Yue Chen 28166149
Yuxuan Han 19543140
James (Mengzhe) Huang 30019129
Tao Li 40596141
Xueyang Zhao 50197145

Predicting National Export Value

Abstract

In our study, we look to construct a prediction equation for national export value, to illustrate relationship with various national and economic statistics, by examining the national export value of numerous countries and their corresponding data. For our response variable, we use national gross export in USD. For our explanatory variables, we use national gross import in USD, GDP per capita in USD, GNI per capita in USD, annual growth in percentage points, time to trade in days, number of exported products, cost to import per shipping container in USD, and whether said country is a member of the World Trade Organization as a binary variable. All of our variables are from the 2011 fiscal cycle and are scaled according for use in models. We will fit various models using our data, considering both linear models and those with transformations on the explanatory variables and/or the response variable by examining the residual plots. Good models will be selected with criterion based on its CP statistic and/or its adjusted R-squared value, and compared with cross validation, which will be conducted with a training and holdout set, as well as out of sample data from another fiscal year. Cook's distances will be used to determine influential data points, i.e. countries that play an influential global role in trade. From our analysis we conclude that the overall best model involves log of import, log of numbers of export products, binary variable wto and cost to export.

1. Variables

Response variables:

export: Export (US$ in 10^9)

Explanatory variables:

import: Import (US$ in 10^9)

wto: Is the country a WTO member or not. 1 is yes; 0 is no.

gdp: GDP per capita (current US$ in 10^3)

gni: GNI per capita (current US$ in 10^3)

growth: Country Growth (%)

time: Time to trade (days)

noex: No. of exported products (in 10^3)

cost: Cost to export (US$ per container in 10^3)

2.Data Analysis

	export	import	wto	gdp	gni	growth	time	noex	cost
Min	**0.007**	**0.092**	**0**	**0.356**	**0.9**	**-15.146**	**3.60**	**0.026**	**0.275**
1st Qu	**2.202**	**5.175**	**1**	**1.686**	**4.145**	**7.422**	**12**	**1.254**	**0.805**
Median	**10.161**	**15.431**	**1**	**5.540**	**10.29**	**11.024**	**17**	**2.321**	**1.1**
Mean	**94.693**	**102.837**	**0.863**	**11.937**	**14.315**	**12.264**	**20.71**	**2.401**	**1.389**
3rd Qu	**65.244**	**70.109**	**1**	**14.03**	**19.73**	**16.353**	**23.5**	**3.717**	**1.555**
Max	**1898.388**	**2263.619**	**1**	**100.575**	**63.33**	**51.142**	**76**	**4.6**	**5.491**
SD	**263.074**	**305.401**	**0.346**	**17.518**	**13.149**	**9.813**	**13.7**	**1.387**	**0.924**

Table 1: summary statistics of response variable and explanatory variables

	export	import	wto	gdp	gni	growth	time	noex	cost
export	**1**	**0.9616**	**0.0552**	**0.2735**	**0.3335**	**-0.1028**	**-0.2395**	**0.4555**	**-0.2037**
import	**0.9616**	**1**	**0.0858**	**0.2787**	**0.3455**	**-0.1080**	**-0.2469**	**0.4346**	**-0.2046**
wto	**0.0552**	**0.0858**	**1**	**0.1082**	**0.0391**	**-0.0857**	**-0.1587**	**0.2473**	**-0.1030**
gdp	**0.2735**	**0.2787**	**0.1082**	**1**	**0.9249**	**-0.0978**	**-0.3730**	**0.4430**	**-0.1834**
gni	**0.3335**	**0.3455**	**0.0391**	**0.9249**	**1**	**-0.1081**	**-0.4629**	**0.5418**	**-0.2737**
growth	**-0.1028**	**-0.1080**	**-0.0857**	**-0.0978**	**-0.1081**	**1**	**-0.1110**	**-0.0705**	**0.0317**
time	**-0.2395**	**-0.2469**	**-0.1587**	**-0.3730**	**-0.4629**	**-0.1110**	**1**	**-0.4620**	**0.7631**
noex	**0.4555**	**0.4346**	**0.2473**	**0.4430**	**0.5418**	**-0.0705**	**-0.4620**	**1**	**-0.3278**
cost	**-0.2037**	**-0.2046**	**-0.1030**	**-0.1834**	**-0.2737**	**0.0317**	**0.7631**	**-0.3278**	**1**

Table 2: summary of sample correlation of all variables

From the table, import is highly correlated with export (the response variable), and noex is moderately correlated with export. Since gni is very highly correlated with gdp (with correlation=0.9249), one of these variables may be

omitted when fitting our models. Here, we choose to omit gdp because gni has a higher correlation with export than gdp. As some of the explanatory variables have relatively low correlation with the response variable, they may be considered for exclusion when fitting models.

