École Polytechnique Fédérale de Lausanne

Econometrics Project

MASTER IN FINANCIAL ENGINEERING

Augustin Kapps, Lucas Giordano, Gabriel Rechsteiner

8 janvier 2021

Abstract

The Gross domestic product, commonly named GDP is an economic metric measuring the market value of all the products and services produced in a specific period for a given country. Every country has a GDP value. But countries can also be characterized by a lot of other metrics : population, exchange rates, current savings, etc. All these metrics might be (or not) related to the value of the GDP : we can naively expect a country with a big population to have a bigger GDP than smaller countries. Our goal is to find these relations and to build a model capable of explaining and quantify them.

1 Introduction

In our study, we will try to relate some metrics with the real gross domestic product CGDP of three different countries, namely : China, South Korea, and Taiwan. We choose these three countries because they have a lot of similarities but also differences. They are countries with fast-growing economy and are geographically relatively close. But they distinguish in size, population and GDP. In addition, the disposable data goes from 1950 to 2000, therefore covers the entire cold war and its economic repercussion. Hence it would be interesting to take countries of different political blocs : Taiwan and South Korea in the western bloc and China in the eastern bloc until 1961, when it became its own bloc after the Sino-Soviet split. The advantage of picking countries with different political regimes is that it also leads to countries with different economic systems over time. For all of these reasons, we think it would be interesting to compare them and their underlying models. As a first step, we will acquire the necessary datasets from Penn World Table and visually study them with a correlation analysis. Then we will build a panel of models using different techniques such as complete OLS, simple to general and so forth. Once these models are created, we will compare them. And then once we have found the best model we will try to improve it, using feature expansion or ridge regression for example and interpret its result.

1.1 Assumptions

- Since we don't know how to create instrumental variables, we will assume that the underlying data generation processes are exogenous.
- We will first focus on linear dependencies between regressors and regressands. Note that this assumption will be relaxed in the 'Model improvement' section when we will use feature expansion to add non-linearity.
- In order to be able to use the least square model, we assume that the data follow a normal law $\mathcal{N}(\mu, \Sigma)$.

Basically, we will first assume that all assumptions (A1 to A6) hold. But we will perform some diagnostic to test them and then try to relax some of them.

1.2 Important informations

Note that we used the Matlab version : R-2019-a update 9. The VIF analysis has a huge impact on our study and it is highly influenced by the method used to compute eigenvalues of a matrix (small eigenvalues slightly change due to floating-point errors). This method can change according to the version of Matlab. We hard-coded some of the results that depend on the VIF analysis in order to make the rest of the study coherent even when using a different version of Matlab.

2 Data Description and Exploratory Analysis

2.1 Data Description

For this assignment, the data was taken from the Penn World Table which provides purchasing power parity and national income accounts converted to international prices. We chose to perform the analysis for the following countries : China, Taiwan and South Korea (named Korea the rest of the rapport). As requested, all of them contain data for at least 12 topics and 40 years, as summarized in table 1. For all countries, we excluded the $CGNP$ feature since it contained missing values that, if removed, would have led to not enough data in terms of years. We also could have interpolated those values but since we already had 17 features (substantially larger than the required 12), we decided not to pursue in this direction.

Table 1 – Summary of Penn Tables data

		China	Korea	Taiwan
Features	number	17	17	17
	excluded	CGNP	CGNP	CGNP
Years	period	52-00	53-00	51-98
	number	49	48	48

2.2 Exploratory Analysis

In this section, due to report size constrains, we will only describe in detail the explanatory analysis performed for the China dataset. Indeed, since the steps taken will be similar for the last two countries, only the results of the analysis will be presented in this report (cf. code for plots and computation).

2.2.1 Descriptive Statistics

Table 2 contains, a statistical summary of each feature of the China dataset (min,max,mean,median,etc.). In addition, the histogram plot (see figure 1) of these features allow us to make the following conclusions :

1. The scales differ significantly between the variables : POP and $CGDP$ (and possibly PI) have very large maximum, mean and variance compared to the rest. This could potentially cause issues during the regression task (give more weight than we should to those features) indicating that we should look into ways to normalize/rescale the data in the modeling part.
2. The statistics of POP are interesting. The mean is $10x$ bigger than the median and the variance surprisingly high (key features of power-law distribution even though the distribution is not skewed). We could argue that taking the log of this feature could possibly help to "standardize" this feature. Historically, these variations could be explained by the fact that the population in China was known to have a rapid growth which leads China to be the largest country in terms of the number of inhabitants (went from 500M to 1.250B in 50 years).
3. We can clearly notice similarities between variables (OPENK/OPENC, PC/PG,CC2/KC, etc.) and we might suspect some possible collinearity issues.

We did the same analysis for Korea and Taiwan showed in figure 4 and 3. We can see that for every country the scale of POP and $CGDP$ are magnitudes higher compared to the other variables. The mean and the median for the population of Korea and Taiwan are approximately the same. This supports the hypothesis that the difference between China's mean and median population is linked to the exceptional growth of China's Population. Finally, by looking at Korea's XRAT, we can see that the variance of the exchange rate is huge. This variance not seen for China and Taiwan comes from the historical changes in South Korea's currency. Between 1953-61 the South Korean hwan was introduced, during this period there was strong inflation. Then the second South Korean won was reintroduced which slowed down the inflation. But then in 1997, the won was devalued to almost half of its value, as part of the 1997 Asian financial crisis. These big changes in the evaluation of Korea's currency are at the origin of the strong variance of Korea's XRAT.

Table 2 – Basic Descriptive statistics for China

statistics	CC	CC2	KC	CSAVE	XRAT	CG	KG	CI	KI	OPENK	OPENC	POP	PC	PG	P	PI	CGNP	CGDP
min	4.42	48.66	46.98	7.14	1.5	16.38	20.39	6.48	5.35	6.13	4.79	569000	19.14	10.69	19.25	29.56	98.19	112.85
max	10.79	75.43	72.67	25.52	8.62	28.76	27.74	23.22	22.63	53.49	48.36	1258821	57.89	41.72	62.8	108.58	100.5	3843.67
mean	6.19	60.3	60.06	16.65	3.49	23.06	24.6	16.24	14.82	17.46	15.77	913862	41.92	26.29	42.93	74.28	99.8	955.51
median	5.25	58.27	59.49	16.51	2.46	24.27	24.62	16.34	14.87	10.15	9.26	93068	48.91	29.23	49.68	87.54	100	386.14
variance	3.69	62.53	45.42	24.55	4.9	13.55	3.42	20.97	25.04	174.55	157.2	47455059600	180.29	111.55	208.56	828.88	0.35	1145975.41
skewness	1.25	0.46	-0.02	-0.09	1.46	-0.46	-0.33	-0.33	-0.14	1.06	1.04	-0.01	-0.48	-0.1	-0.42	-0.38	-1.86	1.29
kurtosis	3.17	1.93	2.06	2.11	3.61	1.85	2.44	1.93	1.85	2.79	2.76	1.66	1.52	1.52	1.5	1.42	5.06	3.44

