INTRODUCTION

Taxes often have a negative connotation to people, but despite the frustration taxation is one of the essential aspects of a developed society. Tax revenues allow governments to function, funding social programs and public investments. Taxes are the price we pay for a stable, more equitable society. With this project we analyze in detail one of the largest tax costs for businesses, the corporate income tax cost (or Total Tax Rate), on a global level. The data were obtained from the World Bank's Doing Business 2015 report[1]. Our aim is to come to an understanding of what explains this cost best and to what extent the tax cost is related to the Statutory Tax Rate, geographical region, and other features of the country. In doing so, we have chosen to approach the research question from two separate angles: multiple regression analysis and ANOVA. In both analyses the response variable is Total Tax Rate, with ANOVA doing additional testing for response variable Statutory Tax Rate. The cross-sectional observational design of this study does not allow us to make generalizations to either past or future tax rates, or to other taxes. The continuous variables of interest in this project are:

- **Statutory Tax Rate** (*STR*)**:** is the rate of corporate income tax as envisaged in the legislation of a given country. This is the rate that is applied to the tax base, which is the taxable profit of a company. A company's tax liability, in turn, is its tax base multiplied by the statutory tax rate.
- **Total Tax Rate** (*TTR*)**:** is the effective rate of the paid corporate income tax, which is calculated for each company individually through dividing the tax liability by the commercial profit. It is important to remember that the Total Tax Rate is by definition not equal to the Statutory Tax Rate due to the amount of non-taxable income and non-deductible expenses, as well as limitations on tax depreciation of assets.
- **Time:** is estimation of hours required for a company to prepare, file, and pay its corporate income taxes in one year.
- **Number_Taxes:** Number_Taxes is the number of all taxes (including corporate income tax) that the case study company is liable for paying. This is different from Tax Payments which are described here.
- **Population**[2]**:** is a country's population expressed in millions of people.

Apart from the aforementioned variables, we have constructed the following categorical variables:

- **Single rate:** equals 1 if statutory tax rate does not depend on the amount of taxable profit; equals 0 if the country uses progressive corporate income tax rates. This is a categorical indicator variable.
- **Population Category** (PopCat)**:** Population: grouped in three categories (low, medium, high)
- **Time:** grouped in three categories (low, medium, high).
- **Region:** grouped in 7 categories based on geographical region (OECD, East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, South Asia, Sub-Saharan Africa).

REGRESSION

This part of the project covers the results of a regression analysis performed on the dataset. It will concisely explain the important steps taken leading up to the model(s) that we consider to be effective and functional in explaining the variance in the response variable - Total Tax Rate (TTR). Our analysis starts out with all 173 observational units (that is, countries). Considering the fact that our sample closely overlaps with the population (that is, all countries in the world that impose corporate income tax), the proposed models will not be used for making predictions. Instead, we are merely interested in analyzing potential relationships between the variables. We took the following model (Model 1), incorporating most of the variables detailed in the World Bank's Doing Business dataset, as our starting point:

$$TTR = \beta_0 + \beta_1 * STR + \beta_2 * SingleRate + \beta_3 * STR * SingleRate + \beta_4 * Population + \beta_5 * Time\_Medium + \beta_6 * Time\_High + \beta_7 * EAAP + \beta_8 * EACA + \beta_9 * LAAC + \beta_{10} * MENA + \beta_{11} * SOAS + \beta_{12} * SSAF + \beta_{13} * Number\_Taxes + \varepsilon.$$

Model specifications:

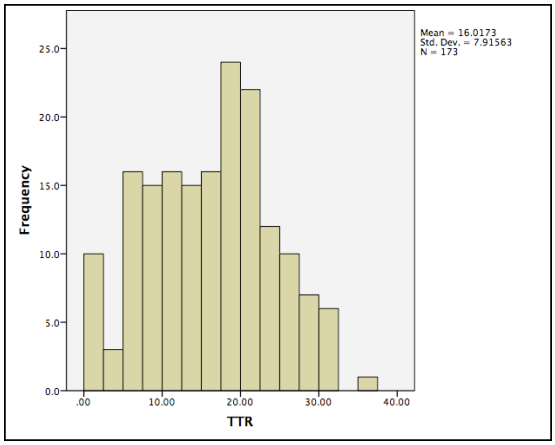
For the predictor *Time:*
Time_Low is 1 if time is < 30 hrs., 0 otherwise. This is the base category in the model.
Time_Medium is 1 if time is between 30 and 83.5 hrs., 0 otherwise;
Time_High is 1 if time is > 83.5 hrs., 0 otherwise.

The countries are grouped in 7 regions, in line with the World Bank's clustering of the data:
EAAP is 1 for countries from East Asia & Pacific, 0 otherwise;
EACA is 1 for countries from Europe & Central Asia, 0 otherwise;
LAAC is 1 for countries from Latin American & Caribbean, 0 otherwise;
MENA is 1 for countries from Middle East & North Africa, 0 otherwise;
SOAS is 1 for countries from South Asia, 0 otherwise;
SSAF is 1 for countries from Sub-Saharan Africa, 0 otherwise;
OECD is 1 for countries from OECD (High Income), 0 otherwise. (Model Base Category)

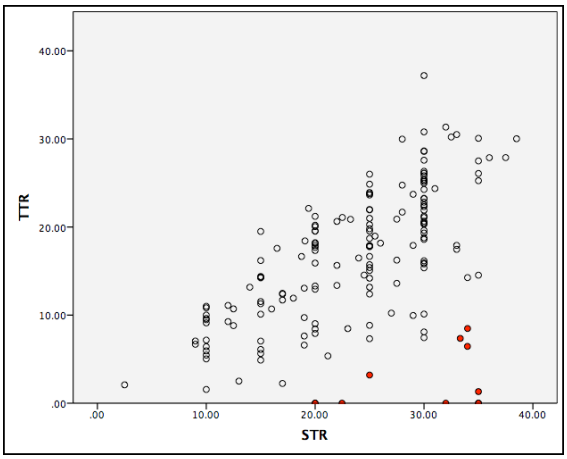
[1] http://www.doingbusiness.org/data/exploretopics/paying-taxes
[2] http://www.econstats.com/wdi/wdiv1072.htm

