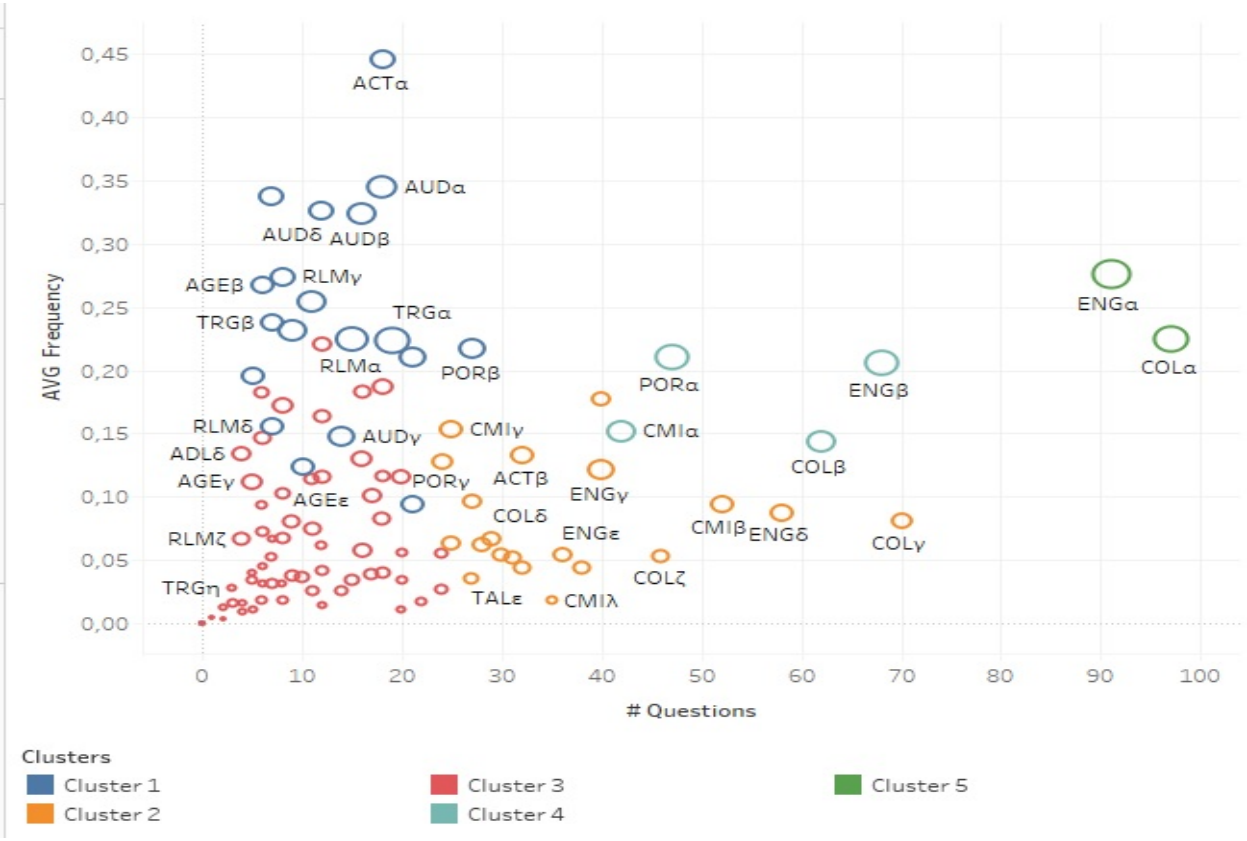
From that classification, 17 classes arose: alpha to pi (alpha, beta, gamma, delta, epsilon, zeta, eta, theta, iota, kappa, lambda, mi, nu, xi, omicron, pi) and then directly to omega, indicating the minimum.

So, the ticker AGE, for General Administration, was assigned as AGEα to its highest frequent topic appearing on tests; AGEω to the least frequent. Equal frequency values were assigned the same Greek letter; i.e., if there are 2 questions in the β frequency range, both of them belong in beta set.