Table 3 – Basic Descriptive statistics for South Korea

statistics	YEAR	CC	CC2	KC	CSAVE	XRAT	CG	KG	CI	KI	OPENK	OPENC	POP	PC	PG	P	PI	CGNP	CGDP
min	1,950.00	43,872.00	51.7	50.56	-0.27	17.82	5.66	5.55	10.21	7.86	44,016.00	8.51	21264.9	24.36	15.96	23.74	25.82	97.77	268.82
max	2,000.00	49.05	87.68	80.27	41.85	1401.44	12.71	16.74	42.15	41.64	86.31	87.18	47,275.00	83.81	93.13	80.09	73.4	101.04	14936.69
mean	1,975.00	43,975.00	67.81	66.18	22.94	505.12	9.25	9.95	27.35	24.86	29.34	48.59	35264.57	51.36	46.11	49.65	48.37	99.36	910,948.00
median	1,975.00	21.36	63.91	64.16	26.35	484.00	44,021.00	9.47	29.28	25.15	29.55	58.54	36130.5	51.36	45.7	50.13	47.06	99.49	1884.61
variance	221.00	165.65	141.94	85.85	195.82	120500.64	4.96	10.56	96.16	111.09	484.86	547.72	64620488.42	247.67	484.08	228.77	179.04	0.6	25167690.54
skewness	0.00	0.63	0.29	0.15	-0.23	0.35	-0.23	0.47	-0.34	-0.12	0.65	-0.37	-0.24	0.09	0.46	0.09	0.01	-0.03	43,831.00
kurtosis	44,044.00	1.91	1.53	1.53	1.53	2.42	1.66	44,014.00	1.87	1.83	2.65	1.75	1.81	2.14	43,832.00	44,045.00	1.93	2.47	43,953.00

Table 4 – Basic Descriptive statistics for Taiwan

statistics	YEAR	CC	CC2	KC	CSAVE	XRAT	CG	KG	CI	KI	OPENK	OPENC	POP	PC	PG	P	PI	CGNP	CGDP
min	1,950.00	8.27	50.71	54.21	2.61	43,900.00	43,843.00	13.33	8.56	6.74	13.16	20.83	8,255.00	41.47	34.22	44.05	63.43	NaN	180.66
max	2,000.00	55.05	75.52	69.4	34.12	40.06	26.69	31.79	26.63	43,913.00	95.74	106.24	21777.1	95.61	93.19	96.48	144.15	NaN	17742.8
mean	1,975.00	25.34	62.32	61.87	18.13	32.53	19.55	44,033.00	43,999.00	15.46	51.62	68.51	15628.81	60.43	57.18	64.02	87.48	NaN	4258.48
median	1,975.00	20.97	61.97	62.33	22.93	36.59	18.89	43,909.00	17.84	16.92	58.05	85.45	15,855.00	57.77	55.03	60.9	81.88	NaN	1471.81
variance	221.00	221.74	44.2	14.85	91.65	74.54	15.13	37.93	24.42	27.81	844.78	927.2	17549193.49	214.97	367.37	241.29	345.88	NaN	26887836.05
skewness	0.00	0.65	0.34	-0.03	-0.37	-1.08	0.09	0.28	-0.18	-0.33	-0.01	-0.39	-0.19	0.37	0.45	0.23	0.94	NaN	1.23
kurtosis	44,044.00	44,076.00	2.13	2.38	1.81	3.19	1.94	1.59	1.89	44,013.00	1.48	1.45	1.76	2.00	1.93	1.79	3.61	NaN	3.25

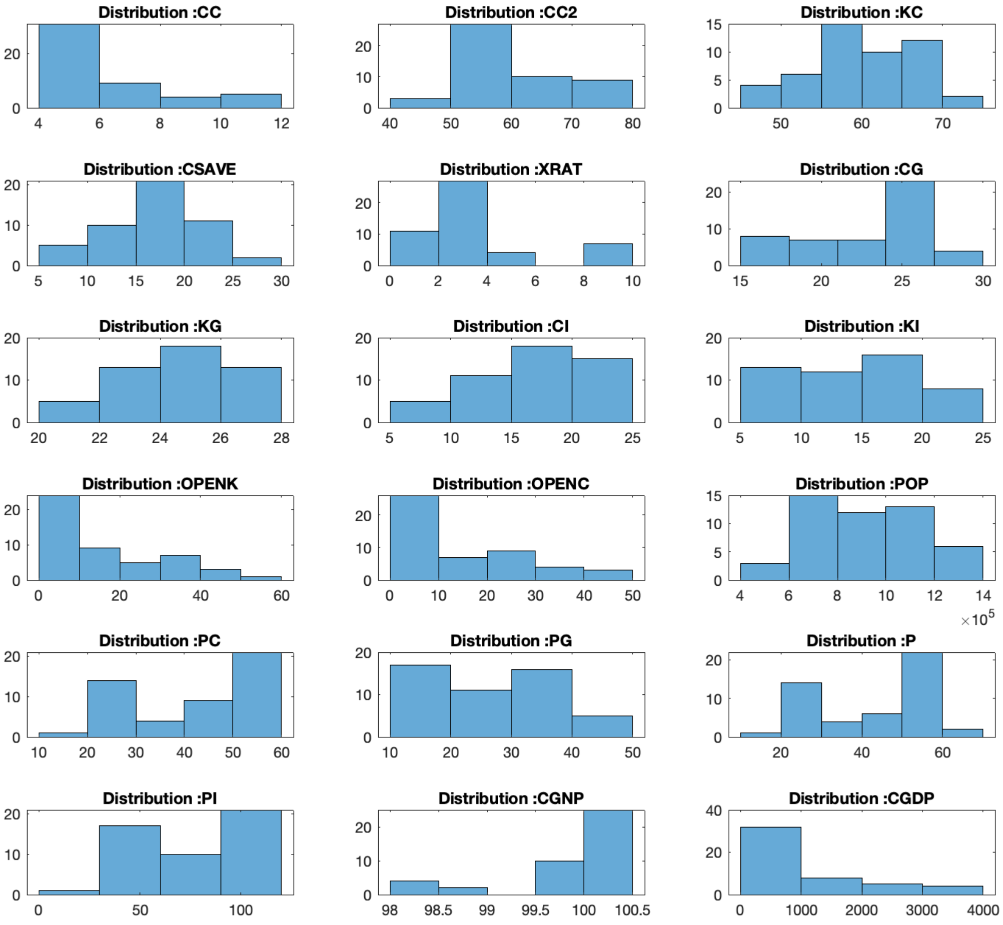


FIGURE 1 – Distribution plot China

2.2.2 Correlation analysis

In the previous section, we noticed some similarities between the features. In this section, we will be looking into ways to quantify this observation. We chose to compute the correlation matrix of the dataset. This is conveniently plotted as a heatmap in figure 2a. Note that the axis corresponds to the feature that we would like to predict.
As expected, this plot reveals a lot of correlated features (for China) :

- the group of variables : $[PC, PG, P, PI]$ are all strongly correlated ($corr \approx 1$) but in the same time strongly negatively correlated with the target $CGDP$ ($corr \approx -1$). At first glance, they seem to constitute a good choice of features when it comes to GDP prediction (if we only consider linear relationships) but we should only take one of the groups to reduce collinearity. In the further model selection section, we will discuss how to make this decision.
- Similar conclusion can be drawn for the group $[CC2, KC]$, on top of the fact that they are

Table 5 – Condition number for the design matrix

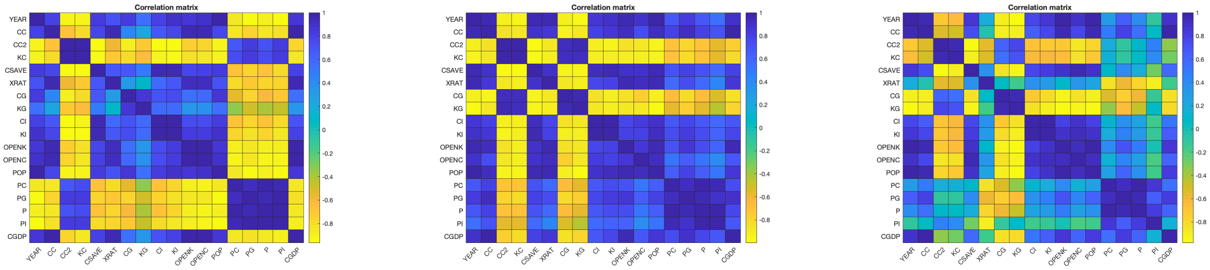
	China	Korea	Taiwan
condition number	5.73e13	1.36e12	7.59e11

also correlated with the previously discussed group at level $0.5 \leq corr \leq 0.9$.

