Jordan Coursey
A11605033

**Motivation and theoretical Underpinnings:**

The goal of this report is to validate methods of fostering small and medium enterprises to promote entrepreneur-driven economic growth in GPSistan. The first assumption is that lowering the cost of starting a business is critical to the number of successful small to medium sized businesses founded because they do not have as much capital as a large company and are more susceptible to small changes. The goal is to find what business practices lead to the creation of more small and medium businesses. We do not have the number of businesses started so we will use the proxy variable "Starting a business cost (% of income per capita.)"

The general hypothesis this paper is that people would start more small and medium businesses if the process of getting your business started was expedited.

**Data selection:**

I assume that the cost of starting a new business is not directly in the control of the government and so is logically our observed effect (i.e. dependent variable.) The goal is to find evidence that decreasing the amount of time it takes to get a business running decreases the cost of starting a business.

I used four time-related variables in the data set and see how they affect the cost of starting a business; time to start a business, time to get a construction permit, time to get electricity and time to register property in days. The model controls for the time it takes for the government to perform four basic necessities of a business. The final model takes into account all four time variables to isolate the effect of each governmental process separately from the rest. This is necessary to eliminate the ability for these times to overlap with each other. When they are regressed together

The observations were done on a national level across 190 countries. Cities and countries with insufficient data were thrown out. I took the natural log of the dependent variable, cost of starting a business, later in the model to fix issues with homoscedasticity.

**Regression analysis and results:**

Figure 1 demonstrates the progression of the variables effects on the cost of starting a business. There is sufficient data to show a positive correlation between the cost of starting a business and the time it takes to start a business in days. For every day increase in start time the cost goes up .57 percentage points (as cost is in percent of income per capita) with a 99.9% degree of confidence. I wanted to see whether there would be any omission bias so I added the variables for construction permit time and the time to get electricity. The time it takes to get a construction permit actually had a negative correlation with cost (-.02) and time to get electricity had a correlation of .02 with cost, however, neither statistically significant. Neither effected the significance or impact of the start time variable. Lastly, I added property registration time which had a correlation coefficient of .085 with 95% confidence. (figure 2)

Start time and property registration time were the most statistically significant, so I ran a regression with a full interaction between these two variables to see if there was an amplified effect with the combination of these two. There was a significant interaction term between the two (.0014), meaning that increasing these two variables has an effect on cost in addition to their individual effects.

Jordan Coursey
A11605033

**Threats to validity, regression diagnostics**

The mean "vif" comes out to be 1.13 suggesting low collinearity meaning independent variables aren't affecting each other. The hettest yielded a "p value" of .000, giving us enough proof to reject its homoscedasticity, which violates a Gauss-Markov assumption threatening the validity of the model. In figure 3 on the left graph there is a visual trend suggesting that there is an omitted variable bias, meaning an unaccounted for variable affecting our data. The graph on the right is the result of taking the natural log of the dependent variable (the cost of starting a business as a % of income per capita.) This creates a log-level regression and is interpreted as a 1 unit change in x results in a percentage change in y. The resulting p value is .6460 meaning the model is homoscedastic. The apparent lack of an omitted variable would suggest that the zero conditional mean assumption holds, however, I would suggest more tests to be run to prove that.

Venezuela has the highest leverage and the second highest R-Stud followed by the C.A.F. and could be thrown out. (figure 4) There were 5 observations that met every outlier test and are also candidates for being thrown out. (figure 5)

**Discussion and conclusion**

The analysis should control for GDP growth in the future as the growth of the economy would likely change how easy it is to start a new business. It was not clear as to how the cost of starting a new business was derived in the write-up. This made its use as a proxy variable troublesome because we do not know whether these costs were exogenous to the government or if they were costs that were imposed by the government. In the future the data sets should include the number of small and medium businesses.

The cost of starting a business is not clearly explained in the write-up. The conclusions of this study rest on the assumption that that this cost is distinct from any governmental impositions on starting a business as those can be controlled by policy directly. If this were not the case, then the first step would be to reduce the costs instead.

There is sufficient evidence to suggest that lowering the time it takes to start a business and the time it takes to register property will lower the cost of starting a business. I would suggest that you further test the hypothesis using the amount of small and medium businesses started as the dependent variable instead of the proxy variable of cost.