SYMBOL	AGE	AUD	CMI	ACT	ADL	COL	TAL	ENG	TRG	POR	RLM
α	10,00%	15,789%	15,217%	10,000%	15,000%	22,454%	null	27,576%	22,353%	21,076%	18,987%
β	8,57%	14,035%	9,420%	8,889%	11,667%	14,352%	6,000%	20,606%	8,235%	12,108%	13,924%
γ	7,14%	12,281%	9,058%	5,556%	8,333%	10,648%	5,333%	12,121%	7,059%	9,417%	10,127%
δ	5,71%	10,526%	6,159%	4,444%	6,667%	6,713%	4,667%	10,909%	5,882%	7,175%	8,861%
ε	4,29%	8,772%	5,797%	3,333%	5,000%	5,556%	4,000%	7,576%	4,706%	6,278%	6,329%
ζ	2,86%	7,018%	5,435%	2,222%	3,333%	5,324%	3,333%	7,273%	3,529%	5,381%	5,063%
η	1,43%	5,263%	3,986%	null	null	4,398%	2,667%	6,667%	1,176%	4,933%	3,797%
θ	null	3,509%	3,623%	null	null	3,935%	2,000%	3,333%	null	4,036%	null
ι	null	null	2,536%	null	null	3,472%	1,333%	2,121%	null	3,587%	null
κ	null	null	2,174%	null	null	2,778%	0,667%	null	null	3,139%	null
λ	null	null	1,812%	null	null	2,546%	null	null	null	2,691%	null
μ	null	null	1,449%	null	null	2,315%	null	null	null	2,242%	null
ν	null	null	0,725%	null	null	1,389%	null	null	null	1,794%	null
ξ	null	null	null	null	null	1,157%	null	null	null	1,345%	null
ο	null	null	null	null	null	null	null	null	null	0,897%	null
π	null	null	null	null	null	null	null	null	null	0,448%	null
ω	0,00%	1,754%	0,362%	1,111%	1,667%	0,926%	0,000%	1,818%	0,000%	0,000%	0,000%

Table 3: 1999 questions frequency ranges Alpha to Pi and Omega, separated by subject.

Each of the frequencies were then used to classify the sets. The idea here is quite similar to the famous “Moneyball” case; to look for the highest payoff using how frequently the players scored. In our case, how frequently the topics appear.



Visualization 5: Frequency class assigned topics, clusterized. Sizes indicate expected score by topic study. Made with Tableau.

The scatterplot shows how frequent the topics are for study reference. Also, with the frequencies at hand and also the possible score list presented in table 1, it is possible to ascertain individually which topics will provide more incremental scores if thoroughly studied.

Following this idea, which topic class should one study more thoroughly or focus on to guarantee a highest score (frequency multiplied by possible points in the test), I have separated a special analysis for how each set contribute for approval.