— $Year$ and CC also share a significant positive correlation ($corr \approx 0.8$) but, contrary to the previous group, they are positively correlated which is also good for our prediction task. We should also note that $Year$ is correlated to almost all other variables which were to be expected since they are all time-series variables.

— Generalising the previous points, we understand that all variables seem to be predictive when it comes to $CGDP$, except KG and (possibly) CG.

If we compare these results with the other countries (figure 2), we find the similar conclusion that some variable are highly correlated. In the end, this analyse clearly shows that it would be a good idea to perform feature reduction (though VIF,for example) to reduce the collinearity present in the data. At as last step, we computed the condition number of the design matrices (without CGDP). The result can be found in table 5 were we clearly observe tremendously large values, supporting our analysis.



(a) China (b) Korea (c) Taiwan

Figure 2 – Correlation matrices (last column is GDP)

2.2.3 Scatter plots : features vs CGDP

In addition, an important feature to study is the correlation between the different variables and the GDP. Indeed the stake of the whole model depends on it since we need correlation to predict the GDP from the variables. As we can see in figure 2, we did a scatter plot for every variable with the GDP to evaluate correlation for the case of China. We can see that every variable is correlated with the GDP even if the trend is more clear for some variables than the other (e.g. $YEAR$,POP,CC,$OPENK$, etc.). For the other countries, we observe similar patterns and there is no point in listing them all. Overall, we are confident in the fact that we will be able to come up with a model that explains the GDP to some extent in this analysis.

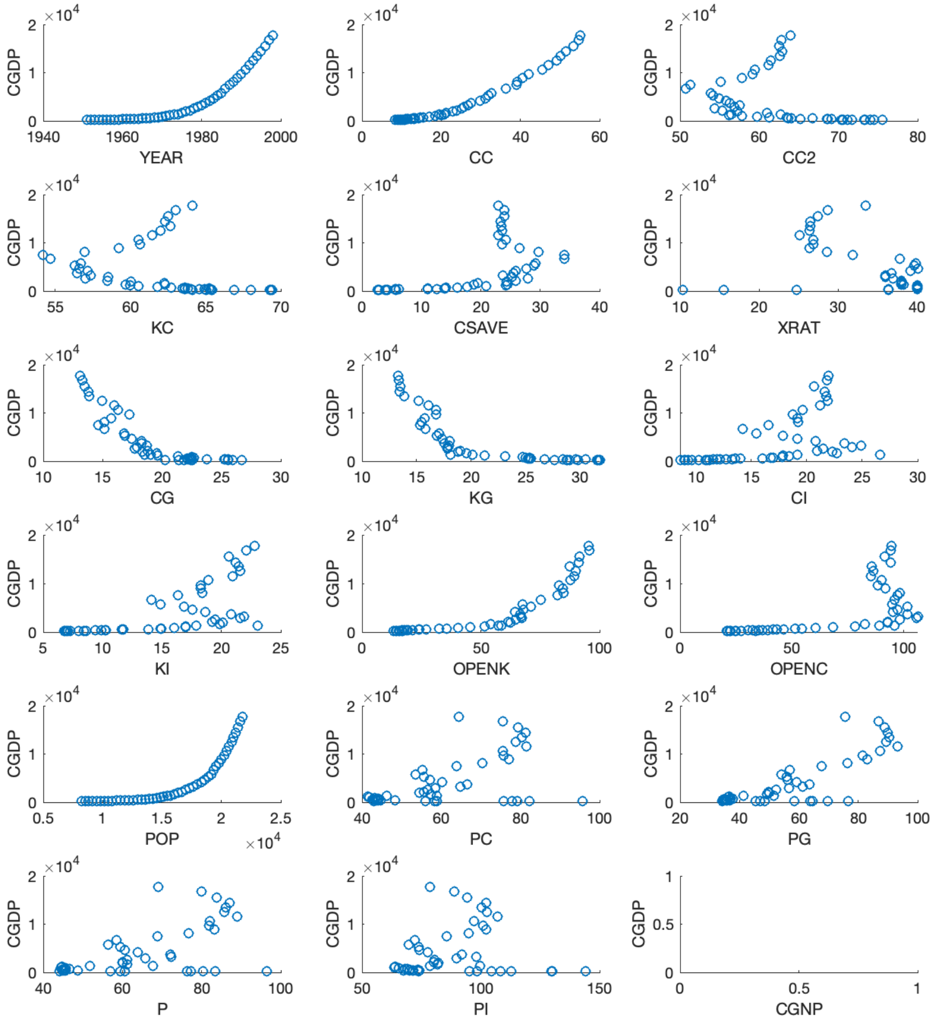


Figure 3 – Scatter plot target (CGDP) vs all features

3 Feature selection

In this section, we will try to reduce the complexity of our model by selecting the subset of the features that contain the smallest number of features while still being able to explain (fit) the dependent variable. This is important for our task since we want to predict the GDP. In other terms, we do not want our model to be too close to our training data. This feature selection will thus help to avoid overfitting as well as reduce collinearity in the design matrix. We want to emphasize that this section only aims at describing the methodology used to select this subset of features not at providing good results in the test-set. In that regard, we will only provide numbers (R^2 for example) to justify the relevance of the introduced methods. The real usage of those will be part of the modeling section. The shared numbers will be generated using a simple OLS method but, later on, when we will study a more complex model ($FGLS$, etc.) we will re-do this analysis using those updated models.

3.1 VIF analysis

In the explanatory analysis, we understood that our design matrix was highly colinear. In order to allow the models that will be introduced in the following section to work correctly, we need to reduce the value of the condition number of its design matrices, since all the models need to inverse theirs. We chose to use the variance inflection factor (VIF) to determine which columns cause collinearity in the matrix. We hope that this procedure will help us select a good subset of the original features such that the overall condition number stays within reasonable bounds that we fixed under 20 as seen in class.