*Figure 1: This graph shows at the margin how much each of the time variables affects the cost of starting a new business.*

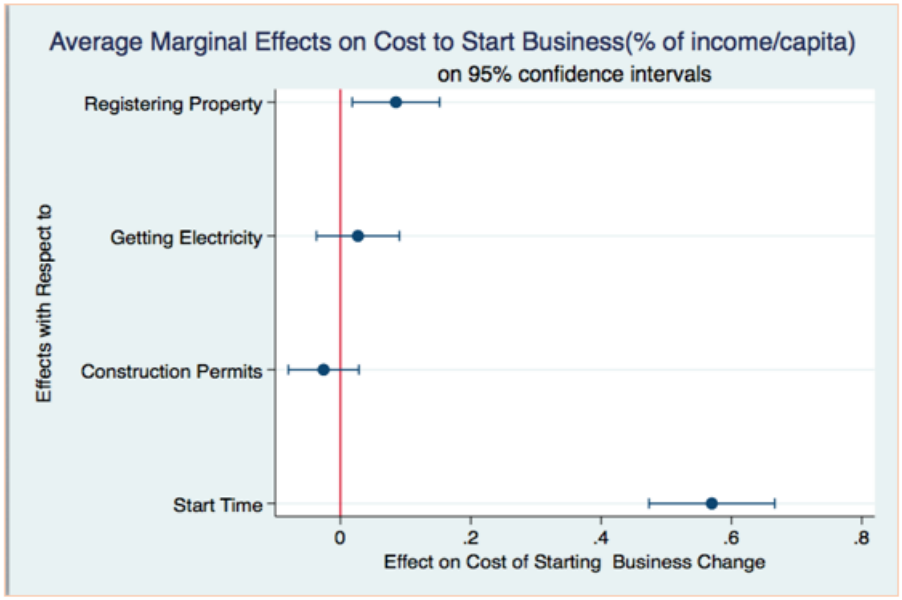
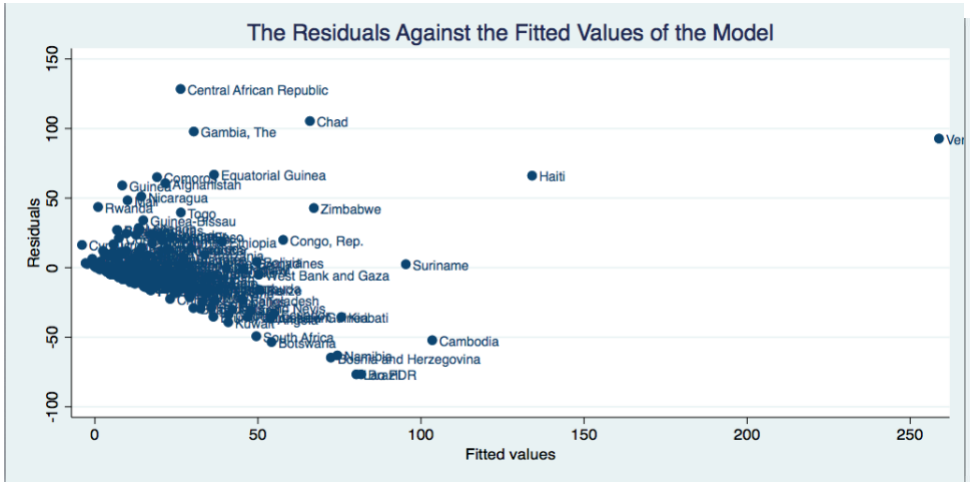



*Figure 2: the table shows the progression of the model from the first variable to the last with the significance and the standard error in parenthesis.*

Cost of starting a new business regressed on start time, construction permit time, getting electricity time, property registration time and the interaction between start time and property registration time

| | new business cost | new business cost | new business cost | new business cost | new business cost |
|---|---|---|---|---|---|
| start time | 0.573*** | 0.586*** | 0.588*** | 0.570*** | 0.476*** |
| | (9.99) | (9.32) | (11.49) | (11.65) | (7.13) |
| construction permit time | | -0.0317 | -0.0305 | -0.0258 | -0.0143 |
| | | (-0.91) | (-1.06) | (-0.94) | (-0.52) |
| getting electricity time | | | 0.0376 | 0.0269 | 0.0304 |
| | | | (1.11) | (0.83) | (0.95) |
| property register time | | | | 0.0852* | 0.0106 |
| | | | | (2.51) | (0.21) |
| start time * property register time | | | | | 0.00141* |
| | | | | | (2.06) |
| constant | 2.618 | 6.882 | 1.604 | -2.175 | -0.337 |
| | (0.74) | (1.20) | (0.31) | (-0.43) | (-0.07) |
| N | 190 | 185 | 184 | 181 | 181 |
| t statistics in parentheses | | | | | |
| * p<0.05 | ** p<0.01 | *** p<0.001" | | | |