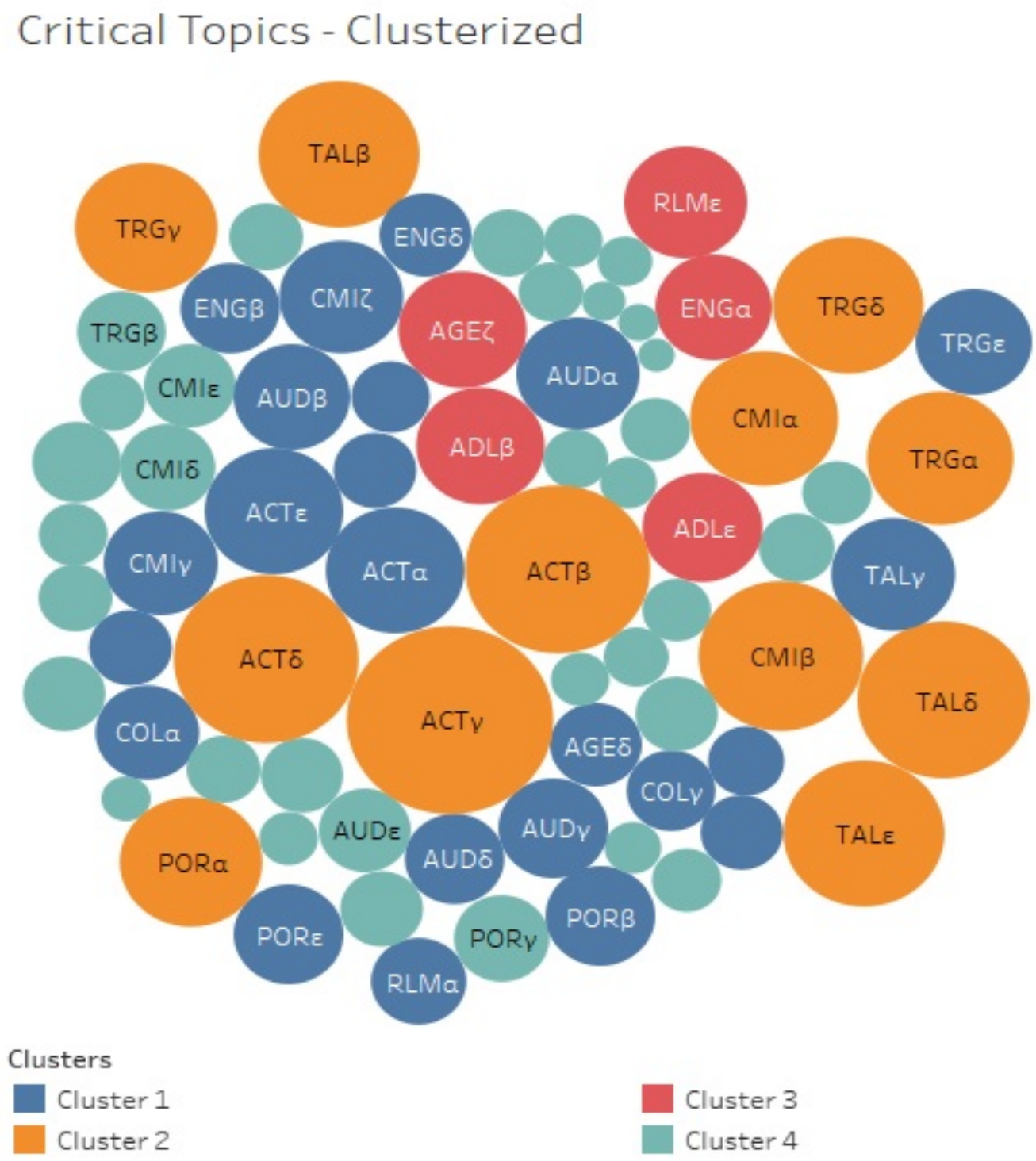
SETS	LETTER	CUMULATIVE POSSIBLE SCORE	CUMULATIVE TOPIC COUNT
α	ALPHA	29,81	10
β	BETA	64,09	27
γ	GAMMA	92,87	45
δ	DELTA	121,97	68
ε	EPSILON	148,74	100
ζ	ZETA	166,06	127
η	ETA	177,82	151
θ	THETA	184,72	161
ι	IOTA	189,12	170
κ	KAPPA	191,68	178
λ	LAMBDA	196,27	187
μ	MI	198,49	193
ν	NU	199,78	198
ξ	XI	200,44	201
ο	OMICRON	200,61	202
π	PI	200,70	203
ω	OMEGA	210,00	268

Table 4: Cumulative Possible Scores and Count by set. Made with Excel for Mac.

The Table 4 shows the number of subjects in each of the Greek letter sets and their contribution for the final score. It is quite clear that after the Eta set, there are increasingly diminishing returns on each topic set studied. The last set, Omega, guarantees possible 9.3 additional points in the test. However, for that it is necessary to study 65 more topics, including the ones that had never appeared in a test before for the last 13 years.

Sorting by sets is already a good strategy. However, subjects are not linear. Test 1 of basic knowledge is weighted as 1; test 2 is weighted 2. So, it might pay better on average to study for a topic in Accounting with a lower frequency than Portuguese or English, for example.

With that in mind, a selection of topics was made in order to determine the highest of the highest payoffs; it is not a strategy that can move one to “laziness” given the competitiveness of the examination. However, it certainly gives someone edge over others who did not see this trend, saving a lot of effort.



Visualization 6: Critical Topics, Clusterized. Sizes indicate expected score by topic study. Made with Tableau.

TOPIC	KA	FREQUENCY	INCREMENTAL SCORE
ACTγ	ACT	22%	8,89
ACTβ	ACT	18%	7,11
ACTδ	ACT	18%	7,11
TALδ	TAL	21%	6,20
CMIβ	CMI	19%	5,65
TALε	TAL	18%	5,40
TALβ	TAL	18%	5,40
TRGδ	TRG	24%	4,71
CMIα	CMI	15%	4,57
TRGα	TRG	22%	4,47
TRGγ	TRG	21%	4,24
PORα	POR	21%	4,22
ACTα	ACT	10%	4,00
ACTε	ACT	10%	4,00
ADLβ	ADL	35%	3,50
CMIζ	CMI	11%	3,26
TALγ	TAL	11%	3,20
AUDα	AUD	16%	3,16
TRGε	TRG	14%	2,82
AUDβ	AUD	14%	2,81

Table 5: Cumulative Possible Scores and Count by Subject, 20 largest. Made with Excel for Mac.