Observing the scatter plots of each explanatory variable against the response variable and the linear model residuals , we can see that some of the variables require a transformation. Figures 1-4 below show that residual plots are improved when we take log transformations of variable ex, im, gni and noex. Figure 5 show that the scatter plots of growth, time and cost against the log transformed response variable appear random, asides from heteroscedasticity.

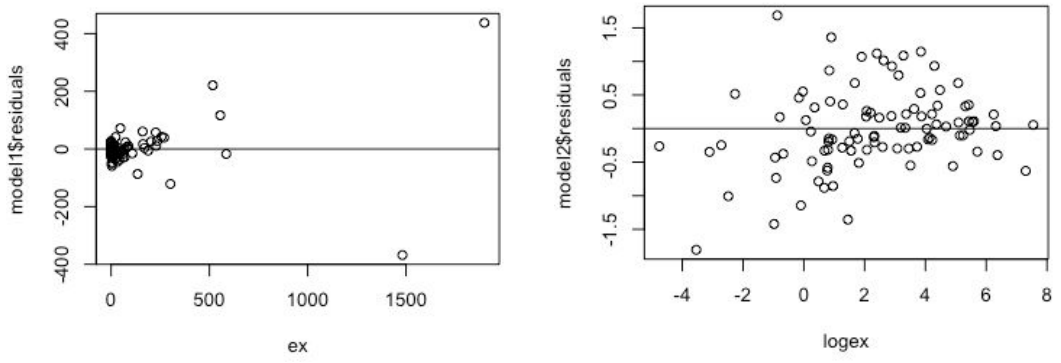


Figure 1: residual plots of ex and logex

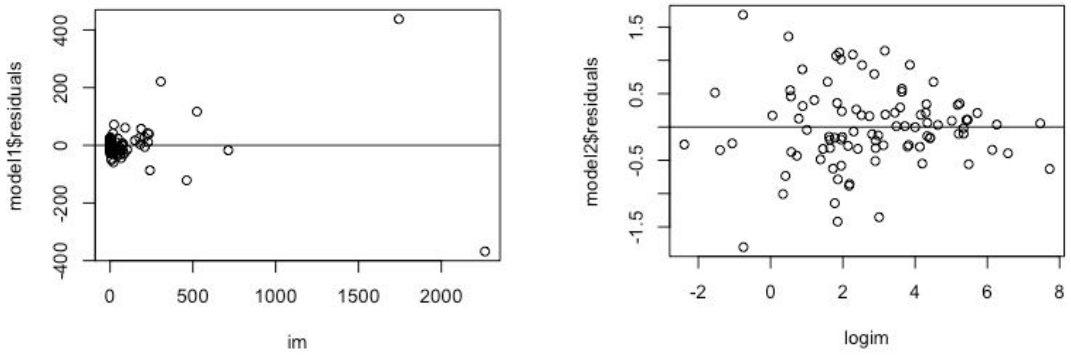


Figure 2: residual plots of im and logim

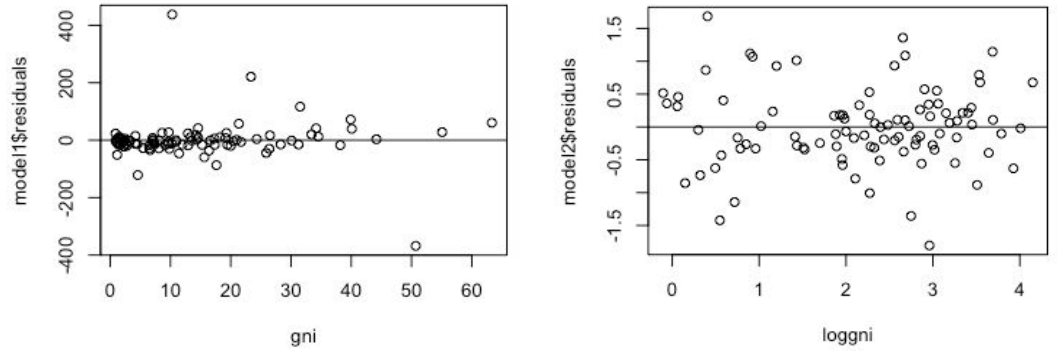


Figure 3: residual plots of gni and loggni