Figure 3.1: The residuals by the fitted values before logging has a clear trend (heteroskedastic)
Figure 3.2: the residuals by the fitted values after logging has no trend (homoscedastic)



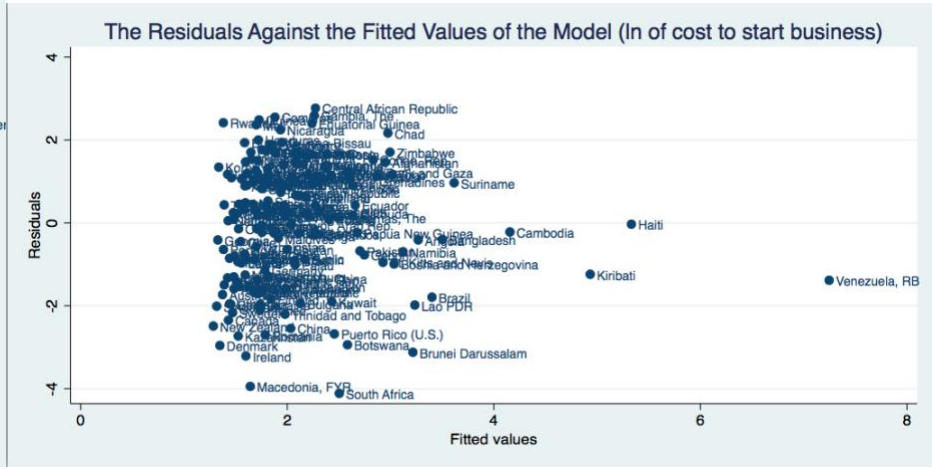



*Figure 4: This shows the relationship between two outlier tests with our observations*

Observations on Leverage by the Absolute Value of Studentized Residual

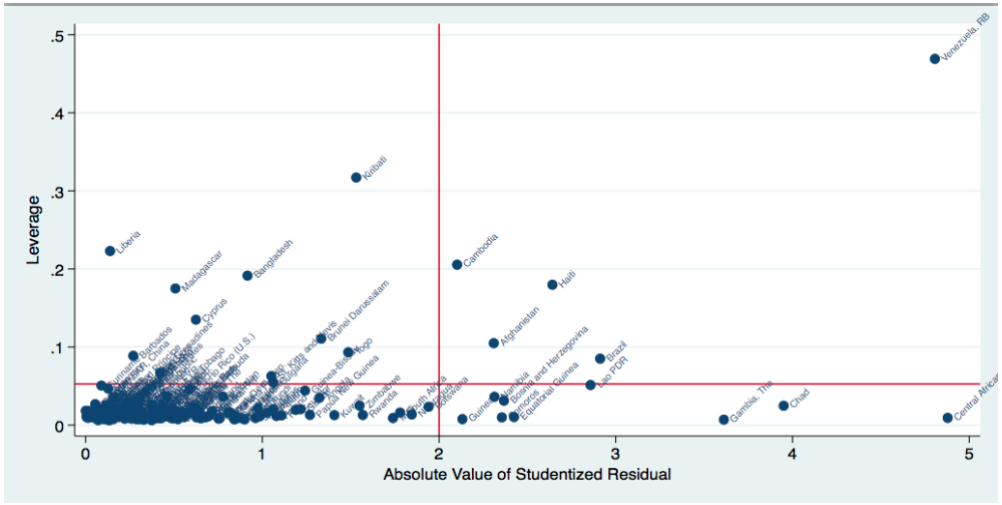



*Figure 5: Amount of observations that met each outlier criteria then the ones that met all four*

| Outlier Test | Residual stud | Leverage | Cooksd | dfit | 4 criteria met |
|---|---|---|---|---|---|
| # of Observations | 14 | 16 | 14 | 7 | 5 |