Exploring the impact of economic and social factors on stock market performance

Hallberg, Vincent  Hjorth, William

May 10, 2023

Abstract

This study seeks to investigate the relationship between human development factors and domestic stock markets using a multiple linear regression model. Despite efforts to improve the model’s explanatory power, the findings indicate that the model fails to confirm the research question. Nevertheless, the model uncovers a discernible trend in the dataset, albeit with limited explanatory capacity. These results highlight the complexity of the interplay between human development factors and domestic stock markets and suggest the need for further research and alternative modeling approaches to deepen the understanding of this relationship.

Keywords — regression analysis, societal development, domestic stock market, bachelor thesis, applied mathematics, macroeconomic factors
Sammanfattning

Denna studie syftar till att undersöka sambandet mellan faktorer som rör mänsklig utveckling och inhemska aktiemarknader genom användning av en multivariat linjär regressionsmodell. Trots försök att förbättra modellens förklaringskraft, visar resultaten att modellen misslyckas med att bekräfta forskningsfrågan. Trots det upptäcker modellen en urskiljbar trend i datasetet, även om dess förklaringskraft är begränsad. Dessa resultat betonar komplexiteten i samspelet mellan faktorer som rör mänsklig utveckling och inhemska aktiemarknader, och föreslår behovet av ytterligare forskning och alternativa modeller för att fördjupa förståelsen av detta förhållande.

Nyckelord— regressionsanalys, samhällsutveckling, inhemska aktiemarknad, kandidatexamsarbete, tillämpad matematik, makroekonomiska faktorer
Table of contents

1 Introduction 5
   1.1 Background .............................................. 5
   1.2 Previous research .......................................... 5
   1.3 Purpose and Aim ........................................... 6
   1.4 Research question ........................................ 6

2 Mathematical Theory 7
   2.1 Coefficient of determination ($R^2$) ...................... 7
   2.2 P-value .................................................... 7
   2.3 Variance Inflation Factor (VIF) .......................... 7
   2.4 Residual analysis .......................................... 8
      2.4.1 Studentized residuals ......................... 8
      2.4.2 Plots for residual analysis .................. 8
   2.5 Influential data points (DFFITS, COVRATIO, Cook's distance) 10
   2.6 Autocorrelation ......................................... 10
      2.6.1 Durbin-Watson test .............................. 11
      2.6.2 Lagged Dependant Variable (LDV) ............. 11
   2.7 Uncentered regression models ......................... 11
   2.8 Yeo-Johnson Transformation ............................ 11

3 Methodology 13
   3.1 Software ................................................ 13
   3.2 Response variable ....................................... 13
   3.3 Regressor variables ...................................... 13
      3.3.1 Human Development Index .......................... 13
      3.3.2 Gross Domestic Product per Capita (PPP adjusted) 14
      3.3.3 Inflation rate ..................................... 14
      3.3.4 Gini Coefficient .................................. 14
   3.4 Country selection ....................................... 15

4 Results 16
   4.1 First Model ............................................. 16
   4.2 Second Model ........................................... 17
   4.3 Third Models ............................................ 17
      4.3.1 Third model for full country set .................. 18
      4.3.2 Third model for Developed countries ............ 20
      4.3.3 Third model for Developing countries ........... 21
   4.4 Fourth Models .......................................... 23
      4.4.1 Fourth model for full country set .................. 23
      4.4.2 Fourth model for Developed countries ............ 25
      4.4.3 Fourth model for Developing countries ........... 26

5 Discussion 28
   5.1 Model Adequacy ......................................... 28
   5.2 Limitations ............................................... 28
   5.3 Further Research ......................................... 29
6 Conclusion
7 Appendix
References
1 Introduction

1.1 Background

In the 1600s, the creation of the first real stocks emerged through the establishment of various 'East India Companies' in the Netherlands. These companies ventured on long and risky voyages, mitigating their risks by dividing ownership of their vessels and selling parts to investors. Investors would contribute capital and receive a percentage of the profits if the voyage succeeded[8].

This historical context underscores the enduring importance of financial asset trading for speculation and hedging, benefiting the economic interests of a country’s citizens, companies, and governments. A well-functioning and developing domestic stock market has been demonstrated to contribute to a country’s overall economic well-being[1].

Research suggests that macroeconomic factors, such as GDP, inflation, and interest rates, have a significant impact on the stock market’s performance[1]. Simultaneously, human development levels, measured by indicators such as the Human Development Index (HDI), can also influence macroeconomic factors[15]. Understanding the complex interplay between these variables and their impact on the stock market is crucial for policymakers and investors to make informed decisions on investment strategies, economic policies, and social development initiatives.

This study focuses on specific factors of societal development, including HDI, GDP per Capita (PPP adjusted), inflation rate, and the Gini coefficient. By examining the relationship between these indicators and macroeconomic factors, the study aims to illuminate how they collectively influence the domestic stock market.