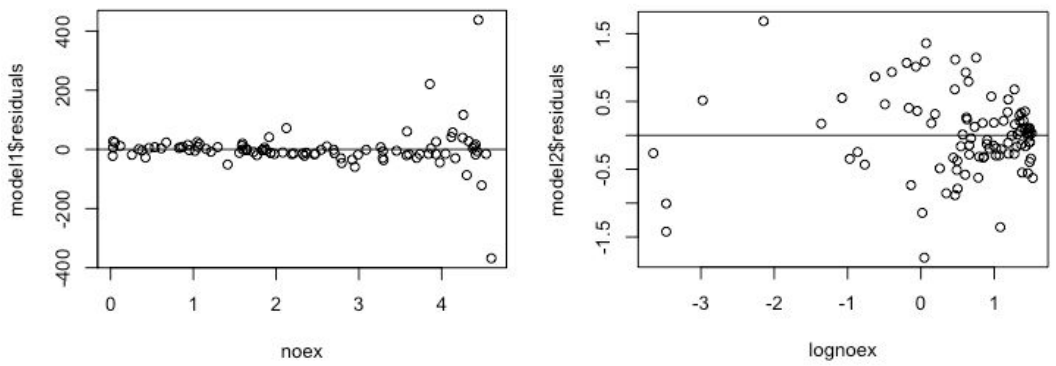


Figure 4: residual plots of noex and lognoex

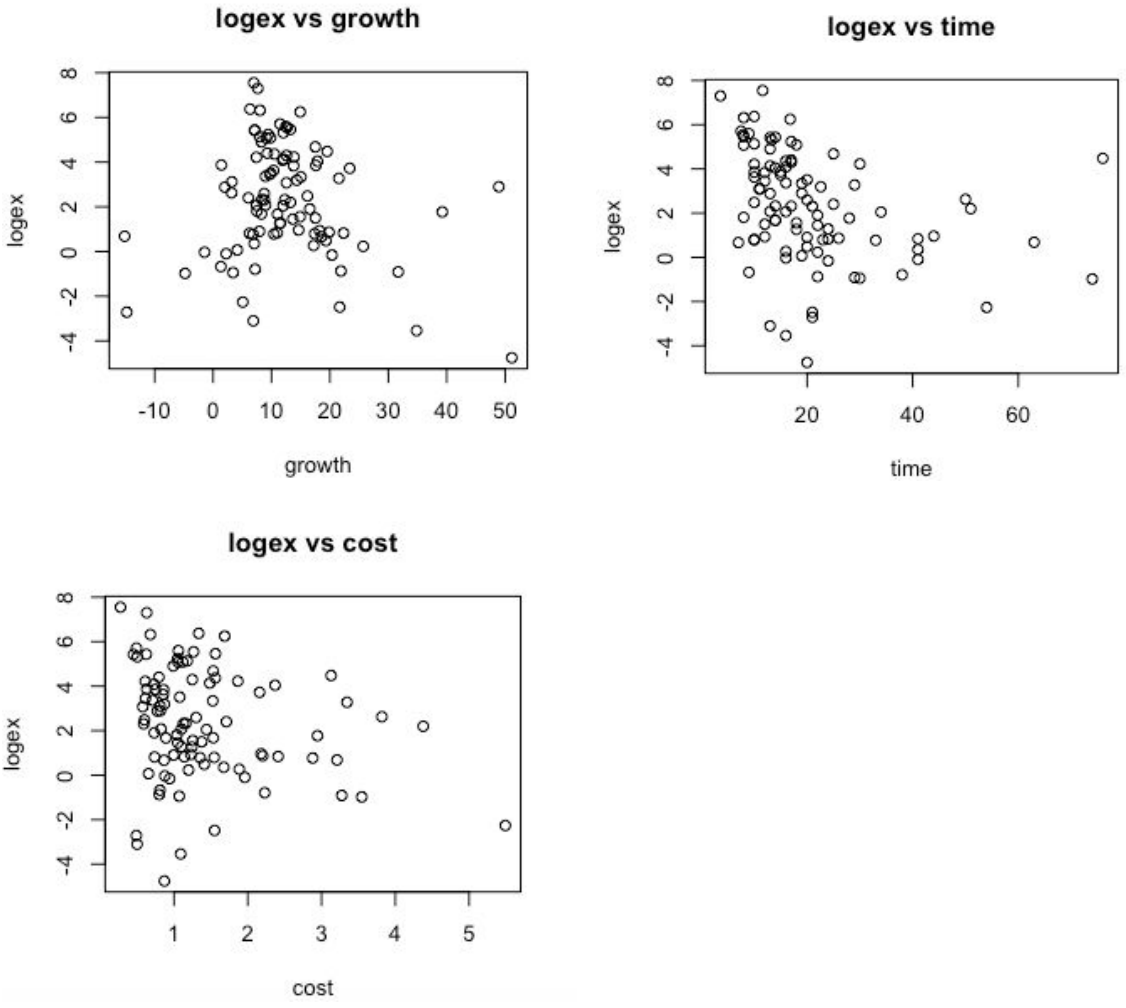


Figure 5: scatter plots of logex vs growth, time and cost

Normal Q-Q plot

Our Normal Q-Q plot performed on original dataset(model1) shows that most points fall along a line in the middle of the graph, but curve off in the extremities, which suggests that most of our data come from Normal distributions except for a few extreme values. However, the normal Q-Q plot for the log-transformed data frame(model2) shows

improvement comparing to the normal Q-Q plot of Model 1. Therefore, the log-transformations of some explanatory