Looking at the data. Before fitting the regression model, we looked at univariate plots of each of the quantitative predictors (*STR*, *Population*, and *Number of Taxes*) and the response variable, *TTR*. The histogram of the Total Tax Rate (*TTR*) showed several outliers for which the *TTR* was either 0, or very close to 0 (see below). These data points were related to countries where the case study company would have to pay no or a very low (given the statutory tax rate) corporate income tax. We have identified two major reasons for these outliers: first, in some countries the main tax cost for the companies would be represented by a tax different from the corporate profit tax (for example, the sales tax); secondly, in some countries tax policy on deductibility of expenses could result in a case study company having nil or very low taxable profit. For the time being, we decided not to remove these 11 extreme points before we gained a better understanding of their impact. In addition, a histogram of the predictor Statutory Tax Rate (STR) demonstrated that most of the economies legally impose a tax rate greater than 20% on companies (with the peak at 30%). Interestingly, however, the effective tax rate (TTR; see histogram) is mostly between 5% and 25%. Furthermore, we include the scatterplot showing a clear linear relationship between TTR and STR, with few influential points and outliers (marked in red).

Response Variable, Total Tax Rate in % *(TTR)*



Scatterplot of TTR and STR

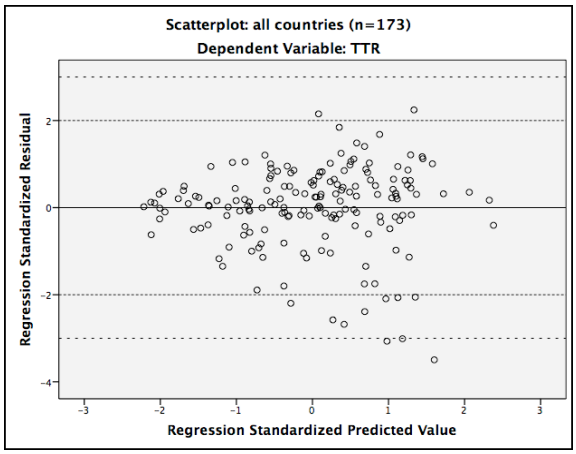


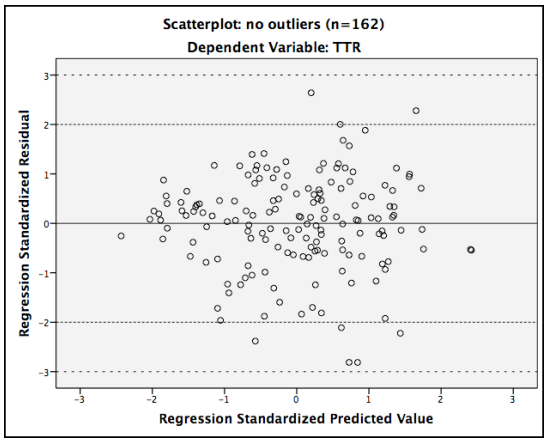
Checking for multicollinearity. The Variance Inflation Factor (VIF) is a statistic to capture multicollinearity. VIF > 5 can be an indicator of an issue; > 10 is very concerning. After fitting the full regression model, we found that there were three variables that showed multicollinearity and had a VIF above 18: the *Statutory Tax Rate* (*STR*), *Single_Rate*, and interaction of these two variables. Our assumption was that the multicollinearity was caused by the (higher-order) interaction term. We confirmed that after removing the interaction term all variables had a VIF below 2.1, and there was no concern regarding multicollinearity in lower-term variables. For the purposes of the further analysis the interaction term was included back to the model.

The assumptions.

- The residuals should have a mean of zero. This assumption is being met, since we are using the least squares method in our analysis.

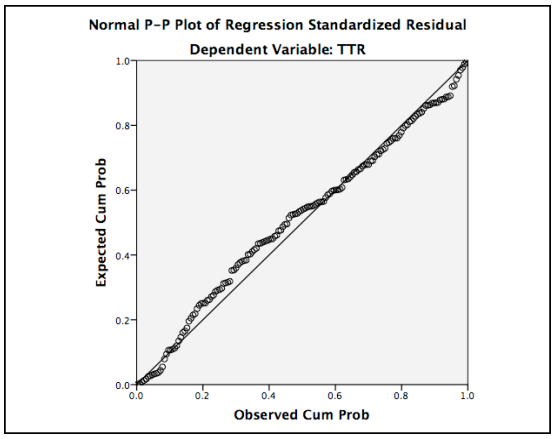
- The assumption of constant variance of the error terms, or homoscedasticity. There is a clear wedge-shape in the residuals, which indicates a violation of homoscedasticity (see residual plot below, left). In an attempt to even out the spread of the errors, several transformations were tested, including log(TTR), STR^2, $\sqrt{STR}$, and $1/STR$. None of these transformations changed the shape of the residual plot considerably. Therefore, we chose to proceed without transforming any of the variables. We decided to remove 11 economies from our dataset that were (a) identified as extreme points in the univariate plot of the response variable (*TTR*; see above), and (b) were more than 3 standard deviations away from the fitted residuals line (see residual plot below, left). The removed economies (also highlighted in red on the scatterplot of TTR and STR) are: Afghanistan, Argentina, Croatia, The Gambia, Guinea, Luxembourg, Zambia, Bolivia, France, Belgium and Venezuela. This reduced our sample from 173 observations to 162. This improved the shape of the residual plot a lot; the wedge-shape is far less obvious (see residual plot below, right). While it continued to show some sign of heteroscedosticity, we considered this assumption to be met for the purposes of our regression analysis. Considering the fact that our sample closely overlaps with the population (that is, all countries in the world that impose corporate profit tax), this model will not be used for making predictions. Instead, we are merely interested in analyzing potential relationships between the variables.