1.2 Previous research

Previous research has examined the relationship between the economy and stock market performance from various angles. Antonios (2010) found a unidirectional causal relationship between stock market development and economic growth, using the stock market index as a response variable and GDP and bank credits to the private sector as regressor variables[1]. Wongbango and Sharma (2002) investigated the causal links between macroeconomic variables and stock prices in ASEAN-5 countries, concluding that past values of GNP, CPI, money supply, nominal interest rate, and exchange rate could predict changes in stock prices[21]. Sehrawat and Giri (2014) found a unidirectional causal relationship between financial development indicators and HDI in India, using the ratio of domestic credit to private sector to GDP, domestic credit as a share of GDP, the ratio of broad money supply as a percentage of GDP, and GDP per capita as regressor variables[13].

Other research has also examined the relationship between economic developments and human developments. In their article "Paths to success: The relationship between human development and economic growth" from 2011, Suri et Al, find that human development is an important factor for continous economic growth[15]. Similarly G.Ranis (2004) finds that economic growth and human development feeds into each other[11]. Both of these articles arrive at a similar conclusion, that investments in human development ultimately pays off in increases in economic developments [11] [15].

The important contributions of these studies show that there exists some link between the economy and stock market performance as well as a link between the
1.3 Purpose and Aim

This project is founded on the principles of humanistic and liberal thought. The underlying belief is that economic development and societal progress are closely intertwined, with each influencing and reinforcing the other. This liberal theory asserts that greater human development should result in increased economic prosperity, and conversely, a prosperous economy should promote societal advancement. Studies conducted by N. Zehra and S.W. Khattak[24] as well as Ben Shepherd and G. Pasadilla[14] lend support to this idea.

There is also some research suggesting that the stock market plays a critical role in a country’s overall economy, both in the advanced economy and the emerging market. This has been shown in various research papers, e.g. in [9] the author found a strong positive correlation between the stock market and the growth of the economy.

From this we conclude that it may be possible to predict the future economic prosperity given the societal development. Furthermore this conclusion suggests a possible investment strategy, where investors invest in nations which successfully improve on a societal level. If such a strategy exists, it would further reinforce the traditional liberal theory that investing in a nation’s population leads to prosperity for both state and non-state entities.

Therefore the purpose of this study is to investigate the possibility to use societal progress in a country as a predictor of the domestic stock markets performance, in order to see how measuring societal development could serve as a basis for investment opportunities and strategies.

To achieve this a quantitative study will be performed, with some elements of qualitative work. Using a multiple linear regression model, which will examine how social development and domestic stock markets correlates.

1.4 Research question

The research question is:

Does a country’s societal development affect the domestic stock market?


2 Mathematical Theory

For this study we will be using a multiple linear regression model. Regression analysis is a method to observe the correlation between regressor variables and a response variable. The theory is that the regressors affect the response. So if the model is “true”, an increase in the regressors will correspond with an increase in the response variable. In our case, an increase in the developmental factors, will lead to an increase in the domestic stock market performance. The specific regression model we will be performing will be an Ordinary Least Squares (OLS) model. We will evaluate our the accuracy of our model by a few metrics.

2.1 Coefficient of determination ($R^2$)

The first metric for evaluating a regression model is the coefficient of determination, usually referred to as $R^2$. It is a value between 0 and 1, where 0 indicates that the independent variables do not explain any of the variance in the dependent variable, and 1 indicates that the independent variables completely explain the variance in the dependent variable. In other words, $R^2$ measures how well the regression model fits the data. A higher $R^2$ value indicates a better fit, meaning that the independent variables in the model are more effective in explaining the variation in the dependent variable.

2.2 P-value

The p-value is a statistical measure used to assess the significance of the regression model. It quantifies the probability of obtaining the observed data or more extreme results under the assumption that the null hypothesis is true. In the context of regression analysis, the null hypothesis usually states that the coefficients of the independent variables are equal to zero, indicating no relationship with the dependent variable.

The p-value ranges between 0 and 1. A small p-value (typically less than a predetermined significance level, e.g., 0.05) suggests strong evidence against the null hypothesis. It indicates that the independent variables in the model have a statistically significant impact on the dependent variable. On the other hand, a large p-value implies that there is not enough evidence to reject the null hypothesis. In such cases, the independent variables are considered not statistically significant in explaining the variation in the dependent variable.

Interpreting the p-value is crucial in regression analysis, as it helps determine the statistical significance of the relationship between the response variables and the dependent variable. Lower p-values provide stronger evidence for the significance of the model, while higher p-values suggest a lack of statistical significance.

In this study the significance level of $p \leq 0.05$ will be deemed statistically significant.