variables(ex, im, noex, gni) improve our regression model. Also, the larger adj R2 of model2 compared to model1 indicates that the log-transformations are reasonable.

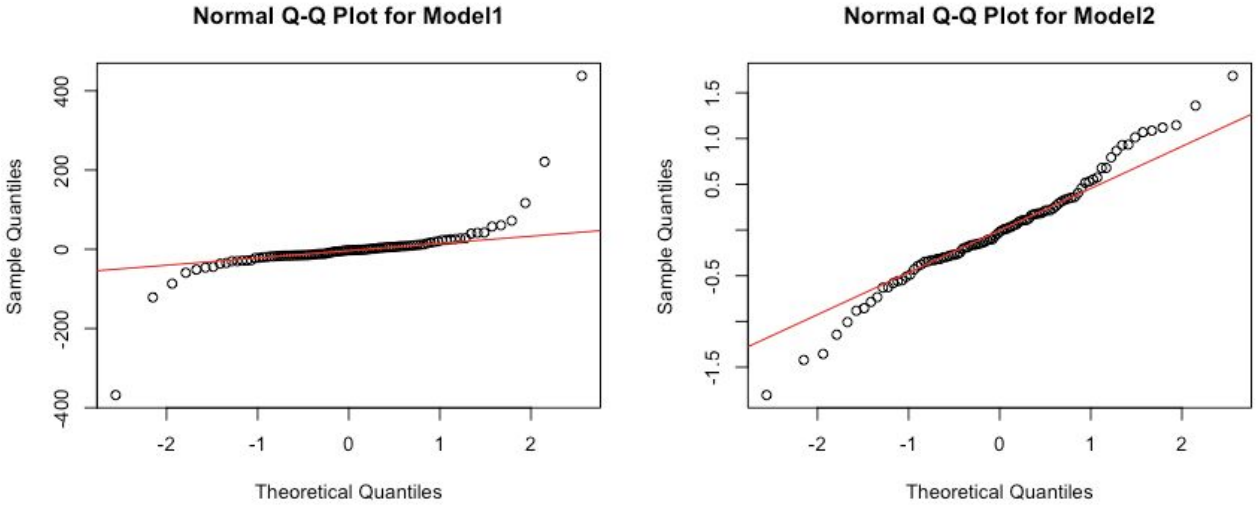


Figure 6: normal Q-Q plots of residuals for Model1 and Model2

3.Variable Selection Methods

We use the leap package in R to perform exhaustive and backward selection methods to select variables. In addition, we calculate the *Cp* and *adj* R^2 to determine as numeric criteria in which we judge our models on. The model with the highest *adj* R^2 and the lowest Cp value is the best fitted one. From the output in R, we choose models 3, 4 and 5 which have moderately high *adj* R^2 and low Cp values.