By analyzing the clustering made with Greek letters, we can see that Accounting-gamma set is the best set of subjects to study. Even though they do not present the highest frequency by definition (ACT-alpha) is the highest, the frequency appears in a balanced, consistent way.

Finally we can move on to the regression analysis in order to verify the behavior of the sets.

For this regression, the incremental score was considered by studying each set individually. Also, Ramsey RESET tests were performed to determine the best fit for the number of variables in the regression.

Multiple Regression: Topic Sets.

Dependent Variable: TOTAL
Method: Least Squares
Date: 08/20/17 Time: 19:42
Sample: 1 32
Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.084858	0.235193	0.360803	0.7215
ALPHA	0.993936	0.125508	7.919325	0.0000
BETA	0.792178	0.178756	4.431621	0.0002
GAMA	1.601432	0.238406	6.717253	0.0000
DELTA	1.107694	0.170281	6.505101	0.0000
EPSILON	1.046465	0.217138	4.819366	0.0001
ZETA	1.112809	0.236073	4.713824	0.0001
ETA	0.539137	0.291841	1.847366	0.0776
THETA	2.857099	0.551089	5.184459	0.0000

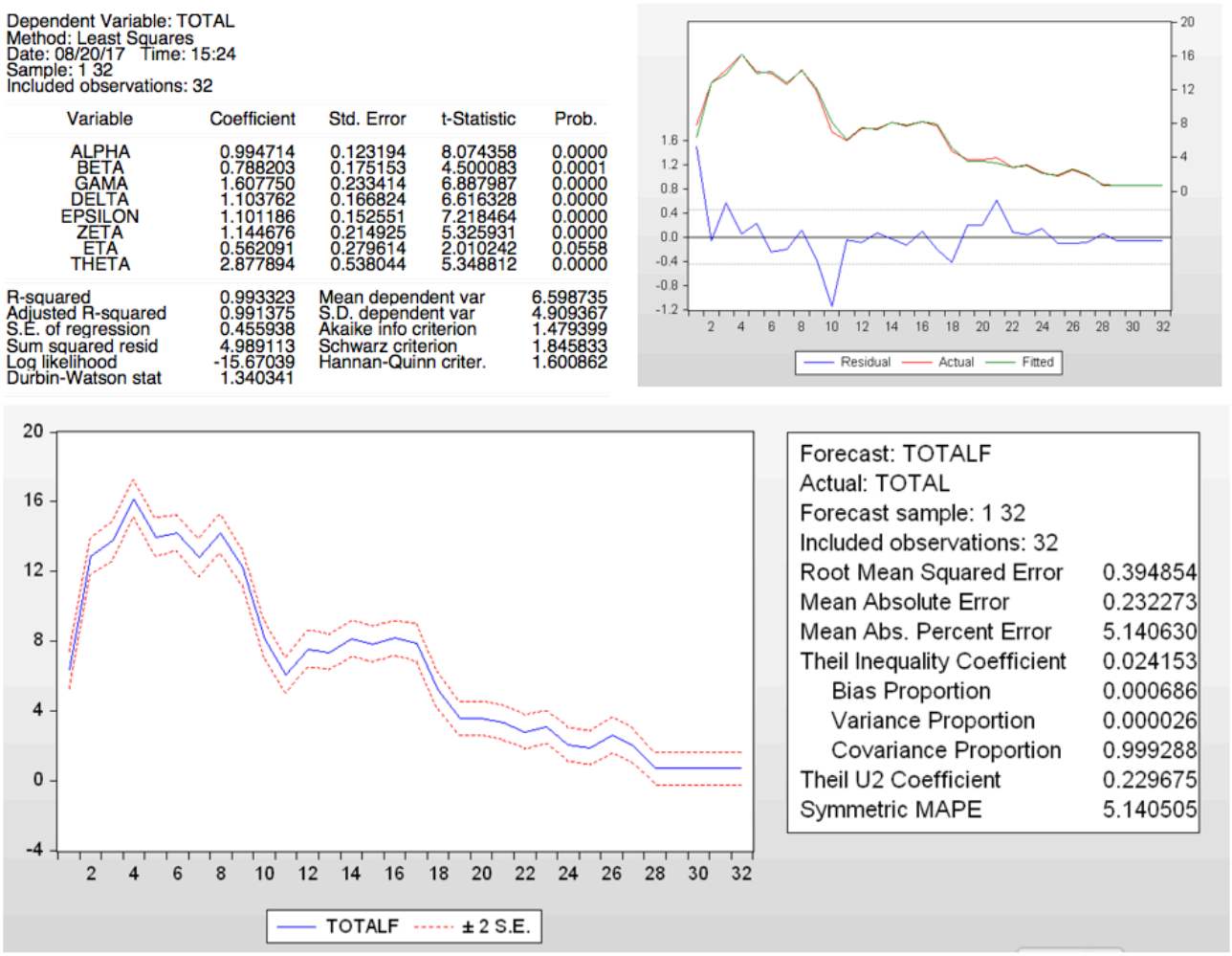
R-squared	0.993360	Mean dependent var	6.598735
Adjusted R-squared	0.991051	S.D. dependent var	4.909367
S.E. of regression	0.464432	Akaike info criterion	1.536255
Sum squared resid	4.961034	Schwarz criterion	1.948494
Log likelihood	-15.58009	Hannan-Quinn criter.	1.672901
F-statistic	430.1154	Durbin-Watson stat	1.358933
Prob(F-statistic)	0.000000		