2.3 Variance Inflation Factor (VIF)

Variance inflation factor (VIF) is a measure of multicollinearity among the predictor variables in a regression model. Multicollinearity occurs when the predictor variables in a regression model are highly correlated with each other. In such cases, the effect of each predictor on the response variable cannot be estimated accurately, leading to unreliable and unstable results.
The VIF measures the degree to which the variance of the estimated regression coefficients is inflated due to multicollinearity. Specifically, it measures the increase in the variance of the estimated regression coefficients relative to what would be expected if the predictor variables were not correlated with each other. A VIF value of 1 indicates that there is no multicollinearity among the predictor variables, while a value greater than 1 indicates that there is some degree of multicollinearity. A commonly used threshold for identifying high levels of multicollinearity is a VIF value of 5 or greater.

To compute the VIF for a predictor variable, a regression model is fitted with that variable as the response variable, and all other predictor variables as predictors. The VIF is then computed as the ratio of the variance of the estimated regression coefficient for the variable of interest to the variance of the estimated regression coefficient for the intercept term.

We will be using several VIF-tests in our model to verify whether or not there is multicollinearity between our regressor variables.

2.4 Residual analysis

Residual analysis is a powerful tool to analyze the adequacy of the model or the data. Doing a OLS-regression requires some assumptions about the residuals to hold, and they can be compacted as normality which states that $\epsilon \sim N(0, \sigma^2 I)$ [10, p. 129]. Doing a residual analysis gives information about the assumptions.

2.4.1 Studentized residuals

In this report the scaling method studentized residuals will be used. The studentized residuals are calculated as:

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

This is beneficial for analyzing points with large residuals and large $h_{ii}$ because they can be possible influential points in the model [10, p. 133]

2.4.2 Plots for residual analysis

To analyze the residuals in the data QQ-plots and Tukey-Anscombe plots will be used.

The QQ-plot is a quantile-quantile scatterplot which plots two sets of quantiles on the x and y axis, with sample quantiles on the y-axis and theoretical quantiles on the x-axis. It takes the sample quantiles from the data and sorted in ascending order and plots it against the theoretical quantiles. If the QQ-plot forms a straight line the assumption that the residuals are normally distributed is not rejected [3].

The Tukey-Anscombe plot plots the residuals $r_i$ on the y-axis and the fitted value $\hat{Y}_i$ on the x-axis. The reason for doing this is that the sample correlation between them should be zero. The residuals should be distributed equally in the plot if there is no correlation between the residuals and the fitted values. If that is not the case and there is a systematic relation in the plot, some adjustments to the model has to be made or influential data points have to be removed[2, p. 12]. Below are some figures of the different Tukey-Anscombe plots:

In Figure 3 a) there is a linear increase in standard deviation which can be a problem of heteroskedasticity, in b) there is a non-linear increase in standard deviation that can also be a problem of heteroskedasticity, in c) there are two groups with different
Figure 1: QQ-Plot of quantiles which comes from truly Normal distributions

Figure 2: Tukey-Anscombe plot with no correlation between residuals and fitted values

Figure 3: Tukey-Anscombe plots
standard deviations and lastly in d) there is asymmetry which suggest that the model might need a quadratic term or another regressor variable[2, p. 12]. If there is some serial correlation between the residuals the assumption $E[\epsilon] = 0$ is not met.

2.5 Influential data points (DFFITS, COVRATIO, Cook’s distance)

To deal with the influential data points detected in the QQ and Tukey-Anscombe plots this study will use the influence measurements DFFITS, COVRATIO and Cook’s distance, every observation that doesn’t pass these tests will be deleted from the model.

The definition of DFFITS for the $i$:th observation is:

$$DFFITS_i = \frac{Y_i - \hat{Y}_{i(i)}}{s(i)\sqrt{h_{ii}}}$$

An observation $i$ will be deemed an influential data point if the DFFITS value for it is bigger than a certain threshold which is $|DFFITS_i| > 2\sqrt{p/n}$. Where $p$ is the number of regressors and $n$ is the number of data points. DFFITS measures the shift in $Y_i$ if the $i$:th observation is deleted.

The definition of COVRATIO for the $i$:th observation is:

$$COVRATIO_i = \frac{\det(s^2(X'_{(i)}X_{(i)}^{-1}))}{\det(s^2[X'X]^{-1})}$$

An observation $i$ will be deemed an influential data point if the COVRATIO value for it is bigger than a certain threshold which is $|COVRATIO_i - 1| > 3p/n$. Where $p$ is the number of regressors and $n$ is the number of data points. COVRATIO is used to measure the impact of an observation on the coefficients in the model.