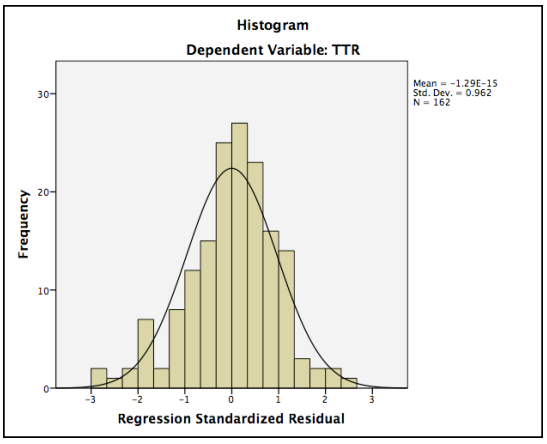




- The assumption of independent errors has been met, since there is no reason to expect that in different countries the tax regulations of interest will be dependent on each other. While one can argue that tax regulations can be to some extent driven by global and regional trends, each country would still establish internal rules at its own discretion.

- The assumption of normality of the error terms has been met. The NPP-plot below shows that the data points follow the normal line fairly accurately. This is further confirmed by the histogram of residuals, which shows a bell-shaped distribution centered around zero.





As a bottom-line, we consider the assumptions of a mean of zero, independence and normality of the error terms to be fully satisfied. The assumption of equal variance is partially satisfied after removing the outliers. While the plot still slightly thickens, we have decided against taking the log of our response variable for interpretation purposes. Despite the fact that inferences may thus be less accurate for high values of our predictors, we consider this assumption to be met for the purposes of our analysis.

Fitting the regression model. We started with fitting the full regression model (Model 1; see above).

1) Fitting Model 1:

*TTR^ = 4.774 + 0.766 * STR - 1.82 * SingleRate - 0.035 * STR * SingleRate - 0.0000000054 * Population - 0.777 * Time_Medium - 0.876 * Time_High + 0.465 * EAAP + 0.077 * EACA + 2.575 * LAAC + 0.433 * MENA + 4.171 * SOAS - 0.893 * SSAF - 0.331 * Number_Taxes.*

The Overall F-test (testing overall usefulness of the model):

H_0: $ß_1 = ß_2 = \ldots = ß_{13} = 0$; H_a: at least one beta is not zero.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5051.847	13	388.604	16.698	.000[b]
	Residual	3444.325	148	23.272		
	Total	8496.172	161			

Since F = 16.7 and p is close to zero, we reject H_0. This suggests that at least one variable in the model is useful in predicting the average tax cost (Total Tax Rate or *TTR*). This model explains 55.9% of variability in *TTR*, after correcting for the sample size and number of predictors (i.e., R^2_{adj} = 0.559), and has a standard error of estimate (SSE) of 4.82.

2) Through removing non-significant predictors from our model, one at a time, we tried to increase the accuracy of our model.

The significance of each variable was tested using the following hypotheses:
H_0: $ß_i = 0$; H_a: $ß_i \neq 0$.

The decision to reject or fail to reject the null-hypothesis was based on the t-statistic and the respective p-value of beta of the predictor in question, with 5% level of significance ($\alpha = 0.05$). The predictors for which the probability of obtaining a given t-value was higher than 5% were thus considered as non-significant. For categorical variables, we removed the complete set of predictors when one or more of its levels yielded very high p-value.
First, we removed the interaction term *STR * SingleRate* as it had the highest p-value amongst the predictors. This increased R^2_{adj} to 56.2%, and reduced the standard error of estimate to 4.81.
After refitting the regression model, we continued the process of identifying and removing other non-significant predictors, using the same approach. Second, we removed the time categories (*Time_Medium* and *Time_High*). This increased R^2_{adj} to 56.5%, and reduced the standard error of estimate to 4.79.
After refitting the model, once again we used the same approach for identifying the most non-significant predictors in the model. However, this time, whereas the predictors *SingleRate*, *Population*, and the region categories were rendered non-significant, removing any of them was resulting in a model with a lower R^2_{adj} and a higher standard error of the estimate. That being the case, we concluded that despite the fact that these variables are non-significant on their own, after accounting for other variables, their combination increases the accuracy of the model. Therefore, we decided to keep these three predictors in the model. Below is a table summarizing change in R^2_{adj} and the standard error of estimate (SSE) after removing each of the non-significant predictors. (Note that this is not a cumulative change.)

Predictor removed	**None**	**SingleRate**	**Population**	**Region (6 categories)**
R^2_{adj}	56.5%	56.2%	55.7%	54.6%
SSE	4.79	4.81	4.83	4.89

Accordingly, we found that the following fitted model provides the most accurate estimation of the average tax cost:

Model 2:
*TTR^ = 4.794 + 0.733 * STR - 2.499 * SingleRate - 0.0000000057 * Population + 0.457 * EAAP + 0.108 * EACA + 2.584 * LAAC + 0.464 * MENA + 4.11 * SOAS - 0.958 * SSAF - 0.33 * Number_Taxes.*

The Overall F-test (testing overall usefulness of the model):
H_0: $ß_1 = ß_2 = \ldots = ß_{10} = 0$; H_a: at least one beta is not zero.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5033.831	10	503.383	21.954	.000[b]
	Residual	3462.341	151	22.292		
	Total	8496.172	161			

Since F = 21.95 and p is close to zero, we reject H_0. This suggests that at least one variable in the model is useful in predicting the average tax cost (Total Tax Rate or *TTR*). This model explains 56.5% of variability in *TTR* after correcting for the sample size and number of predictors (i.e., R^2_{adj} = 0.565), and has a standard error of estimate (SSE) of 4.79.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.770[a]	.592	.565	4.78847

The nested model

Given that the total tax cost (or Total Tax Rate) is often associated with the statutory tax rate only, we decided to use the Nested F-test to further investigate relationship between *TTR* and *STR* as well as the testing the usefulness of adding the other predictors to the final model (Model 2; see above).