Visualization 7: Greek letter sets Regression. Made with eviews 9.5, student edition.

This regression shows staggering 0.99 in terms of R-Squared. Once again, p-values are close to zero, Akaike and Schwartz criteria are relatively low, which determines if there are non-tested alternatives. The constant C or error shows a very high p-value, so it shold be therefore discarded.

Note that for this regression it was understood by table 4 that after theta there was little score explanation for the remaining sets, adding noise to the model. In this case, they were removed.

However, in order not to omit the variables and suffer from regrets; after running the model once more, there will be the addition to verify if the omitted variables IOTA to OMEGA are relevant.



Dependent Variable: TOTAL
Method: Least Squares
Date: 08/20/17 Time: 15:24
Sample: 1 32
Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ALPHA	0.994714	0.123194	8.074358	0.0000
BETA	0.788203	0.175153	4.500083	0.0001
GAMA	1.607750	0.233414	6.887987	0.0000
DELTA	1.103762	0.166824	6.616328	0.0000
EPSILON	1.101186	0.152551	7.218464	0.0000
ZETA	1.144676	0.214925	5.325931	0.0000
ETA	0.562091	0.279614	2.010242	0.0558
THETA	2.877894	0.538044	5.348812	0.0000

R-squared	0.993323	Mean dependent var	6.598735
Adjusted R-squared	0.991375	S.D. dependent var	4.909367
S.E. of regression	0.455938	Akaike info criterion	1.479399
Sum squared resid	4.989113	Schwarz criterion	1.845833
Log likelihood	-15.67039	Hannan-Quinn criter.	1.600862
Durbin-Watson stat	1.340341		

Forecast: TOTALF
Actual: TOTAL
Forecast sample: 1 32
Included observations: 32

Root Mean Squared Error	0.394854
Mean Absolute Error	0.232273
Mean Abs. Percent Error	5.140630
Theil Inequality Coefficient	0.024153
Bias Proportion	0.000686
Variance Proportion	0.000026
Covariance Proportion	0.999288
Theil U2 Coefficient	0.229675
Symmetric MAPE	5.140505

Visualization 8: Greek letter sets Regression, revised. Made with eviews 9.5, student edition.

Running the model one more time in eviews, there were similar results. Theta shows a strange coefficient, however. Returning to the data, theta is a set that shows relevant frequencies in Portuguese, Auditing, Tax Law and International Commerce. This means it shows important frequencies for important subjects, even though with little significance overall.