The definition of Cook’s Distance for the $i$:th observation is:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_i)'(X'X)(\hat{\beta} - \hat{\beta}_i)}{ps^2}$$

where $s^2 = SSRES/(n - p)$ and $n$ is the number of observations and $p$ is the number of regressor variables, $\hat{\beta}_i$ is the estimation the OLS makes for $\hat{\beta}$ if the $i$:th data point is deleted. A data point $i$ will be deemed influential if $D_i > F_{0.5, p, n-p}$

[12, p. 361-367]

2.6 Autocorrelation

One of the basic presumptions in regression analysis is that all the error terms $\epsilon$ are uncorrelated. Any breach of this premise could harm the model. Errors that are correlated are referred to as autocorrelation with one another. OLS estimates may result in significantly underestimated error variances $\sigma^2$ if autocorrelation is present, even though they are still unbiased. Autocorrelation implies that OLS estimates are no longer the minimum variance estimates. This suggests that hypothesis tests, prediction intervals, and confidence intervals are less exact techniques [10, p. 475]. Time series
data frequently shows some autocorrelation. The addition of a lagged dependent variable is one method for dealing with the autocorrelation problem [10, p. 494-495].

2.6.1 Durbin-Watson test
To investigate if the residuals have a serial correlation the Durbin-Watson test may be used and is defined as:

\[
d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \approx 2(1 - \hat{\rho}),
\]

Here \(\hat{\rho}\) is the sample residual correlation (between \(e_i\) and \(e_{i-1}\) [12, p. 354-355]. The value of \(\hat{\rho}\) lies between 1 and -1, so the Durbin-Watson test has a value between 0 and 4. Where a 0 indicates a positive serial correlation between the residuals, 4 indicates a negative serial correlation and 2 indicates no serial correlation. Normally a value between 1.5 and 2.5 is considered low autocorrelation [12].

2.6.2 Lagged Dependant Variable (LDV)
A lagged dependent variable refers to a variable in a regression model that is taken from a previous time period and used as a predictor variable in the current time period. The use of lagged dependent variables can be useful in predicting future values, especially if there is a strong correlation between past and present values. However, use of LDVs can introduce issues of autocorrelation and violate some of the assumptions of regression models.

We will introduce a LDV to our model and explain why and how this affects our results in the Results section (see Third Models in the results section).

2.7 Uncentered regression models
An uncentered regression model is a model where the predictor variables and the response variable are used in their original scales. In contrast to a centered regression model, where the mean of the predictor variables and the response variable are subtracted from each observation to make the mean zero, and the resulting variables are used in the regression model. The difference between these two approaches is that in a centered regression model, the intercept represents the predicted value of the response variable when the predictor variables are at their mean values. In short, in a uncentered regression model, intercept represents the predicted value of the response variable when the predictor variables are equal to zero.

Our model will begin as a centered model but we will transform it to an uncentered one and explain why this fits and doesn’t violate any of our assumptions (see Third Models in the Results section).

2.8 Yeo-Johnson Transformation
The Yeo-Johnson transformation is a power transformation that can be used without restrictions for the dependant variable \(y\), which means that it can transform values for \(y\) that are zero or negative. The transformation is defined as[23]:
\[ \psi(\lambda, y) = \begin{cases} 
\frac{(y+1)^{\lambda-1}}{\lambda} & \text{if } \lambda \neq 0, y \geq 0 \\
\log(y+1) & \text{if } \lambda \neq 2, y \geq 0 \\
\frac{(-y+1)^{2-\lambda-1}}{(2-\lambda)} & \text{if } \lambda = 0, y < 0 \\
-\log(-y+1) & \text{if } \lambda = 2, y < 0 
\end{cases} \]

This transformation will give different values for \( \lambda \) when ran in the software python 3. These are some values of lambda that corresponds to different transformations:

\( \lambda = -1 \): This corresponds to the reciprocal transformation, where the variable is raised to the power of -1 \((1/y)\). It is suitable for data that is heavily right-skewed.

\( \lambda = -0.5 \): This corresponds to the reciprocal square root transformation \((1/\sqrt{y})\). It is useful for moderately right-skewed data.

\( \lambda = 0 \): This represents the logarithmic transformation \((\log(y))\). It is appropriate for data that is positively skewed.

\( \lambda = 0.5 \): This corresponds to the square root transformation \((\sqrt{y})\). It is suitable for data that is moderately left-skewed.

\( \lambda = 1 \): This represents the identity transformation \((y)\). It is used for variables that are normally distributed or do not require transformation.[22]
3 Methodology

3.1 Software

In this study Python 3 was used to analyze the data, the libraries used were:

```python
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.stats.outliers_influence import OLSInfluence
from statsmodels.stats.outliers_influence import variance_inflation_factor
from scipy import stats
from sklearn.metrics import r2_score
```

3.2 Response variable

To measure the response variable, stock index data was collected from 25 countries (see Appendix). The aim was to use indices that emulate the entire domestic stock market, prioritizing indices with the broadest exposure. However, in many cases, there was no to alternatives, and therefore the available indices were used. Additionally, while our goal was to use indices between 1990-2021, some indices did not track that far back (see Appendix).

Once we collected the indices, they were converted into a simple 5 year moving average and then changed to represent the percentage growth of the 5 year average. This became the response variable.

We chose to use a 5-year simple moving average and growth percentage to account for the fluctuation in stock markets. By changing to a 5-year moving average, the response variable is less likely to provide extreme results that provide little information.