We proceeded as follows : while the condition number of the design matrix is higher than 20, we compute VIF_i, $\forall_{2 \leq i \leq k}$ and we remove the feature leading to the highest of the VIF's.
Surprisingly, only the intercept remained after having applied this procedure (the last condition number for China was 145.6). This clearly indicates an issue with our design matrix. If we take a look back to table 2, we observe a huge difference in terms of variance and scale between the variables. Those differences imply an unfair estimation of the model parameters giving more importance to variables having a large scale. To deal with this issue, we standardized our features :

$$\tilde{X_i} = \frac{X_i - \bar{X}_i}{STD(X_i)} \tag{1}$$

Only by doing this transformation, the condition number for China drop from $5.73e13$ to $3.78e8$, testifying of the relevance of the method. After this small detour, we ran our VIF analysis again and we got the following set of 9 non-colinear variables with a condition number of 18.1 for China :

$$\tilde{C} = \{INTERCEPT, CC, KC, XRAT, CG, KG, CI, OPENK, PC\}$$

It is interesting to note that this method selected only one feature of each of the similar groups presented in the correlation analysis section (for China). We are thus confident that this output is relevant.

One should also note that this standardization process can also be applied when we will split our dataset into train/test tests. Indeed, we have to normalize our training set first and then record the mean and standard deviation. Then for the test-set we have $\tilde{X_i^{(te)}} = \frac{X_i^{(te)} - \bar{X_i^{(tr)}}}{STD(X_i^{(tr)})}$. In that sense, we make sure that the model trained on the training set can also be applied to the test set, without any additional bias.
Applying this method to the three countries gives table 6.

3.2 Best possible model

In this section, we will construct the optimal set of features to use, given the initial set of non-linear features resulting from the VIF analysis. To do so, we can simply compute all the possible model combinations and use the test set R2 to select the best model (optimality criterion for our prediction task). This would give us an upper bound for the best fit we could hope to get. This procedure results in table 7.

Table 6 – VIF analysis : condition numbers and selected features

	China	Korea	Taiwan
condition number	18.11	18.47	8.83
selected features	INTERCEPT, CC, KC, XRAT, CG, KG, CI, OPENK, PC	INTERCEPT, CC2, CG, CI, OPENK, OPENC, PC, PG, PI	INTERCEPT, CC, CSAVE, CG, CI, PI
number	9	9	6

Table 7 – Optimal set of features

		China	Korea	Taiwan
Opti.	selected features	INTERCEPT, CC, KC, KG, CI	INTERCEPT, CG, OPENK, OPENC, PC, PG	INTERCEPT, CC, CSAVE, CG, PI
	number	5	6	5
	Test set R^2	0.9303	0.7010	0.8927

3.3 Model complexity reduction

We plan to tackle this problem by first using the model sections procedure seen in class :$simple - to - general$ and $general - to - simple$. Using the R^2 selection criteria, we got table 8.
We can observe that China is the only country for which this model selection procedure leads to a higher R^2 on the test set. However, at least for China, we would hope to be able able to achieve at least a positive R^2 on the test set. That's why we introduce a third model selection technique derived from the General to Simple approach. Indeed, it seems that there is still overfitting present for China even after the selection of features. We hope to be able to reduce this by doing the following : iterate while there is p-value greater than a given significance level α and fit a model then remove the column leading to the highest p-value (if $p_{values} < \alpha$) on the fitted model. Therefore, this procedure will greedily remove all features that are not statistically significant. After having performed this, we got table 9. We see that when there was no overfitting present before this procedure might be overkill but for China, it worked well.

3.4 Vuong's test

We ran a Vuong test on every possible pair of models for each of the selected countries to see it could help us to discriminate some of them. Unfortunately for China and Taiwan, the statistics were not extreme enough to reject any of the null hypotheses (equivalence of models) thus we didn't discard any model using this method. For Korea The p value and the general to simple model (g2s) are both equivalent and better than the others.

Table 8 – S2G and G2S approaches

		China	Korea	Taiwan
All	selected features	INTERCEPT, CC, KC, XRAT, CG, KG, CI, OPENK, PC	INTERCEPT, CC2, CG, CI, OPENK, OPENC, PC, PG, PI	INTERCEPT, CC, CSAVE, CG, CI, PI
	number	9	9	6
	Test set R^2	-0.6917	0.3899	0.7116
S2G	selected features	INTERCEPT, CC, CG, KG, CI, OPENK, PC	INTERCEPT, CC2, CG, PG	INTERCEPT, CC, CSAVE, CG, CI
	number	6	4	5
	Test set R^2	-0.4189	-7.025	0.6573
G2S	selected features	INTERCEPT, CC, KC, CG, KG, CI, OPENK, PC	INTERCEPT, CC2, CG, OPENK, OPENC, PC, PG	INTERCEPT, CC, CSAVE, CG, CI
	number	8	7	5
	Test set R^2	-0.1999	0.3376	0.6573

Table 9 – P-value selection approaches

		China	Korea	Taiwan
p-val	selected features	INTERCEPT, CC, XRAT, CG, OPENK	INTERCEPT, CC2, CG, OPENK, OPENC, PC, PG	INTERCEPT, CC, CSAVE, CG, CI
	number	5	7	5
	Test set R^2	0.7938	0.3376	0.6573

4 Models

In this section, we will study and apply different models to our datasets and study which assumptions are valid or not.

4.1 OLS on the complete dataset

As a first step and to have a comparison basis, we will simply try to fit an ordinary least square model on the complete dataset (we don't remove any feature) and study the result. First, we split the datasets into a training and testing set (20 %). Then, using the OLS method that we implemented we obtained three (on each country data) fits having the following R2 scores on the train set (figure 10) :

We can see that our model explains well the variance of the target variable within the train set. But the whit is more interesting is the distribution of the residuals produced by this model. Note that a high value of R2 doesn't mean that our model corresponds to the real underlying model (see the testing R2 in the figure **??**). In our case, the OLS is probably overfitting the data.

Table 10 – OLS R2 scores (complete dataset)

Country	China	Korea	Taiwan
Training R2	0.9991	0.9988	0.9978
Testing R2	-0.1751	-3.2596	0.4775

4.2 OLS on reduced dataset and basic assumption analysis

In the last section, we saw the initial set of feature that we are given is expressive enough to be able to accurately model the GDP. However, as seen in the VIF analysis section, we need to reduce this set to a subset in which collinearity does not occur. From this point onward, only features resulting from the VIF will be used (for each country). Note that this selection will obviously reduce the overfitting issue. The purpose of this section is to understand which least square assumptions are valid for our dataset. In order to do that, an OLS model is fitted. We will first look at :

Table 11 – OLS R2 scores (reduced dataset)

Country	China	Korea	Taiwan
Training R2	0.98	0.96*	0.99
Testing R2	-0.69	0.38	0.71

1. Normality of residuals
2. Homoscedasticity

In figure 4, we first plotted the distribution of the residuals. The QQ-plot directly tells us that the residuals are likely to come from a standard normal distribution. This is good news : if the other assumptions (homoscedasticity, non-autocorrelation, exogeneity) holds then we have an unbiased and consistent estimator. We also performed this check for various different subset of features and countries and we arrived at the same conclusion.

On the bottom part of the figure, residuals against $CGDP$ and $Year$ are plotted (China). It allows to diagnosing if homoscedasticity is a valid assumption to make. Indeed, if homoscedasticity was true, we would observe no changes in the variability of the residuals when the value on the $x - axis$ varies. However, it is not the case. We do not see a clear pattern but we can still observe higher variability when CGDP grows (residuals are more concentrated around 0 when CGDP is low) and similarly for $year$. Therefore, we suspect heteroscedasticity in our data. Note that again this pattern is recurrent across countries and we will not include all the plots in this report. As seen in class, the Breusch Pagan test can be used to formalize this intuition. We got that for each country the test rejected the null hypothesis (homoscedasticity). For instance, for china, we had $BP_{stat} = 12.69 > 3.32 = BP_{crit}$.