Omitted Variables Test
Null hypothesis: OMEGA are jointly significant
Equation: UNTITLED
Specification: TOTAL ALPHA BETA GAMA DELTA EPSILON ZETA ETA THETA
Omitted Variables: OMEGA

	Value	df	Probability
t-statistic	1.224290	23	0.2332
F-statistic	1.498887	(1, 23)	0.2332
Likelihood ratio	2.020271	1	0.1552

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	0.305243	1	0.305243
Restricted SSR	4.989113	24	0.207880
Unrestricted SSR	4.683870	23	0.203647

LR test summary:

	Value	df
Restricted LogL	-15.67039	24
Unrestricted LogL	-14.66025	23

Unrestricted Test Equation:
Dependent Variable: TOTAL
Method: Least Squares
Date: 08/20/17 Time: 15:22
Sample: 1 32
Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ALPHA	1.023230	0.124138	8.242676	0.0000
BETA	0.830687	0.176799	4.698474	0.0001
GAMA	1.489733	0.250329	5.951098	0.0000
DELTA	1.011143	0.181622	5.567295	0.0000
EPSILON	1.058519	0.154960	6.830927	0.0000
ZETA	0.955408	0.262967	3.633191	0.0014
ETA	0.401102	0.306403	1.309066	0.2034
THETA	2.668135	0.559419	4.769471	0.0001
OMEGA	1.278542	1.044313	1.224290	0.2332

R-squared	0.993731	Mean dependent var	6.598735
Adjusted R-squared	0.991551	S.D. dependent var	4.909367
S.E. of regression	0.451272	Akaike info criterion	1.478766
Sum squared resid	4.683870	Schwarz criterion	1.891004
Log likelihood	-14.66025	Hannan-Quinn criter.	1.615411
Durbin-Watson stat	1.383297		

Visualization 9: Omitted Variables Test, Ramsey RESET. Made with eviews 9.5, student edition.

Redundant Variables Test
Null hypothesis: ETA THETA EPSILON are jointly insignificant
Equation: UNTITLED
Specification: TOTAL ALPHA BETA GAMA DELTA EPSILON ZETA ETA THETA
Redundant Variables: ETA THETA EPSILON

	Value	df	Probability
F-statistic	33.57047	(3, 24)	0.0000
Likelihood ratio	52.73436	3	0.0000

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	20.93586	3	6.978621
Restricted SSR	25.92498	27	0.960184
Unrestricted SSR	4.989113	24	0.207880

LR test summary:

	Value	df
Restricted LogL	-42.03757	27
Unrestricted LogL	-15.67039	24

Restricted Test Equation:
Dependent Variable: TOTAL
Method: Least Squares
Date: 08/20/17 Time: 15:25
Sample: 1 32
Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ALPHA	1.620112	0.157182	10.30726	0.0000
BETA	1.086312	0.357373	3.039718	0.0052
GAMA	1.259756	0.471102	2.674063	0.0126
DELTA	1.761455	0.312558	5.635607	0.0000
ZETA	1.703239	0.385917	4.413488	0.0001

R-squared	0.965302	Mean dependent var	6.598735
Adjusted R-squared	0.960161	S.D. dependent var	4.909367
S.E. of regression	0.979890	Akaike info criterion	2.939848
Sum squared resid	25.92498	Schwarz criterion	3.168869
Log likelihood	-42.03757	Hannan-Quinn criter.	3.015762
Durbin-Watson stat	1.233677		

Visualization 10: Redudant Variables Test, Ramsey RESET. Made with eviews 9.5, student edition.

Both redundant variables test and omitted variables test show that variables shouldn't be included or removed. In this case, the model will be accepted as fit.

Conclusion

Breaking down one of the toughest selection processes in the country. How valuable is that? On the personal side, I believe that it will actually help some people close to me achieving their personal goals.

On the analysis side, it is trying to find out how to succeed by cutting down in half the steps needed to be successful in a complex, time consuming endeavor.

From the regression and the analyses made, it becomes clear that an optimal strategy is focused study in 8 sets (alpha to theta) in order to increase the chances of approval. This represents studying 40% less topics and give emphasis to what really matters to be get there.

This could mean a stable life for some of my closest or even for me, who started thinking this might be a card to hold for a rainy day or even to be applied in things that catch my attention a little more: for example, diplomatic career (which I had tried in the past and got really close). Things seem to get easier when organized and seen through statistics.

There are many ways the model could be improved, for example, weighting difficulty level of each question and adding study hours necessary to master each subject. This might be an interesting extension for this study.

All in all, should this strategy prove successful in practice, at least I'll get a couple of drinks paid.