3.3 Regressor variables

3.3.1 Human Development Index

Human Development Index (HDI) is an index that is meant to track the well being of a nations citizenry. The index is supposed to measure the well being of people in the following three categories: A long and healthy life, Knowledge and a decent standard of living [18] (see Appendix for formula calculating HDI). We will use this as one of our societal factors in our analysis.

The method for calculating HDI changed in 2010. Post 2010, HDI is calculated from Gross National Income(GNI) per capita at purchasing power parity (PPP), life expectancy at birth and mean years of schooling as well as the expected years of schooling. These variables are then weighed differently by separate factors[18].

Before 2010, the method involved: life expectancy at birth, education and standard of living. Education was measured by the adult literacy rate and the combined primary, secondary, and tertiary gross enrollment ratio. Which were then weighted differently. The standard of living was calculated by the natural logarithm of gross domestic product by capita at PPP (See Appendix for a full explanation of the calculations)[18].
The data was collected from the UNDP (United Nations Development programme) which regularly performs this analysis. However this analysis is quite recent and as such only stretches as far back as 1990 which is why we start our analysis at that year and move forwards[18].

Like the stock indices the HDI measurement was changed to a growth percentage over from a 5-year moving average to synchronise with the rest of the model.

3.3.2 Gross Domestic Product per Capita (PPP adjusted)

Gross Domestic Product per Capita (PPP adjusted) is a measure that indicates the average economic output per person in a country, taking into account differences in purchasing power between nations. It provides insights into the standard of living and economic well-being of a country’s population.

In our analysis, we included GDP per Capita (PPP adjusted) (which will be refered to as GDcP aswell) as one of the economic metrics to examine its relationship with the stock market. By considering the per capita value, we can assess the individual economic contribution and potential purchasing power of the population.

We obtained the data for GDP per Capita (PPP adjusted) from reliable sources such as the World Bank [16]. It’s worth noting that we ensured minimal gaps in the data, and any missing values primarily corresponded to countries that were not included in our model.

Similar to the stock indices and HDI measurements, we transformed the GDP per Capita values into growth percentages over a 5-year moving average. This transformation allows us to focus on analyzing the changes in GDP per Capita as a predictor of stock market behavior.

In summary, the inclusion of GDP per Capita (PPP adjusted) in our model helps us assess the economic conditions and potential influences on the stock market by considering the average economic output per person.

3.3.3 Inflation rate

The inflation rate is the rate of which a country’s money loses value[6]. We collected our data from the International Monetary Fund. This data included a lot of gaps for certain countries and years. Fortunately most of the countries in the model were unaffected, but some still needed to be cleaned up and removed from the data set. Unlike the other variables used in the model the inflation rate was not changed to a percentage. Instead we used a 5-year moving average and then used the numeric change between each average, unlike earlier variables which use a percentage change. This because inflation can vary from positive to negative, which when turned to percentage growth turns into infinite values.

The data for the inflation rate was found through the International Monetary Fund [5].

3.3.4 Gini Coefficient

The Gini coefficient, or Gini Index, is a statistical measure for the wealth inequality of a nation. The higher the coefficient higher the wealth disparity between individuals in the nation (see Appendix for formula calculating Gini Index). We gathered the data of the Gini coefficient from the World Bank. This data set included large gaps in the data frame mostly from developing countries, usually there was some data but
with gaps for certain years. This data was however still included in the analysis. The reason for still including this data is that ruling it out would hamper the ability to create regression and study the research question in a meaningful way.

The data for GDP was found through the World Bank [17].

3.4 Country selection

To answer the research question properly, it was necessary to include a diverse range of countries from both the developed and developing worlds. However, the choice of countries was constrained by the lack of a domestic stock market and the length of time data was available. Ultimately, the study included countries with a domestic stock exchange that extended at least until the late 2000s, with the shortest data available stretching back to 2008. A complete list of the countries used in the study can be found in the appendix. To account for the differing opportunities and challenges faced by developing and developed countries, we categorized the countries accordingly. The cutoff point for the developed category was an HDI (Human Development Index) of above 0.85, which is classified as ‘very high human development’ by the United Nations Development Programme[18]. A list of countries and their categories can be found in the Appendix.
4 Results

4.1 First Model

Table 1: Summary of the first Model

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-2.77</td>
<td>se(1.69)</td>
<td>t(-1.64)</td>
</tr>
<tr>
<td>HDI%</td>
<td>1.43</td>
<td>se(2.01)</td>
<td>t(0.71)</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>2.42</td>
<td>se(0.51)</td>
<td>t(4.78)</td>
</tr>
<tr>
<td>Gini%</td>
<td>0.51</td>
<td>se(0.62)</td>
<td>t(0.83)</td>
</tr>
<tr>
<td>Inflation change</td>
<td>-0.9626</td>
<td>se(0.70)</td>
<td>t(-1.376)</td>
</tr>
</tbody>
</table>