In conclusion, for the next steps, it will be required to use a more general approach to deal with the negative effects of heteroscedasticity.

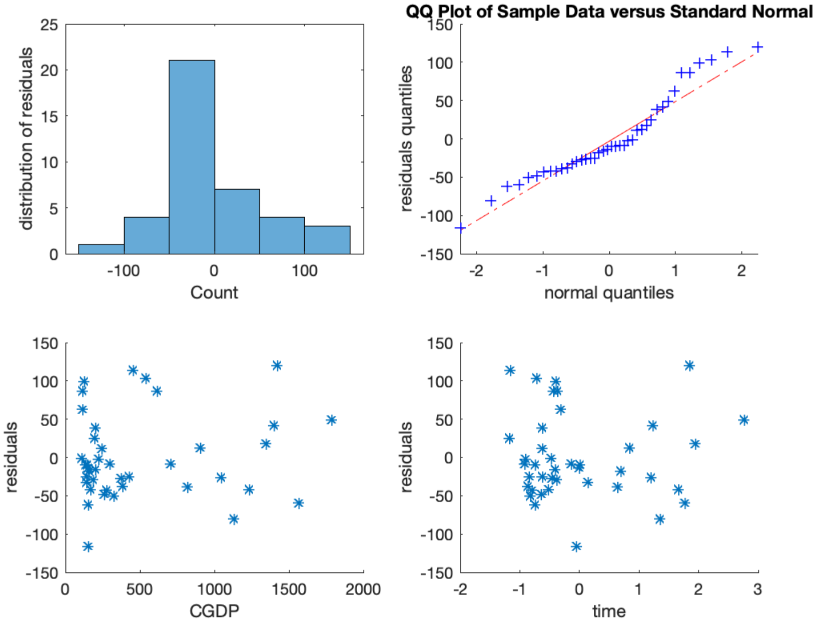


Figure 4 – China : assumptions

4.3 General least squares

As we saw in class, the general way to lift the homoscedasticity assumption is to use a transformed model so that, in the transformed space, the latter holds and a simple OLS can be applied. This method is called *general least squares*. This procedure usually requires to come up with an estimator of the covariance matrix Ω of the residuals. However, this is a rather cumbersome task and we will impose some structure on this matrix to make a more efficient estimator (FGLS). We will look at the following approaches (from the course material) :

1. Standard : $\hat{\Omega}$ is diagonal with $\hat{\Omega}_{ii} = \hat{\epsilon}_i = e_i^2$
2. Heteroscedasticity. : $\hat{\Omega}$ is diagonal with $\hat{\Omega}_{ii} = \hat{\epsilon}_i = e_i^2 = exp(z_i\theta)$ where θ needs to be estimated. In this set up, let's take $z_i = \log x_i$ (found online).

Let's follow the same training procedure as for the previous model but using a feasible generalized least squares method. We obtain the table 12.

Table 12 – FGLS scores

		China	Korea	Taiwan
Initial	BT stat	12.69	18.08	3.10
	BT crit	3.32	3.32	1.63
Standard	BT stat	4.5	4.5	3.00
	BT crit	3.32	3.32	1.63
Heterosc.	BT stat	39.30	11.72	4.42
	BT crit	3.32	3.32	1.63

We observe that for all methods tried (and all countries), we always rejected the null hypothesis (homoscedasticity) with the Breusch Pagan test. This means that we failed to come up with a model of the covariance matrix for which in the transformed model a simple OLS can be applied (assuming homoscedasticity). We decided to stop looking for more complex models in this direction since we already used 2 different approaches for the structure of Ω and that we still have other areas to cover in this report. As the last step, we also compared our results with the $fgls$ method of the Matlab library. Of course, we had a slight variation in the model parameters found but that's also due to the fact that the algorithm used is not the same. Indeed, we look at the source code of the function and we saw that they used a QR matrix factorization to estimate Ω which is something clearly different than what we did in our code. Despite that point, we found out that even with the parameters estimated using that function, we were not able to get rid of heteroscedasticity. There could be several explanations for that :

- The Breusch Pagan test is not limited to a particular form of heteroscedasticity. Indeed it assumed a parametric form for the residuals ($f(\alpha_0 + \alpha^T z_i)$) which, in theory, could be anything. Therefore, it could simply mean that the form hidden in our dataset is not one of those we tested.
- With the concepts seen in class, we are not competent enough to design the right form of the structure of Ω.
- We have to think carefully about the standardization process applied in the section on collinearity reduction (equation (1)). Indeed, let's write this transformation using the diagonal matrix

$$D = \begin{bmatrix} STD(X_1) & & \\ 0 & \ddots & 0 \\ & & STD(X_k) \end{bmatrix} \tag{2}$$

using D, equation (1) can be written in matrix form as

$$\tilde{X} = (X - \mu)D^{-1} \tag{3}$$

the OLS estimator of the latter matrix is

$$\tilde{\beta} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y = (D^{-1}X^T X D^{-1})D^{-1}X^T y \tag{4}$$

we almost recognize a GLS estimator. Indeed, if $D^{-1}X^TXD^{-1} \approx X^TD^{-2}X = X^T\Omega^{-1}X$, we would have a sort of GLS. Even if it is not true here, this gives us insights that the transformation applied to have helped reduce the heteroscedasticity in the first place, implying that we were not able to do that better in this section.

The last point raises one question, what if we ran a Breusch Pagan test on the result of the OLS without normalization the data first, would be getting a higher statistic value ? Actually, this is not the case, both methods yield the same BT test value 12.69, meaning that they are equivalent in terms of heteroscedasticity.

Furthermore, we should note that the analysis performed here can not be generalized to the test set. Indeed, the $\hat{\Omega}$ matrix is of size n_{train} by n_{train} and there is no direct way to apply this transformation to the test set which has dimensions n_{test} by n_{test}. Nevertheless, it was still interesting to study the effect of this FGLS method on this dataset.

4.4 Transformed models

As a different approach to the heteroscedasticity problem, we will try to perform an ordinary least square on transformations of the inputs. By doing so we hope that impact of the residual heteroscedasticity will be reduced. More precisely we will try 3 different types of transformation and for each country, we select the one that produces the most interesting results. These transformations are the following :

— LogLog : take the log transformation of the target variable y and the features X
— LinLog : take the log of the features and keep the target variable
— LogLin : take the log of the target variable and keep the features

Note that, to avoid computational (log of negative values) problems we subtract the min values and added 1 to all variables subject to the transformation. And we only used features obtained by the VIF analysis. During the analysis, we also test the obtained model on the test set (we look at the R2 score). To be consistent we need to apply the exact same transformation on the test set as the one applied on the train set. Namely, we have to subtract the minimum values (and add 1) of the train set to the test set before applying the log.

1. For the country of China, we visually deducted (from the residual graphs) that the log-log transformation was the one that produced the best results. On figure 5d we indeed see that the transformation has improved homoscedasticity (compare to figure 5a). However, after applying the transformation we wanted to see how predictive our model his so we tried it on the test set but we obtained an R2 score of -0.31.