Model 3: logex ~ logim + lognoex + cost

Model 4: logex ~ logim + lognoex + wto + cost

Model 5: logex ~ logim + lognoex + wto + loggni + cost

	Criteria	Model 3	Model 4	Model 5
Exhaustive Selection	Cp	1.5415	2.4367	4.1736
	Adjusted R2	0.9393	0.9394	0.9389
Backward Selection	Cp	1.5415	2.4367	4.1736
	Adjusted R2	0.9393	0.9394	0.9389
	AIC	-88.8982	-88.0909	-86.3773
	Residual SE	0.6136	0.6131	0.6156

Table 3: the Cp value, adj R^2, AIC and residual SE for models 3, 4 and 5 using exhaustive and backward selection methods.

From table 3, model 3 has the lowest Cp value while Model 4 has the highest adj R2. In the backwards selection, Model 3 has the lowest AIC while Model 4 has the lowest residual SE. Overall, Model 3 and Model 4 are better fit than Model 5.

4.Out-of-Sample Cross Validation

We use the two-fold cross validation method to compare models 3, 4 and 5.

We randomize our initial sample and divide them equally into two sets-- training set and holdout set. We fit each model in the training set and predict on the holdout set. In addition, we use R to calculate the cross-validation root mean square error to compare each model. The model with the lowest CV $RMSE_{holdout}$ is the best regression model.

	Model 3	Model 4	Model 5
CV $RMSE_{holdout}$	0.5678	0.5640	0.6132

Table 4: the CV $RMSE_{holdout}$ for models 3,4 and 5.

From the above table, we can see that Model 4 has the lowest CV $RMSE_{holdout}$ value compared to Model 3 and Model 5. Using this criteria, Model 4 is a better fit than Model 3 and Model 5.

After considering both table 3 and table 4, we chose Model 4 as the overall best fit for our criteria, based on high adj R^2, low Cp value, low AIC, low residual SE and low CV $RMSE_{holdout}$.

5.Validity of Regression Model

We will validate the regression equation for the model of our choice, Model 4, by examining its residuals and its normal-QQ plot. The normal-QQ plot indicates that the spread of the residuals is approximately normal, as the vast majority of the residuals fall on or close to the standardized residual line, although the spread at the highest and lowest quantiles indicate that the distribution is slightly heavier tailed than a normal distribution. However, the residuals vs fitted values plot show that the residuals get gradually closer to the fitted values as export value increases. This disputes our assumption that the data is homoscedastic, as the plot appears to indicate that the data is heteroscedastic, though it can be reasonably explained, as smaller countries are more numerous, and tend to have more specialized economies, making their export values harder to predict, whereas larger, more industrialized countries tend to have more standardized economies that are similar to each other, resulting in smaller variance as export value increases.