Dependent Variable: $SI$
Model: $SI = c_0 + c_1 \cdot HDI + c_2 \cdot GDPpC + c_3 \cdot GC + c_4 \cdot IC$
Method: Ordinary Least Squares
Sample: 170 observations
R-squared: 0.222
Adjusted R-squared: 0.204
F-statistic: 11.8
Prob (F-statistic): $p(1.87 \cdot 10^{-8})$
Durbin-Watson: 0.472

The first model suffers from several issues. The low R-squared value indicates that the model poorly explains the data. Additionally, the Durbin-Watson test reveals significant autocorrelation among the regressors. Most notably, removing observations based on missing data for the Gini index results in a model with few observations. Because of the amount of datapoints that couldn’t be used because of the lack of data for the Gini index that variable was excluded from the other models.
4.2 Second Model

Table 2: Summary of the second model

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.63</td>
<td>se(1.22)</td>
<td>t(-0.51)</td>
<td>p(0.608)</td>
</tr>
<tr>
<td>HDI%</td>
<td>3.00</td>
<td>se(1.14)</td>
<td>t(2.63)</td>
<td>p(0.009)</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>1.64</td>
<td>se(0.30)</td>
<td>t(5.46)</td>
<td>p(0.000)</td>
</tr>
<tr>
<td>Inflation change</td>
<td>-0.11</td>
<td>se(0.02)</td>
<td>t(-5.54)</td>
<td>p(0.000)</td>
</tr>
</tbody>
</table>

Dependent Variable: SI
Model: \( SI = c_0 + c_1 \cdot HDI + c_2 \cdot GDPpC + c_3 \cdot IC \)
Method: Ordinary Least Squares
Sample: 524 observations
R-squared: 0.147
Adjusted R-squared: 0.142
F-statistic: 29.8
Prob (F-statistic): \( p(8.63 \cdot 10^{-18}) \)
Durbin-Watson: 0.581

Table 3: VIF Values for the second model

<table>
<thead>
<tr>
<th>Predictor</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>6.048</td>
</tr>
<tr>
<td>HDI%</td>
<td>1.209</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>1.205</td>
</tr>
<tr>
<td>Inflation change</td>
<td>1.009</td>
</tr>
</tbody>
</table>

The second model includes significantly more observations, increasing from 170 to 524. However, several issues remain with the model. The R-squared value is lower than in the first model, indicating poor explanatory power. The Durbin-Watson test reveals significant autocorrelation, and a VIF-test indicates multicollinearity between the constant and the regressor variables.

4.3 Third Models

In these models a lagged dependant variable \( SI_{t-1} \) has been added and the constant has been removed because of the poor VIF-value it produced. The removed constant will lead to an uncentered \( R^2 \). Here we will present three separate models. One model with the complete set of 25 countries, one model including developing countries and one model including developed countries.
4.3.1 Third model for full country set

Table 4: Third model for all countries

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI%</td>
<td>1.33</td>
<td>se(0.77)</td>
<td>t(1.73)</td>
<td>p(0.085)</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>0.57</td>
<td>se(0.15)</td>
<td>t(3.70)</td>
<td>p(0.000)</td>
</tr>
<tr>
<td>Inflation change</td>
<td>-0.05</td>
<td>se(0.01)</td>
<td>t(-3.94)</td>
<td>p(0.000)</td>
</tr>
<tr>
<td>SI_{t-1}</td>
<td>0.67</td>
<td>se(0.03)</td>
<td>t(23.61)</td>
<td>p(0.000)</td>
</tr>
</tbody>
</table>

Dependent Variable: SI
Model: SI = c_1 \cdot HDI + c_2 \cdot GDPpC + c_3 \cdot IC + c_4 \cdot SI_{t-1}
Method: Ordinary Least Squares
Sample: 524 observations
R-squared: 0.713
Adjusted R-squared: 0.711
F-statistic: 322.9
Prob (F-statistic): p(2.22 \cdot 10^{-139})
Durbin-Watson: 1.563

Table 5: VIF Values for the third centered model

<table>
<thead>
<tr>
<th>Predictor</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>6.178</td>
</tr>
<tr>
<td>HDI%</td>
<td>1.212</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>1.216</td>
</tr>
<tr>
<td>Inflation change</td>
<td>1.0366</td>
</tr>
<tr>
<td>SI_{t-1}</td>
<td>1.045</td>
</tr>
</tbody>
</table>

The third model for the complete set of 25 countries. The major notable changes to the model were the exclusion of the constant, making the model uncentered, and the inclusion of a the lagged dependent variable.

The constant caused issues both with multicollinearity and model accuracy. As can be noted from Table 5: VIF Values for the final centered model, the VIF for the constant was at 6.178, above our cutoff at 5. Removing the intercept removes the centering of the model. Which means that our model assumption now changes so that the assumption is that only a change in the regressor variables will create a change in the response variable, meaning that the response variable cannot change on its own. While this assumption is unrealistic, it still accurately reflects the purpose of this report.