Full model (Model 2): $TTR = \beta_0 + \beta_1 * STR + \beta_2 * SingleRate + \beta_3 * Population + \beta_4 * EAAP + \beta_5 * EACA + \beta_6 * LAAC + \beta_7 * MENA + \beta_8 * SOAS + \beta_9 * SSAF + \beta_{10} * Number\_Taxes + \varepsilon$.

Nested model: $TTR = \beta_0 + \beta_1 * STR + \varepsilon$.

The table below shows the sum of squares, the number of predictors, and R^2_{adj} both for the full and nested models.

	Sum of Squares (Regression)	Sum of Squares (Errors)	Number of predictors	R^2_{adj}
Full model	5033.831	3462.341	10	56.5%
Nested model	4442.680	4053.492	1	52.0%

H_0: β_i=0 for all predictors in the nested model; H_a: $\beta_i \neq 0$ for at least one predictor in the nested model.

$F = ((SSM_{full} - SSM_{nested})/\#\text{ of new predictors})/(SSE_{full}/(n-k-1))$

where *n* is the number of observations = 162; *k* is the number of predictors in the full model = 10.

F = ((5033.8-4442.7)/9)/(3462.3/(162-10-1)) = 2.86

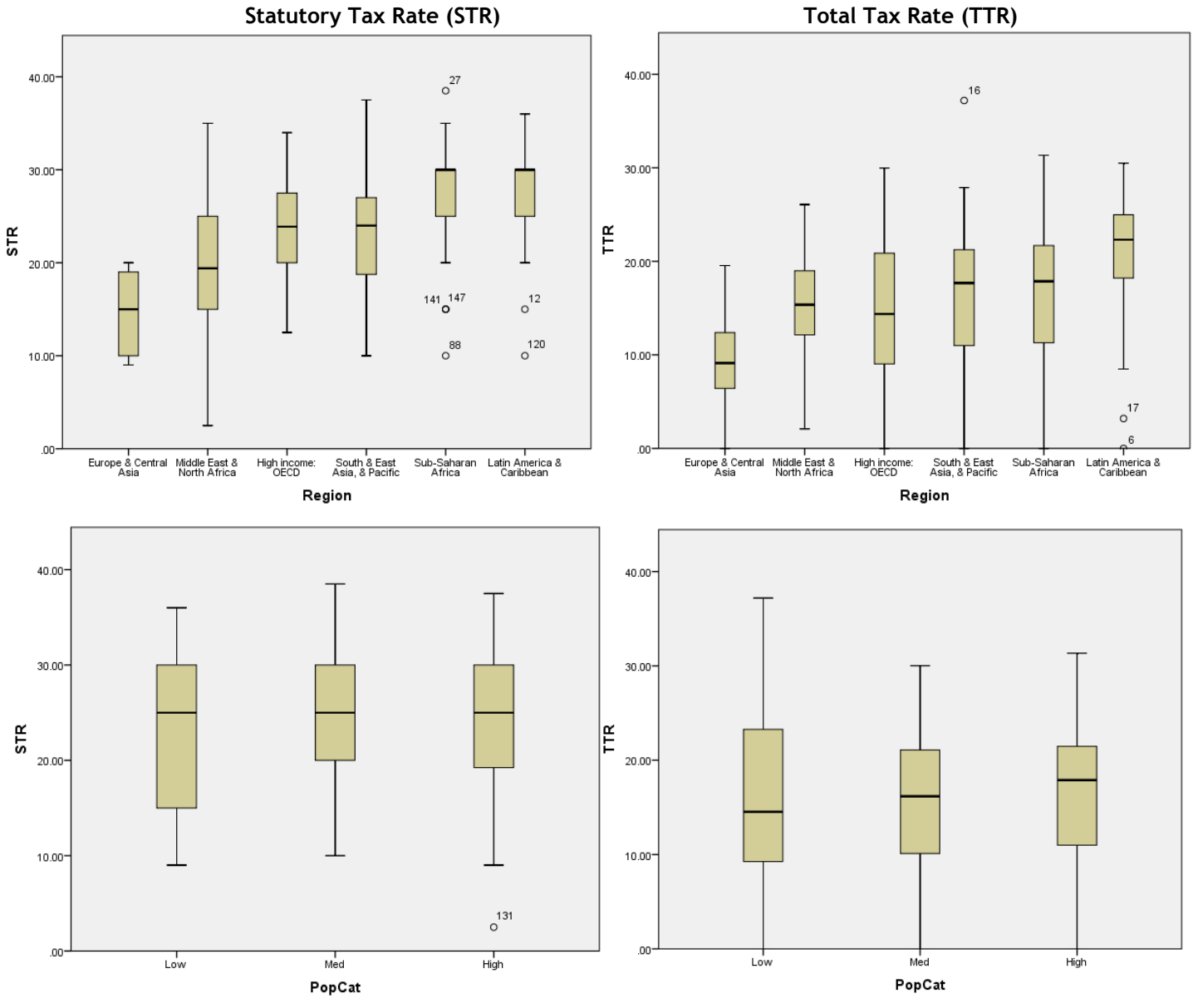
Critical value of F for α = 0.05, based on numerator = 9 and denominator (162-10-1) = 151, is 1.94.