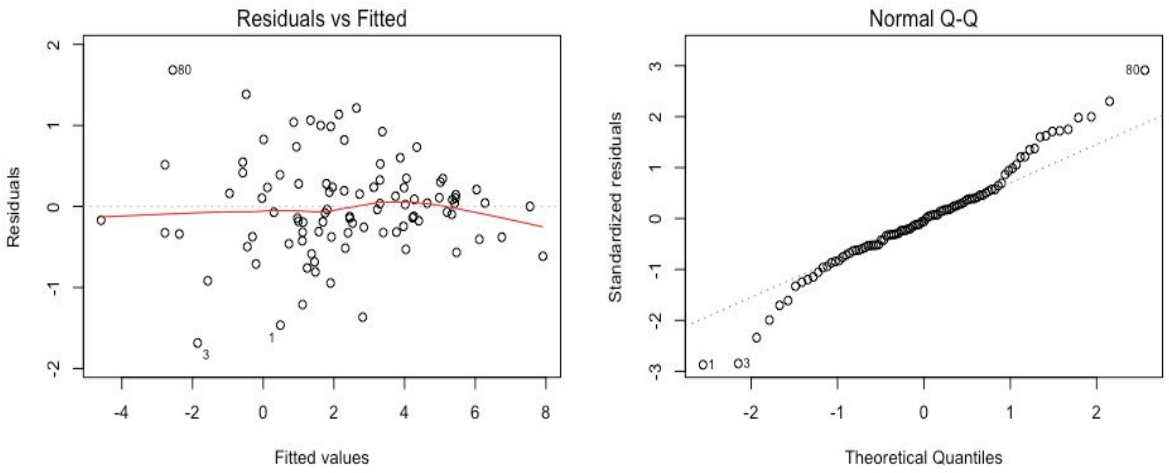
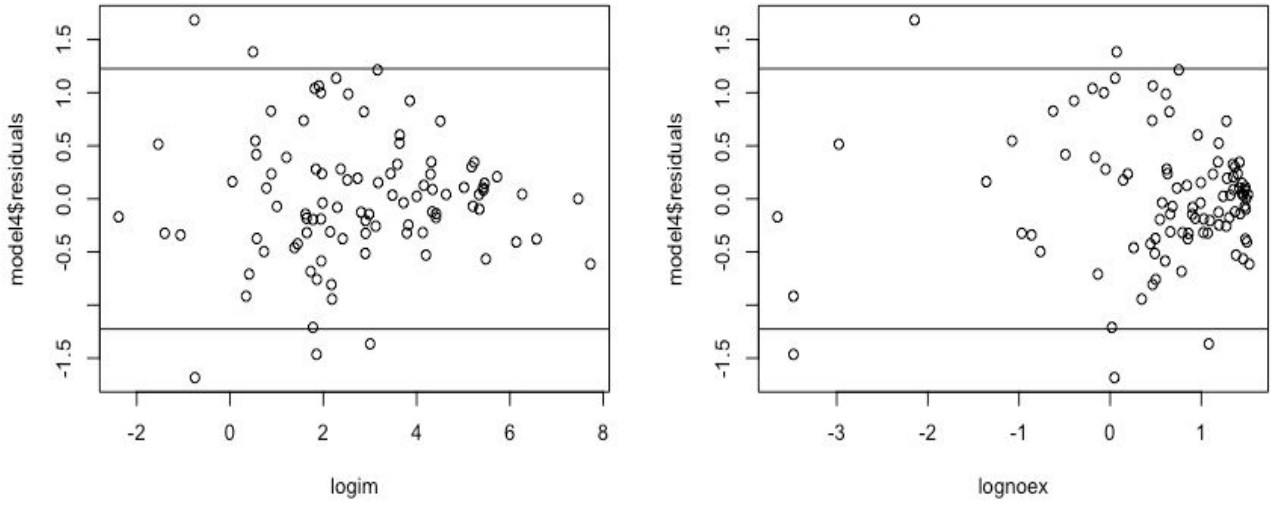


Figure 7: residuals vs fitted value and normal Q-Q plot of residuals of Model 4

Partial Residual Plots for Model 4

Here each variable used on our selected model is plotted against the residuals to verify that the relationships between the variables are linear. Asides from the heteroscedasticity addressed in the previous section, the residual points seem to follow a random pattern, which is indicative of a linear relationship. Each graph has a 95% confidence interval for the residuals constructed and are illustrated by two horizontal lines. Only a handful of points fall outside this confidence interval for each graph, which is reasonable for a sample size of almost 100, which would be around the expected 5%.