The inclusion of the lagged dependent variable significantly improves the R-squared value of the model. This suggests that any investment strategy based on the conclusion of this report should consider both the typical developmental factors and the stock market change from the previous year. The introduction of the lagged dependent variable also addresses issues with autocorrelation, as indicated by the Durbin-Watson test.

The models F-statistic and Prob (F-statistic) shows that the model is of statistical significance. The problem here is that the p-value for the HDI% regressor is
0.085 which is not low enough to reject the null hypothesis, and that implies that the regressor is not statistically significant for this model.

Table 6: VIF Values for the uncentered third model

<table>
<thead>
<tr>
<th>Predictor</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI&amp;</td>
<td>3.48</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>3.69</td>
</tr>
<tr>
<td>Inflation change</td>
<td>1.04</td>
</tr>
<tr>
<td>$SI_{t-1}$</td>
<td>1.47</td>
</tr>
</tbody>
</table>

Furthermore, a VIF test (seen in table 6) was conducted, which did not reveal any problematic levels of multicollinearity. The R-squared value is not perfectly satisfactory but still carries a reasonable level of model accuracy.

From Figure 4 it can be seen that the QQ-plot displays a heavy tail, which suggest that the residuals are not normally distributed. To try to fix this the model was subjected to a Yeo-Johnson transform, which resulted in a slight improvement in model accuracy. However, the lambda value of 0.8456 suggests that there is no significant benefit in transforming the model to a non-linear one. Also there is dispersion in the Tukey-Anscombe plot which suggest that here are groups of residuals with different variance, hence the assumption that $E[c] = 0$ is violated. To address these issues in the plots the deletion of influential data points will be done, which can be seen in section 4.4.1.
### 4.3.2 Third model for Developed countries

Too see countries included see Appendix 7

#### Table 7: Developed countries

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI%</td>
<td>-3.36</td>
<td>se(1.81)</td>
<td>t(−1.855)</td>
<td>p(0.065)</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>0.84</td>
<td>se(0.22)</td>
<td>t(3.81)</td>
<td>p(0.000)</td>
</tr>
<tr>
<td>Inflation change</td>
<td>2.23</td>
<td>se(1.15)</td>
<td>t(1.94)</td>
<td>p(0.053)</td>
</tr>
<tr>
<td>$SI_{t−1}$</td>
<td>0.73</td>
<td>se(0.04)</td>
<td>t(16.71)</td>
<td>p(0.000)</td>
</tr>
</tbody>
</table>

Dependent Variable: $SI$

Model: $SI = c_1 \cdot HDI + c_2 \cdot GDPpC + c_3 \cdot IC + c_4 \cdot SI_{t−1}$

Method: Ordinary Least Squares

Sample: 225 observations

R-squared: 0.686

Adjusted R-squared: 0.680

F-statistic: 120.4

Prob (F-statistic): $p(2.33 \cdot 10^{-43})$

Durbin-Watson: 1.566

In Table 7 it can be seen that the *F-statistic* and *P*ob (F-statistic) shows that the significance level of the model to be enough to be statistically significant. The significance of the regressor variables, however, is limited by the regressors HDI% and Inflation Change, which has a p-value of 0.065 and 0.053 respectively. These p-values do not provide sufficient evidence to reject the null hypothesis, which means that these regressors are not statistically significant in this model.

#### Table 8: VIF Values

<table>
<thead>
<tr>
<th>Predictor</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI%</td>
<td>6.01</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>6.04</td>
</tr>
<tr>
<td>Inflation change</td>
<td>1.15</td>
</tr>
<tr>
<td>$SI_{t−1}$</td>
<td>1.33</td>
</tr>
</tbody>
</table>

Table 8 shows that there exists multicollinearity for the regressors HDI% and GDPpC% which is problematic, but exclusion of these variables would lead to a meaningless interpretation of how societal development affects the stock market, therefore they will still be kept in the model in section 4.4.2 where the influential data points have been deleted.

Figure 6 of the QQ-plot looks good, which suggest that the residuals are normally distributed. This suggests that no transformation should be done, which is accurate with the $\lambda$ obtained when the Yeo-Johnson transformation was done, which yielded $\lambda = 1$ and therefore no transformation is needed. The Tukey-Anscombe plot in figure 7 also look symmetric which means that no violation to the assumption that $E[\epsilon] = 0$. However the model will still be subject to deletion of influential data points, to try to improve the $p$-value. This new model can be seen in section 4.4.2
4.3.3 Third model for Developing countries

To see countries included see Appendix 7

<table>
<thead>
<tr>
<th>Table 9: Developing countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>HDI%</td>
</tr>
<tr>
<td>GDPpC%</td>
</tr>
<tr>
<td>Inflation change</td>
</tr>
<tr>
<td>SI_{t-1}</td>
</tr>
</tbody>
</table>