F = 2.86 > critical value = 1.94 => Therefore, p < 0.05

Given the p-value is below α = 0.05 and the F-value (2.86) is high, we have enough evidence to reject H_0. Therefore, the data suggest that one of the tested betas, or some combination of them, is useful in predicting average tax cost, after accounting for the statutory tax rate. Therefore, while *STR* only explains most of the variability in *TTR*, adding the combination of *SingleRate*, *Population*, *Number_Taxes* and regional categories have been found to improve the model by increasing its accuracy.

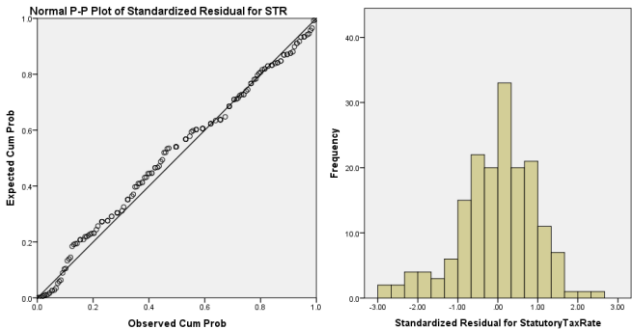
ANOVA

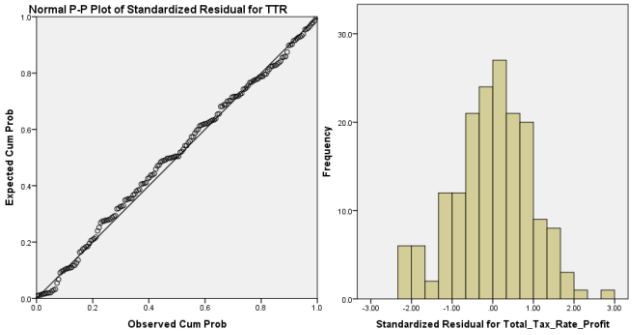
In addition to regression, we used two-way ANOVA to explore the potential for different average statutory tax rate (STR) and total tax rate (TTR) for the various regions and population sizes throughout the world. For this analysis we returned to the original data set before any observations were removed, however a few adjustments were necessary to accommodate the ANOVA modeling. Due to the low overall size of the dataset some groups had few observations. This became problematic when crossing the factors for two way ANOVA because it yielded even smaller block sizes. For this reason we merged the South Asia region, which had the lowest number of total observations (n=8), with the East Asia & Pacific region (new region: SEAP). This gave a better opportunity to explore a possible interaction between region and population. Next we had to categorize the populations into groups for the ANOVA model. The following categories were defined to give the best overall balance of group sizes: Low = less than 5 million (n=65), Medium = 5 million to less than 20 million (n=53), High = more than 20 million (n=55). Throughout the ANOVA portion of this report we will show output for Statutory Tax Rate on the left and Total Tax Rate as a percentage of corporate profit on the right. This will allow us to look at both of these response variables together and efficiently. We begin with boxplots of each response variable against the two factors of Region & Categorized Population to check for any apparent relationships.



The box plots show some obvious relationships between the response variable and the levels of the predictor Region, while PopCat shows only slight deviations among the levels. We also see a few outliers in the dataset that may need further investigation for possible errors at a later stage. Keeping this in mind, we look for our data meeting assumptions of two-way ANOVA.

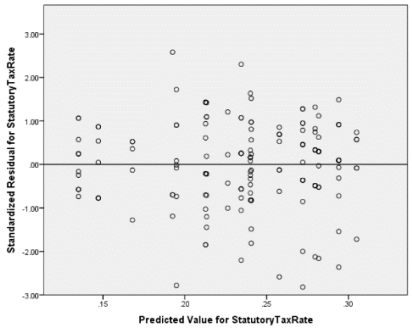
1. **Residuals should have a mean of zero:** This is a given of the ANOVA modeling process, which forces the residuals' mean to be zero.
2. **Data should be obtained via a random process:** This is usually achieved by randomization, but in our case we accept this assumption because our sample is more or less equivalent to the population of interest.
3. **Independence:** Countries are certainly influenced to some degree by the decisions of other governments, however each is free to impose their own tax laws. Each country has their own local concerns and traditions that go into the mix of what becomes their unique taxing strategies and laws.
4. **Normality of errors:** The errors should follow a normal distribution which is verified by the close alignment of data points to the normal line on an NPP-Plot. This is seen in the two plots below for both the Statutory Tax Rate and the Total Tax Rate satisfying the condition of normality. Also, the histogram of standardized residuals shows to be approximately normal.

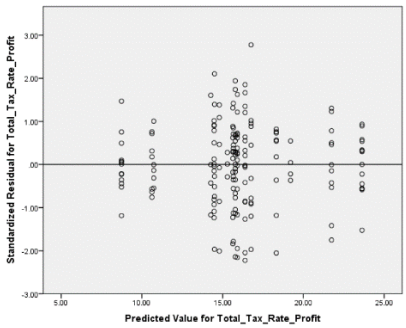




5. **Tests for equal variance among groups**: Three tests for equal variance were performed and these results are included below. For both response variables, two of these tests agree that there is equal variance. The one test that fails is the standard deviation max/min rule of thumb. Since this dataset has a low number of observations for each group, we accept this discrepancy with the rule of thumb and conclude the dataset does pass the condition of equal variances among groups. Also, to investigate this concern, nonparametric Kruskal-Wallis pairwise comparisons were ran on both STR and TTR for the Region factor. Due to space constraints in this report, full details of the nonparametric analysis is kept to a minimum, however these tests overall agreed with Tukey's post-hoc analysis of Region and distributions were checked for a continuous, similar shape.