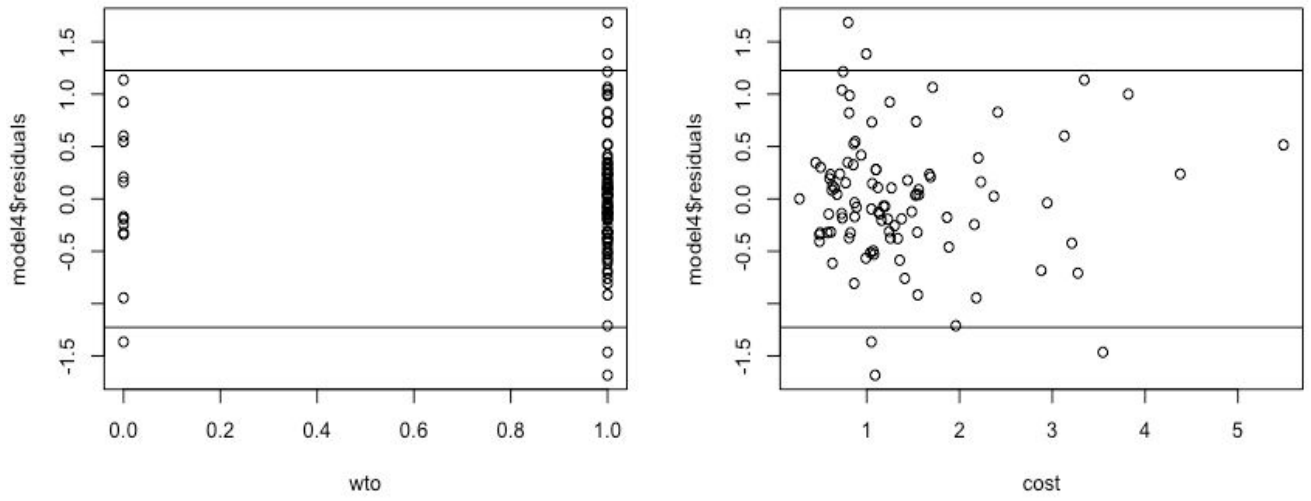


Figure 8: partial residual plots for each variable in Model 4

Significance Analysis

The ls.diag() function was used to obtain Cook's distances and dfits values for each country in our selected model, Model 4. From Cook's distances, using the common criteria that a value over 4/n (0.042) indicates an influential data point, 10 countries were observed to be influential to our model: Afghanistan, Algeria, Antigua and Barbuda, Austria, Central African Republic, Congo, Ethiopia, El Salvador, Lebanon, Maldives, Solomon Islands. These countries generally have high residual values in our model and end up being the most influential, unsurprising, as smaller countries with less orthodox economies would deviate from the norm and influence our fitted model more than large industrial countries. However, a general rule of thumb is that a cook's value over 3 times higher than the mean may indicate an outlier, and out of the 10 countries listed above, all except Congo and Ethiopia fall into that range. The other countries may be considered to be outliers. An analysis of dfits value yielded very similar results. A common criteria for dfits is that a value of over 2*sqrt(p/n) where p is the number of parameters, is an influential data point. Using this criteria, the same 10 countries that were singled out with Cook's distances were again found to be influential, this time with an addition of an eleventh country, Suriname. A plot of Cook's values and a plot of absolute dfits values are shown below.

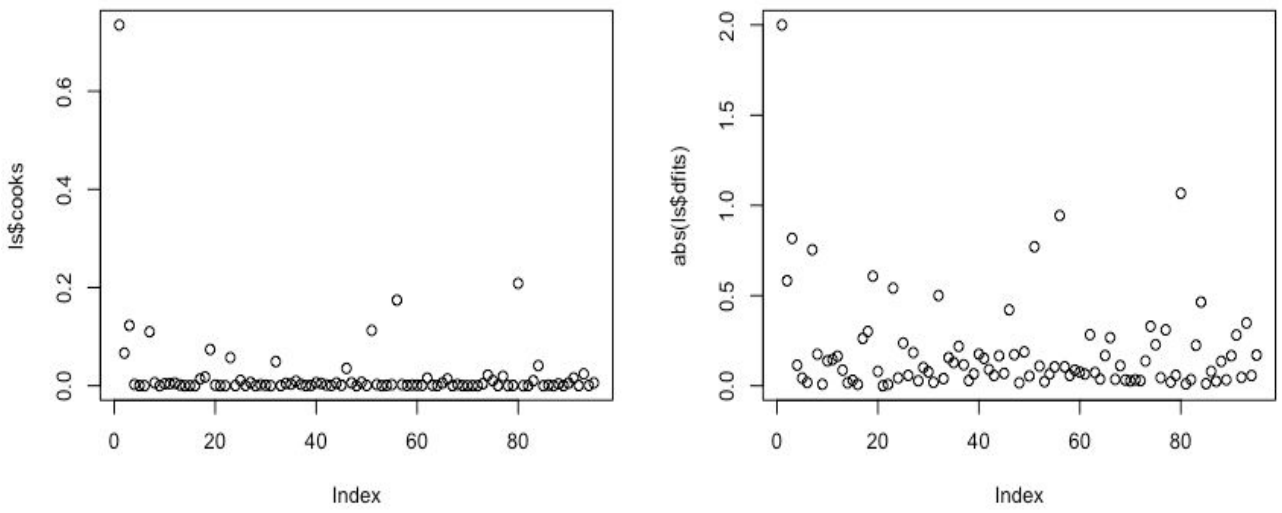


Figure 9: dfits and Cook's distance plots of observations

6. Conclusion

Based on our analysis, the prediction model Model 4 was overall the best fit for our criteria, and the model outputs the following equation to predict national export value:

$$logex = -1.036 + 1.112 \times logim + 0.290 \times lognoex - 0.202 \times wto + 0.188 \times cost$$