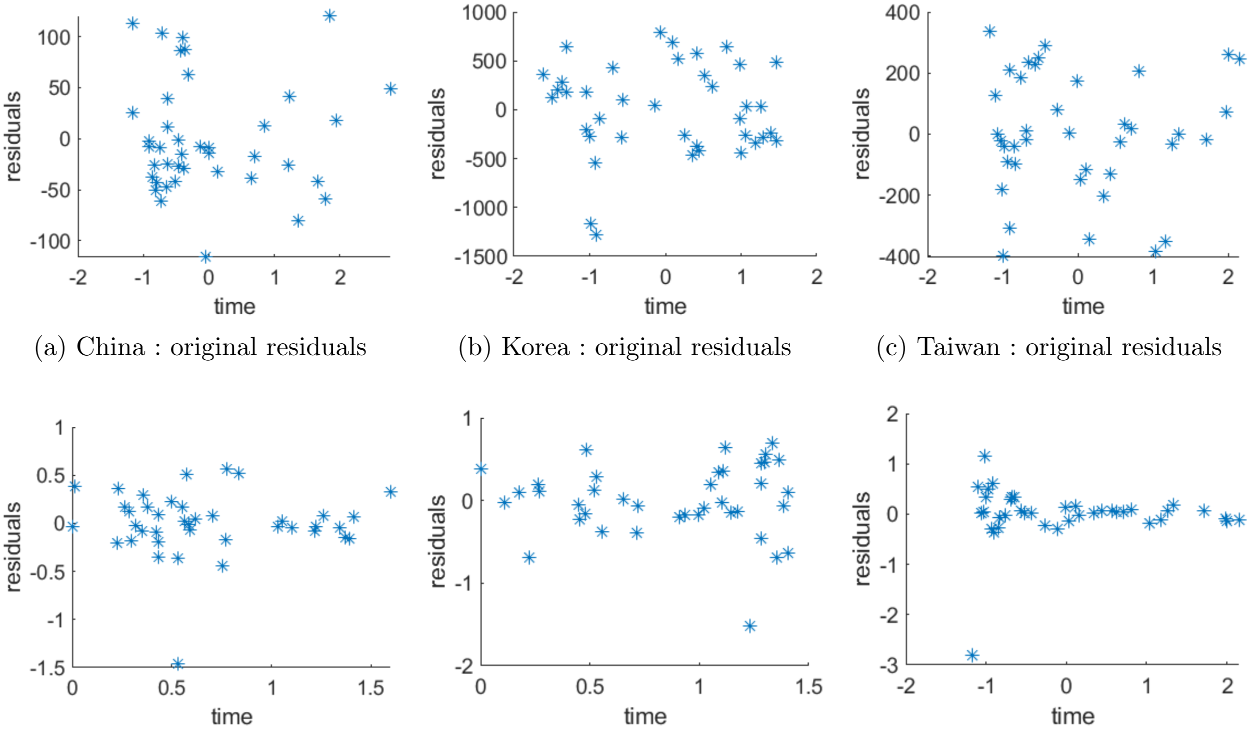
2. The best result that we obtained for Korea was obtained using the LogLog transformation. But this result (figure 5b) is not satisfying at all. We can see some kind of waves in the plot of the residuals what is strong evidence of heteroscedasticity. Moreover, the R2 score obtained on the test set was -1.87 which is low.

3. For Taiwan we obtained the best result using the LogLin transformation. But we see on the obtained residual graph (figure 5f) that the variance of the residuals is higher at the beginning than at the end. Thus the transformation didn't improve homoscedasticity. Again the R2 score on the test set wasn't satisfying (0.29)

We conclude from the obtained results that the log transformations are not satisfying in any of the chosen countries. Thus we will go back to our original set of data.

4.5 Robust Confidence intervals

In the last sections, we tried several approaches to deal with heteroscedasticity, without success. Hence, we decided to keep our original OLS estimator after the previous analysis. However, since we know that heteroscedasticity is present in our dataset, we need to take that into account when the confidence intervals for the parameters are computed. Indeed, we can no longer use the standard confidence interval computation that we used earlier in this report. Instead, we will use *White's* heteroscedastic consistent estimator of the asymptotic variance to achieve our goal.



(a) China : original residuals (b) Korea : original residuals (c) Taiwan : original residuals

(d) China : LogLog residuals (e) Korea : LogLog residuals (f) Taiwan : LogLin residuals

Figure 5 – Log transformation models

5 Model improvement

Now that we found an efficient model we will try to improve it. For now, we have assumed that the data were generated by a linear process, it's time to revoke this assumption. In order to deal with a non-linear generation process, we need to introduce non-linearity in the features that we are using. To do so we will use a polynomial feature expansion.

5.1 Feature expansion

In this section, we studied how feature expansion can improve our models. We expanded the best model working so far for each of the chosen countries. We considered an expansion of degree 2 and including component interactions. Namely, we add the square of each feature and all the possible pairwise multiplication between features in the dataset. after this process the obtained set of data is quite large, we perform a p values selection (select significant features) on the columns to reduce it. As a result, we obtain table 13. In this table, we saved the names of the obtained columns and the obtained R2 score computed on the test set. We see that the only satisfying result is the one obtained for Taiwan.

Table 13 – Expanded features selection

	Korea	China	Taiwan
Selected features	$INTERCEPT$ $CC2$ CG $OPENK$ $OPENC$ $OPENK \cdot CC2$ $OPENK \cdot CG$ $OPENC \cdot CG$ $OPENC \cdot OPENK$ $PC \cdot CC2$ $PC \cdot OPENK$ $PC \cdot OPENC$ $PG \cdot OPENK$ $PG \cdot OPENC$ $PG \cdot PC$	$INTERCEPT$ CC KC CG KG $OPENK$ CC^2 KC^2 CG^2 KG^2 CI^2 $OPENK^2$ $CG \cdot CC$ $CG \cdot KC$ $KG \cdot CC$ $KG \cdot KC$ $CI \cdot CC$ $CI \cdot KC$ $CI \cdot CG$ $CI \cdot KG$ $OPENK \cdot CG$ $OPENK \cdot KG$ $OPENK \cdot CI$ $PC \cdot CC$ $PC \cdot KC$ $PC \cdot KG$	INTERCEPT, CC, CSAVE, CI, CC^2, $CSAVE^2$, $CSAVE \cdot CC$
R2	-22.72	-11.48	0.91

5.2 Structural breaks

If we look at the raw datasets, we observe that for China the feature $XRAT$ stays constant in the period $1952 - 1972$ which let us think that we might have a structural break in 1973. Let's test that. For this experimentation, we need to select a relevant model that we know contains the desired feature. If we go back to section 3.3, we saw that the p-value selection model had a good test R^2, let's pick this one. As done in assignment 6, two methods are used :

1. Include a dummy variable that is zero in the range $1952 - 1972$ and one after and perform a T-test. The result is a $p - value = 0.0011 < \alpha = 5\%$. Therefore, we reject the null.
2. Perform a F-test. Again, we are able to reject the null $F - stat = 12.34 > F - crit = 2.53$.