a) **Residual plots**: Boxplots were provided above, and although in the statutory vs region boxplot Europe & Central Asia shows a slightly lower variance than others, overall both plots show that equal variance is not grossly violated. Residual to fitted value plots reflect similar measures of spread to the boxplots. There are some outliers across the entire sample space, however since the number of observations is small. Since nonparametric analysis agreed with the findings, we chose to leave these in the final ANOVA analysis.





b) **Levene's Test of Equality of Error Variances.** Design: Intercept + Region + PopCat + Region*PopCat
Hypothesis for both response variables: $\mathbf{H_0}$: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \ldots = \sigma_{18}^2$ $\mathbf{H_a}$: not all variances are equal

Dependent Variable: STR

F	df1	df2	Sig.
1.018	17	155	.441

Decision: Pvalue = 0.441 > 0.05 → do not reject H_0
Conclusion: There is equal variance among the groups.

Dependent Variable: TTR

F	df1	df2	Sig.
1.380	17	155	.153

Decision: Pvalue = 0.153 > 0 .05 → do not reject H_0
Conclusion: There is equal variance among the groups.

c) **Rule of Thumb**: S_{max} / S_{min} = .1059/.0382 = 2.765

Rule of Thumb: S_{max} / S_{min} = 10.986/2.086 = 5.267

Fitting the models: **Statutory Tax Rate (STR)** **Total Tax Rate (TTR)**

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4055.800[a]	17	238.576	6.405	.000000
Intercept	69393.565	1	69393.565	1863.135	.000000
Region	3222.326	5	644.465	17.303	.000000
PopCat	99.573	2	49.787	1.337	.265719
Region * PopCat	140.565	10	14.057	.377	.954928
Error	5773.066	155	37.246		
Total	106803.679	173			
Corrected Total	9828.866	172			

a. R Squared = .413 (Adjusted R Squared = .348)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2372.182[a]	17	139.540	2.57	.001158
Intercept	32781.725	1	32781.725	605	.000000
Region	1229.883	5	245.977	4.54	.000690
PopCat	.486	2	.243	.004	.995527
Region * PopCat	692.858	10	69.286	1.28	.247454
Error	8404.856	155	54.225		
Total	55160.633	173			
Corrected Total	10777.038	172			

a. R Squared = .220 (Adjusted R Squared = .135)

Region main effect analysis: H_0 : $\mu_{OECD} = \mu_{SEAP} = \mu_{EACA} = \mu_{LAAC} = \mu_{MENA} = \mu_{SSAF}$ **H_a : the μ are not all equal**

Decision: Pvalue = 1.3778×10^{-13} < 0.001 → reject H_0
Conclusion: Average STR differs among at least one of the regions.

Decision: Pvalue = 0.000690 < 0.001 → reject H_0
Conclusion: Average TTR differs among at least one of the regions.

PopCat main effect analysis: H_0 : $\mu_{LOW} = \mu_{MED} = \mu_{HIGH}$ **H_a : the μ are not all equal**

Decision: Pvalue = 0.265719 > 0.05 → do not reject H_0
Conclusion: Average STR is similar among population levels.

Decision: Pvalue = 0.995527 > 0.05 → do not reject H_0
Conclusion: Average TTR is similar among population levels.

Interaction main effect analysis: H_0 : the main effect of each factor is the same for each level of the other factor
H_a : region & population interact

Decision: Pvalue = 0.954928 > 0.05 → do not reject H_0
Conclusion: The main effect of Region is the same for the different levels of population.

Decision: Pvalue = 0.247454 > 0.05 → do not reject H_0
Conclusion: The main effect of Region is the same for the different levels of population.

Main effect analysis results indicate that no interaction is present and only the main effect of region is significant. With this in mind we followed up with post-hoc analysis using Tukey's honest significant difference to interpret the results of region on average tax rates while mitigating the potential for errors from multiple comparisons. Due to the potential for outlier influence, one-way nonparametric Kruskal-Wallis pairwise comparisons were also ran against Regional medians. The output from these tests is included on the following page, but overall the nonparametric analysis of the medians agreed with Tukey's post-hoc analysis of the means using an adjusted-significance PValue at α = 0.05. What follows is an overview of these differences *emphasizing the regions with the lowest average tax rates and moving up*, using a significance level for these differences of α = 0.05

Pairwise comparison hypothesis: $H_0 : \mu_i = \mu_j$ $H_a : \mu_i \neq \mu_j$ Reject H_0 if PValue is < 0.05
Nonparametric Pairwise Comparisons: $H_0 : \Theta_i = \Theta_j$ $H_a : \Theta_i \neq \Theta_j$ Reject H_0 if Adjusted Significance is < 0.05

- **Europe & Central Asia:** Lowest overall average tax rates, with a significantly lower average Statutory Tax Rate than all other regions except the Middle East & North Africa. Also, a significantly lower average Total Tax Rate than all other regions except High Income OECD, & the Middle East & North Africa.
- **The Middle East & North Africa:** Significantly lower average Statutory Tax Rate than the Latin America & Caribbean, and the Sub-Saharan Africa. However, The Total Tax Rate is similar to all other regions.
- **High Income OECD:** Significantly lower average Statutory Tax Rate than Sub-Saharan Africa. Similar average Total Tax Rates to all other regions.
- **South & East Asia, & Pacific:** Significantly lower average Statutory Tax Rate than Latin America & Caribbean, and Sub-Saharan Africa. Does not have a significantly lower average Total Tax rate from other regions.
- **Sub-Saharan Africa:** Not significantly lower than any other region in either average Tax rate.
- **Latin America & Caribbean:** Not significantly lower than any other region in either average Tax rate.