After refitting the variables from the full linear model according to the residual plots, Model 4 was the best overall fit for our numerical criteria, compared to Model 3 and Model 5, as Model 4 has the highest Adj R2, the lowest residual SE, and the second lowest Cp value. Furthermore, when we performed 2-fold cross validation on all 3 contender models, Model 4 has the lowest CV $RMSE_{holdout}$ value. Examining the residual vs fitted plot, the normal Q-Q plot, and the partial residual plots for Model 4, they appeared as expected, asides from the heteroscedasticity of the residual plots, which is explained in our analysis as smaller countries having more specialized and varied economies compared to larger countries. The significance analysis showed that 8 countries have particularly high Cook's distances, which indicate that they are possible outliers, and so the prediction equation may be improved by their exclusion. Further explanatory variables that could be explored to improve the prediction model include number of major trade partners, number of ports, labour participation rate, and if the country is a landlocked state.

7. Data Reference

http://wits.worldbank.org/

8. Contributions

Our team was formed through acquaintances and mutual friends. The ordering of the names is done alphabetically by surname. The decisions on choosing topic of the project, researching data, choosing variable selection methods, and choosing cross validation methods were made by the whole team. The major contributions of Yue Chen were writing R code for the whole report and explaining some parts of the report(including data analysis and cross validation). The major contribution of Tao Li was searching data, organizing data frame, inserting and explaining some plots in the report. Yuxuan Han organized all the meetings and provided some major ideas for the project(including choosing variables and variables transformation). James Huang performed the significance analysis in R and wrote the analysis, the abstract, and the commentary on our regression validity and partial residual plots, and did the editing for the report. The major contribution of Xueyang Zhao was asking TAs about the problems our group met when wrote report and code, as well as coming up with suggestions according to the coursepack and writing the conclusion for our project. The final formatting was done by Tao Li.

Appendix

```
> model3<-lm(logex~logim+lognoex+cost,data=newdat)
> summary(model3)

Call:
lm(formula = logex ~ logim + lognoex + cost, data = newdat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72396 -0.34227 -0.01043  0.29790  1.61488

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.19961    0.16850  -7.119 2.44e-10 ***
logim        1.11054    0.04758  23.342  < 2e-16 ***
lognoex      0.27617    0.08760   3.153  0.00219 **
cost         0.18925    0.07285   2.598  0.01095 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6136 on 91 degrees of freedom
Multiple R-squared:  0.9413,    Adjusted R-squared:  0.9393
F-statistic: 486.2 on 3 and 91 DF,  p-value: < 2.2e-16
```

```
> model4<-lm(logex~logim+lognoex+wto+cost,data=newdat)
> summary(model4)

Call:
lm(formula = logex ~ logim + lognoex + wto + cost, data = newdat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.68351 -0.32434 -0.03648  0.27988  1.68447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.03577    0.22794  -4.544 1.71e-05 ***
logim        1.11239    0.04757  23.383  < 2e-16 ***
lognoex      0.28962    0.08843   3.275   0.0015 **
wto         -0.20219    0.18961  -1.066   0.2891
cost         0.18802    0.07281   2.582   0.0114 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6131 on 90 degrees of freedom
Multiple R-squared:  0.942,     Adjusted R-squared:  0.9394
F-statistic: 365.5 on 4 and 90 DF,  p-value: < 2.2e-16
```

```
> model5<-lm(logex~logim+lognoex+wto+loggni+cost,data=newdat)
> summary(model5)

Call:
lm(formula = logex ~ logim + lognoex + wto + loggni + cost, data = newdat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74808 -0.32469 -0.00805  0.29163  1.70511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.12915    0.29127  -3.877 0.000202 ***
logim        1.10407    0.05039  21.909  < 2e-16 ***
lognoex      0.28175    0.09008   3.128 0.002382 **
wto         -0.17662    0.19667  -0.898 0.371587
loggni       0.03900    0.07524   0.518 0.605493
cost         0.19817    0.07568   2.619 0.010380 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6156 on 89 degrees of freedom
Multiple R-squared:  0.9422,    Adjusted R-squared:  0.9389
F-statistic: 290.1 on 5 and 89 DF,  p-value: < 2.2e-16
```