Since both methods gets the same result, we can indeed state that there a structural break at year 1973 for China. This structural break can be historically supported by the economic reform that occurred in China in the same period of time. Indeed the economical regime changed in 1978 under Xiaoping and from 1978 until 2013, unprecedented growth occurred, with the economy increasing by 9.5% a year. To compare between 1950 to 1973, Chinese real GDP per capita grew at a rate of 2.9% per year on average

Surprisingly, for Taiwan, the $XRAT$ column also seems not to vary significantly during $1961 - 1977$. We use the same approach, will the full model defined in section 3.3. Indeed, since $XRAT$ is not part of the subset of features returned by the VIF analysis, we want to have a model that can to capture the fact that $XRAT$ is missing. Moreover, since the considered period starts at $1961 > year_0$, two structural break tests are required. The result is the following :

- 1961 : the t-test fails to reject while the F-test does.
- 1977 : the t-test fails to reject while the F-test does.

Therefore, we can reject the null and stay that a structural break exists. However, we are not as confident as we were for China. But we should at lease say that between the year 50 to 60 Taiwan got massive help from the US. And Taiwan became a democracy in 1978. These dates match with the structural brakes and could maybe explain why the exchange rates jumped at those points, especially in a world market run by the US.

Note that we manually looked at the data to find these structural breaks before testing them, we didn't found any for the data coming from Korea but there might exist one that we missed. In addition to the significance test of the structural break, we wanted to check the explaining power of the new models (including the dummy variable). To do so we simply created the models and tested them on the test set. We obtained an R2 score of 0.55 for China, this score is smaller than the one obtained without including the dummy variable of the break which was 0.79 (figure 8). This might be due to the fact our model overfits the training data when using this new variable. Similarly for Taiwan : we obtained an R2 score of 0.69 it is good but not as high as the one obtained by simply using all the features (after VIF analysis) which was 0.71 (figure 8). Again, we suspect the model overfits the training data.

5.3 Ridge regression

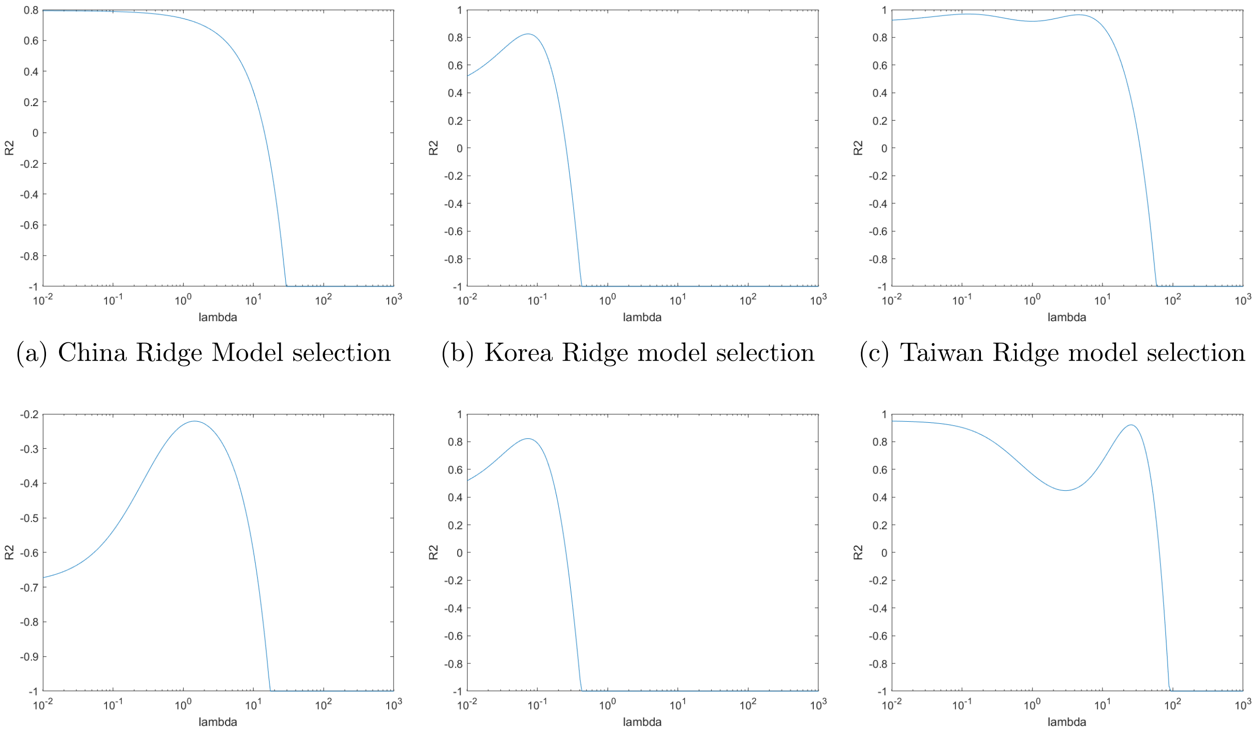
In all the previous analysis, the R^2 that we got on the train were in the order of 0.95 while on the test set we were only able to get a value in the range $[0.39, 0.89]$ depending on the country and the model chosen. Clearly, we have an overfitting problem. As seen in other classes, one typical way to have a better bias/variance trade-off is to use regularization. For this particular analysis, we will implement $L2$ regularization (aka ridge regression) because the optimization problem has a nice close form solution just like ordinary least squares. Mathematically, the idea is to penalise the MSE loss ($MSE = \sum_i (y_i - X\beta)^2$) by a factor $\lambda||\beta||^2$. It has the advantage of shrinking the β to lie in a convex square ball of radius λ centered at 0, hence reducing the complexity of the model and therefore reducing the bias and overfitting. The ridge regression estimator aims to solve :

$$\hat{\beta} = argmin_\beta \; MSE(\beta) + \lambda||\beta||^2 \iff \hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \tag{5}$$

However, to use ridge regression properly we need to find the optimal value of λ. To do so we propose the following : we will first choose the best model we have so far for each country and then we will compute a ridge regression on the features corresponding to the chose model and iterate over possible values of lambda. At each iteration, we save the obtained R2 score (on the test set) and the corresponding values of λ. We then pick the lambda that produced the highest R2 score. We did so on the selected models of tables 14 but also on the models produced from the VIF analysis (row *all* on table 8).

Table 14 – P-value selection approaches

	China	Korea	Taiwan
Selected model	INTERCEPT CC XRAT CG OPENK	INTERCEPT CC2 CG CI OPENK OPENC PC PG	INTERCEPT, CC, CSAVE, CI, CC^2, $CSAVE^2$, $CSAVE * CC$



(a) China Ridge Model selection (b) Korea Ridge model selection (c) Taiwan Ridge model selection

(d) China Ridge VIF (e) Korea Ridge VIF (f) Taiwan Ridge VIF

Figure 6 – R2 evolution

You can see in figure 6 the progression of the R2 score according to the values of lambda. We set the first values of lambda to be zero and then we used a logspace because. We can clearly identify the maxima on all of the figures. In a mode formal way, we used Matlab to find these maximas and we obtained the results shown in figure 15

	China	Taiwan	Korea	Model Type
Best R2	-0.22	0.71	0.8249	VIF
Optimal λ	1.44	0	0.0758	VIF
Best R2	0.79	0.96	0.8223	sel model
Optimal λ	0	0.12	0.0715	sel model

Table 15 – Ridge Results

Note that *sel model* refers to the model described in table 14

6 Model Analysis

6.1 General economical overview

Let us start by analysing the final results for china, shown in table 16. We get that the relevant features are : the Real gross domestic product (CGDP) per capita relative to the united state (CC), the exchange rate (XRAT), the government share of CGDP (CG), and openness in constant prices (OPENK). We see that all of the features parameter estimators are positive and stays positive in their confidence interval. Note the CI is tables 16,18, 17 are all robust statistics. First for CC since it is the real GDP divided by the population, it is reassuring that it grows alongside the GDP, since it is just a rescaled GDP. Seeing that the exchange is positively correlated with the GDP is no surprise, especially for China. Indeed a high exchange rate with the US dollar means a weak currency which boosts the export. When China's economy gradually opened in the 1980s, the Renminbi was devalued in order to improve the competitiveness of Chinese exports. Thus, the official exchange rate increased from ¥1.50 in 1980 to ¥8.62 by 1994. For CG which represent the government expenditure in relation to GDP, the fact that it is positive is interesting. Looking at different literature we see that the correlation between government expenditure and GDP isn't clear and varies between countries. But in the case of China it is positively correlated . The last feature is OPENK and since the Openness Index is an economic metric calculated as the ratio of the country's total trade, the sum of exports plus imports, to the country's gross domestic product. It makes sense that if export increase GDP increases too.