Tukey's Post-Hoc Analysis: Statutory Tax Rate (STR) Total Tax Rate (TTR)

(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound				Lower Bound	Upper Bound
Europe & Central Asia	High income: OECD	-8.9927*	1.65268	.000	-13.7618	-4.2235	-5.2466	1.99411	.096	-11.0011	.5078
	Latin America & Caribbean	-13.3103*	1.66558	.000	-18.1167	-8.5040	-10.4897*	2.00968	.000	-16.2890	-4.6903
	Middle East & North Africa	-5.2933	1.99320	.090	-11.0451	.4585	-5.5351	2.40499	.200	-12.4752	1.4050
	South & East Asia, & Pacific	-8.0086*	1.66558	.000	-12.8150	-3.2023	-7.0212*	2.00968	.008	-12.8205	-1.2218
	Sub-Saharan Africa	-13.7778*	1.52233	.000	-18.1708	-9.3848	-7.8034*	1.83684	.001	-13.1040	-2.5028
High income: OECD	Europe & Central Asia	8.9927*	1.65268	.000	4.2235	13.7618	5.2466	1.99411	.096	-.5078	11.0011
	Latin America & Caribbean	-4.3177	1.58929	.078	-8.9039	.2685	-5.2430	1.91763	.074	-10.7767	.2907
	Middle East & North Africa	3.6993	1.92991	.396	-1.8698	9.2685	-.2884	2.32862	1.000	-7.0082	6.4313
	South & East Asia, & Pacific	.9840	1.58929	.989	-3.6022	5.5703	-1.7745	1.91763	.939	-7.3082	3.7592
	Sub-Saharan Africa	-4.7851*	1.43847	.014	-8.9361	-.6341	-2.5568	1.73565	.682	-7.5654	2.4518
Latin America & Caribbean	Europe & Central Asia	13.3103*	1.66558	.000	8.5040	18.1167	10.4897*	2.00968	.000	4.6903	16.2890
	High income: OECD	4.3177	1.58929	.078	-.2685	8.9039	5.2430	1.91763	.074	-.2907	10.7767
	Middle East & North Africa	8.0170*	1.94097	.001	2.4160	13.6181	4.9546	2.34197	.285	-1.8036	11.7128
	South & East Asia, & Pacific	5.3017*	1.60270	.015	.6768	9.9266	3.4685	1.93382	.473	-2.1119	9.0489
	Sub-Saharan Africa	-.4674	1.45328	1.000	-4.6612	3.7263	2.6862	1.75352	.644	-2.3739	7.7464
Middle East & North Africa	Europe & Central Asia	5.2933	1.99320	.090	-.4585	11.0451	5.5351	2.40499	.200	-1.4050	12.4752
	High income: OECD	-3.6993	1.92991	.396	-9.2685	1.8698	.2884	2.32862	1.000	-6.4313	7.0082
	Latin America & Caribbean	-8.0170*	1.94097	.001	-13.6181	-2.4160	-4.9546	2.34197	.285	-11.7128	1.8036
	South & East Asia, & Pacific	-2.7153	1.94097	.728	-8.3163	2.8858	-1.4861	2.34197	.988	-8.2443	5.2721
	Sub-Saharan Africa	-8.4844*	1.81954	.000	-13.7351	-3.2338	-2.2683	2.19545	.906	-8.6037	4.0671
South & East Asia, & Pacific	Europe & Central Asia	8.0086*	1.66558	.000	3.2023	12.8150	7.0212*	2.00968	.008	1.2218	12.8205
	High income: OECD	-.9840	1.58929	.989	-5.5703	3.6022	1.7745	1.91763	.939	-3.7592	7.3082
	Latin America & Caribbean	-5.3017*	1.60270	.015	-9.9266	-.6768	-3.4685	1.93382	.473	-9.0489	2.1119
	Middle East & North Africa	2.7153	1.94097	.728	-2.8858	8.3163	1.4861	2.34197	.988	-5.2721	8.2443
	Sub-Saharan Africa	-5.7692*	1.45328	.002	-9.9629	-1.5754	-.7823	1.75352	.998	-5.8424	4.2779
Sub-Saharan Africa	Europe & Central Asia	13.7778*	1.52233	.000	9.3848	18.1708	7.8034*	1.83684	.001	2.5028	13.1040
	High income: OECD	4.7851*	1.43847	.014	.6341	8.9361	2.5568	1.73565	.682	-2.4518	7.5654
	Latin America & Caribbean	.4674	1.45328	1.000	-3.7263	4.6612	-2.6862	1.75352	.644	-7.7464	2.3739
	Middle East & North Africa	8.4844*	1.81954	.000	3.2338	13.7351	2.2683	2.19545	.906	-4.0671	8.6037
	South & East Asia, & Pacific	5.7692*	1.45328	.002	1.5754	9.9629	.7823	1.75352	.998	-4.2779	5.8424

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Europe & Central Asia-Middle East & North Africa	-33.867	16.233	-2.086	.037	.554
Europe & Central Asia-South & East Asia, & Pacific	-48.462	13.565	-3.573	.000	.005
Europe & Central Asia-High income: OECD	-52.700	13.460	-3.915	.000	.001
Europe & Central Asia-Latin America & Caribbean	-86.893	13.565	-6.406	.000	.000
Europe & Central Asia-Sub-Saharan Africa	-89.711	12.398	-7.236	.000	.000
Middle East & North Africa-South & East Asia, & Pacific	-14.595	15.808	-.923	.356	1.000
Middle East & North Africa-High income: OECD	18.833	15.717	1.198	.231	1.000
Middle East & North Africa-Latin America & Caribbean	53.026	15.808	3.354	.001	.012
Middle East & North Africa-Sub-Saharan Africa	-55.844	14.819	-3.769	.000	.002
South & East Asia, & Pacific-High income: OECD	4.238	12.943	.327	.743	1.000
South & East Asia, & Pacific-Latin America & Caribbean	38.431	13.053	2.944	.003	.049
South & East Asia, & Pacific-Sub-Saharan Africa	-41.249	11.836	-3.485	.000	.007
High income: OECD-Latin America & Caribbean	-34.193	12.943	-2.642	.008	.124
High income: OECD-Sub-Saharan Africa	-37.011	11.715	-3.159	.002	.024
Latin America & Caribbean-Sub-Saharan Africa	-2.818	11.836	-.238	.812	1.000