Parameter	Coef	Std Err	CI LB	CI UB
INTERCEPT	510.30	13.63	482.63	537.97
CC	162.64	29.71	102.33	222.96
XRAT	96.68	35.19	25.24	168.12
CG	196.23	25.26	144.95	247.51
OPENK	143.10	61.47	18.30	267.89
train R2	0.9675			
test R2	0.7938			
λ	0.000			

Table 16 – China : final model

Now let's analyse Taiwan's final results shown in table 17. We get that the relevant features are : CC, Current Savings (CSAVE) , investment share of CGDP (CI), CC^2, $CSAVE^2$ and $CSAVE * CC$. We have that CC is positively correlated with the GDP like it was for China. We see that current saving is negatively correlated with the GDP. It is surprising that it is negative, as one would expect that an increase in the GDP would lead to an increase in savings. A possibility here is that current savings refer to domestic savings and that Taiwan as an emerging country prefers to save in a foreign currency. It is hard to be sure of the correlation between the GDP and the investment share of GDP, since the parameter estimators have a confidence interval including positive and negative number. For current investment, one could think it should clearly be positively correlated with GDP, since investment boost production. One interpretation of a negative correlation could be that investment came at the beginning of the considered period and triggered an economic growth which didn't stop when investment decreased.

Finally, we analyse Korea's final results shown in table 18.We get that the relevant features are : consumption share of CGDP (CC2), CG, CI, OPENK, openness in current price (OPENC), price level of consumption (PC), price level of government (PG) and price level of investment (PI). We see that the correlation isn't sure for CC2, CI, PC, PI since the confidence interval of the parameter estimator of each feature is positive and negative. Since we have a lot of features for this country let's analyse the feature where we are 95% sure about their correlation. First, we see that

CG is negatively correlated with the GDP, which is the contrary to what we found for China. It is really interesting since it supports what we said before of not knowing the impact of government expenditure. So for Korea, it seems that a decrease in government expenditure, which means fewer taxes increases the GDP whereas for China an increase in CG would better distribute the wealth and increase GDP. Next, we have that OPENK is positively correlated like it was for China. Surprisingly OPENC, which only differs from OPENK because it isn't inflation-adjusted, is negatively correlated. It is strange, but a possible explanation would be that the model uses OPENC and OPENK together to extract inflation and compare the real GDP with it. Finally, the price level of government is positively correlated with the GDP. The price level of government consumption (education, parts of health care, police etc.) consists mainly of wages therefore it is coherent that it increases along with GDP.

Parameter	Coef	Std Err	CI LB	CI UB
Intercept	1581.47	82.04	1414.36	1748.57
CC	2297.41	158.64	1974.27	2620.54
CSAVE	-312.51	108.57	-533.66	-91.35
CI	-70.13	58.18	-188.64	48.38
CC^2	453.26	113.82	221.42	685.10
CSAVE^2	-325.32	143.28	-617.18	-33.46
CSAVE*CC	434.54	230.99	-35.98	905.06
train R2	0.9975			
test R2	0.9632			
λ	0.12			

Table 17 – Taiwan : final model

Parameter	Coef	Std Err	CI LB	CI UB
Intercept	2323.81	87.87	2144.36	2503.26
CC2	1169.10	662.99	-1849.17	2523.11
CG	-1402.74	437.13	-2295.47	-510.00
CI	267.76	298.55	-341.96	877.47
OPENK	3412.41	866.99	1641.79	5183.03
OPENC	-1996.69	459.78	-2935.68	-1057.71
PC	-554.20	412.82	-1397.29	288.89
PG	949.54	372.79	188.21	1710.88
PI	-49.00	332.08	-727.18	629.19
train R2	0.9675			
test R2	0.7351			
λ	0.0715			

Table 18 – Korea : final model

An interesting point we can notice by looking at the 3 countries, is that there isn't an economic feature shared by all the country to predict the GDP. And comparing one to one the features existing in the model between the countries, we observe that there isn't more than two common feature between two particular countries. So even if these countries are Asian countries and near geographically, the factors influencing their GDP are really different. This tells us how complex the prediction of GDP is and how economical, technological and political factors strongly impact which features are the most relevant to explain the behavior of the GDP.

6.2 Country comparison and possible improvements

Overall, we managed to design models that are quite predictive with respect to the test set. Indeed, all tests R^2 are bigger than 0.73. Furthermore, we observe that GDP for *Taiwan* is very accurately predicted compared to the other countries. Since it is the only model for which our polynomial feature expansion actually appeared in the selected features, we understand that it might be worthy to look a bit deeper into this procedure (e.g., do feature selection differently or try other features) to improve the results for *China* and *Korea*. However, given time and size constrain, we decided to stop here.

Regularization can be understood as a tool for overfitting reduction. However, the best λ found for *China* is zero, meaning that there might well be an underfitting problem for this country. That could explain why the test R^2 is smaller for *China* than for *Taiwan* for instance.

If we compare our final models to the best possible model (without regularization and feature expansion) given in table 7, we arrive at the conclusion that our models for *Korea* and *Taiwan* are even better than, what we could achieve there which is a good indicator for the goodness of our findings. It also assesses the importance of regularization and polynomial feature expansion for this GDP modeling task. Nevertheless, the model found for *china* performed significantly worse than the optimal one (test $R^2 = 0.79 < 0.93 = \textit{Optimal } R^2$). Therefore, if we had more time, we should invest time in finding more relevant features for *China*.

7 Conclusion

In this study, we searched the best model to approximate the GDP using the data of 3 different countries : China, Taiwan and South-Korea. We started by assuming that all the requirements to use the classical least square model were fulfilled. We then studied the given data and saw collinearity between the given feature. A VIF analysis was therefore done to remove it. In addition, we removed the non-relevant features with different model selection procedures. The next step was to apply the OLS models to the datasets and study which assumptions are valid or not. It occurred that the homoscedasticity assumption was false and we tried to use FGLS and transformed models to take it into account. But from the complexity of the heteroscedasticity, the best way was to use the OLS model with a robust confidence interval. We then improved our model with feature expansion, structural breaks and ridge regression. This allowed us to have coherent results that we then analysed and contextualised on the economic level.

This study lets us understand the complexity of predicting the GDP and the limit of the classical least square assumptions on a real-world application. It shows especially how hard it is to find a model that correctly apprehends the noise of economic data.