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Europe & Central Asia-High income: OECD	-36.823	13.563	-2.715	.007	.099
Europe & Central Asia-Middle East & North Africa	-37.540	16.357	-2.295	.022	.326
Europe & Central Asia-South & East Asia, & Pacific	-47.971	13.669	-3.510	.000	.007
Europe & Central Asia-Sub-Saharan Africa	-52.673	12.493	-4.216	.000	.000
Europe & Central Asia-Latin America & Caribbean	-72.523	13.669	-5.306	.000	.000
High income: OECD-Middle East & North Africa	-.717	15.838	-.045	.964	1.000
High income: OECD-South & East Asia, & Pacific	-11.148	13.043	-.855	.393	1.000
High income: OECD-Sub-Saharan Africa	-15.850	11.805	-1.343	.179	1.000
High income: OECD-Latin America & Caribbean	-35.699	13.043	-2.737	.006	.093
Middle East & North Africa-South & East Asia, & Pacific	-10.431	15.929	-.655	.513	1.000
Middle East & North Africa-Sub-Saharan Africa	-15.133	14.932	-1.013	.311	1.000
Middle East & North Africa-Latin America & Caribbean	34.983	15.929	2.196	.028	.421
South & East Asia, & Pacific-Sub-Saharan Africa	-4.702	11.926	-.394	.693	1.000
South & East Asia, & Pacific-Latin America & Caribbean	24.552	13.153	1.867	.062	.929
Sub-Saharan Africa-Latin America & Caribbean	19.849	11.926	1.664	.096	1.000

CONCLUSION

In our regression analyses, we have come to two models explaining the Total Tax Rate (TTR). The first model explains the most of the variation (54.6%) in the TTR: TTR^ = 4.794 + 0.733 * STR - 2.499 * SingleRate - 0.0000000057 * Population + 0.457 * EAAP + 0.108 * EACA + 2.584 * LAAC + 0.464 * MENA + 4.11 * SOAS - 0.958 * SSAF - 0.33 * Number_Taxes. The second (nested) model, whilst shorter, explains only 52% of the variation in TTR: TTR^ = 0.599 + 0.702 * STR. We concluded with a Nested F-test showing the usefulness of adding another 9 variables to the subset model. While the subset model with only one predictor, in this case STR, would be easier to interpret and more convenient for practical use, the full model did turn out to explain another 4.5% of the variation in the response variable, TTR. In other words, this is a choice between accuracy and ease of use that really is up to the user of the model and the context in which it will be applied. In building these models, we were surprised not to find a significant interaction between the Statutory Tax Rate (STR) and whether or not the economy upholds a single or progressive rate system. In our preliminary discussions we did expect to find some indication that the two would be significantly related.

The fitted two-way ANOVA models for STR and TTR returned similar results. There was a significant main effect for the regions, a non-significant main effect for population categories and a non-significant interaction between regions and population categories. The main effects are interpreted as different average STR and TTR among regions and similar mean tax rates amongst population levels. This is particularly interesting because one would have thought population categories is analogous to regions in that if a region pays more taxes, then population categories from it should reflect a similar significant tax amount. We also initially predicted there would be an interaction between region and population such that tax rates in different regions would adjust to reflect the various population levels within each region.

The largest issue of concern in analysis of this dataset was the outliers in the data, shown in the scatterplot for both regression and ANOVA. Outlier mediation & nonparametrics were used to mitigate this concern. Overall, we felt that this comes down to the fact that this dataset was relatively small for the number of predictors and levels. Future analysis could incorporate multiple years' worth of data to validate the significance of these findings.

Dataset snapshot:

	Economy	TTR	STR	SingleRate	Number_Taxes	TimeHours	Region	Population	PopCat	SOAS	EAAP	MENA	EACA	LAAC	SSAF	OECD	Time_Low	Time_Medium	Time_High
1	Afghanistan	.00	20.00	1	4	77	South & East Asia, & Pacific	27208325	High	1	0	0	0	0	0	0	0	1	0
2	Albania	9.48	10.00	1	8	119	Europe & Central Asia	3143291	Low	0	0	0	1	0	0	0	0	0	1
3	Algeria	6.59	19.00	1	11	152	Middle East & North Africa	34361756	High	0	0	1	0	0	0	0	0	0	1
4	Angola	25.26	35.00	1	7	75	Sub-Saharan Africa	18020668	Med	0	0	0	0	0	1	0	0	1	0
5	Antigua and Barbuda	26.00	25.00	1	11	23	Latin America & Caribbean	85536	Low	0	0	0	0	1	0	0	1	0	0
6	Argentina	.00	35.00	1	10	105	Latin America & Caribbean	39876118	High	0	0	0	0	1	0	0	0	0	1
7	Armenia	19.55	20.00	1	5	121	Europe & Central Asia	3077087	Low	0	0	0	1	0	0	0	0	0	1
8	Australia	26.10	30.00	1	10	37	High income: OECD	21374000	High	0	0	0	0	0	0	1	0	1	0
9	Austria	15.40	25.00	1	13	47	High income: OECD	8344319	Med	0	0	0	0	0	0	1	0	1	0
10	Azerbaijan	12.94	20.00	1	7	60	Europe & Central Asia	8678851	Med	0	0	0	1	0	0	0	0	1	0