Dependent Variable: SI
Model: SI = c_1 \cdot HDI + c_2 \cdot GDPpC + c_3 \cdot IC + c_4 \cdot SI_{t-1}
Method: Ordinary Least Squares
Sample: 299 observations
R-squared: 0.726
Adjusted R-squared: 0.723
F-statistic: 195.9
Prob (F-statistic): p(9.69 \cdot 10^{-82})
Durbin-Watson: 1.572

In Table 9 it can be seen that the F-statistic and P_{\text{rob}} (F-statistic) shows that the significance level of the model to be enough to be statistically significant. The significance of the regressor variables, however, is limited by the HDI% regressor, which has a p-value of 0.275. This p-value does not provide sufficient evidence to reject the null hypothesis, which means that the regressor is not statistically significant in the model.

Table 10 shows no problems with the VIF-values for any variable.
In figure Figure 8 it can be seen that the QQ-plot suffer from heavy tails, which suggest that the residuals are not normally distributed. It was first tried to fix with a Yeo-Johnson transformation of the response variable, but the $\lambda = 0.825$ suggest that no transformation is needed. Figure 9 which is the Tukey-Anscombe plot shows that...
there is some dispersion in the data and suggests that there are groups of residuals with different variances, hence the assumption that $E[e] = 0$ is violated. Do deal with the problems of these plots the influential data points will be addressed in the updated model in section 4.4.3.
4.4 Fourth Models

In this section all the models have been subject to deletion of influential data points via DFFITS, COVRATIO and Cook’s Distance. The only difference here in the deletion is that the threshold for DFFITS was changed to \( |\text{DFFITS}_i| > 2.9 \sqrt{p/n} \) instead of \( |\text{DFFITS}_i| > 2 \sqrt{p/n} \), which is the most common threshold. This was done since too many points were deemed influential and deleted with the conventional threshold.

4.4.1 Fourth model for full country set

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI%</td>
<td>-1.03</td>
<td>t(-1.68)</td>
<td>p(0.095)</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>0.73</td>
<td>t(6.24)</td>
<td>p(0.000)</td>
</tr>
<tr>
<td>Inflation change</td>
<td>-0.24</td>
<td>se(-0.52) t(-5.54)</td>
<td>p(0.606)</td>
</tr>
<tr>
<td>SI_{t-1}</td>
<td>0.69</td>
<td>se(0.03)</td>
<td>t(21.31)</td>
</tr>
</tbody>
</table>

Dependent Variable: SI
Model: \( SI = c_1 \cdot HDI + c_2 \cdot GDPpC + c_3 \cdot IC + c_4 \cdot SI_{t-1} \)
Method: Ordinary Least Squares
Sample: 392 observations
R-squared: 0.751
Adjusted R-squared: 0.748
F-statistic: 291.9
Prob (F-statistic): \( p(1.49 \cdot 10^{-115}) \)
Durbin-Watson: 1.586

The results in Table 11 shows that the \( F\)-statistic and \( P_{\text{prob}} (F\)-statistic) is good enough to deem the model significant. It can also be seen that the deletion of influential data points have improved the \( R^2 \) value with 3.8% units compared to the model in section 4.3.1. The amount of deleted observations is 132, which is removal of approximately 25.2% of the observations. The p-values for the regressors HDI% and Inflation change are 0.095 and 0.606, respectively. These p-values suggest that there is insufficient evidence to reject the null hypothesis for these regressors, indicating that they are not statistically significant in the model.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI%</td>
<td>3.47</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>3.86</td>
</tr>
<tr>
<td>Inflation change</td>
<td>1.09</td>
</tr>
<tr>
<td>SI_{t-1}</td>
<td>1.64</td>
</tr>
</tbody>
</table>

In the Table 12 of VIF values there is no problem of multicollinearity.
In Figure 10 the QQ-plot suggests the residuals are normally distributed. In Tukey-Anscombe plot which is seen in Figure 11 shows that the residuals are symmetrically scattered around 0 and therefore the assumption that $\mathbb{E}[\epsilon] = 0$ is not violated. The Figure 10 and 11 should be compared to Figure 4 and Figure 5, and the results show that the deletion of influential data points improves the QQ and Tukey-Anscombe plot.
### 4.4.2 Fourth model for Developed countries

To see countries included see Appendix 7

<table>
<thead>
<tr>
<th>Table 13: Developed countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>HDI%</td>
</tr>
<tr>
<td>GDPpC%</td>
</tr>
<tr>
<td>Inflation change</td>
</tr>
<tr>
<td>$S_{t-1}$</td>
</tr>
</tbody>
</table>

Dependent Variable: $S_t$

Model: $S_t = c_1 \cdot HDI + c_2 \cdot GDPpC + c_3 \cdot IC + c_4 \cdot S_{t-1}$

Method: Ordinary Least Squares

Sample: 209 observations

R-squared: 0.678

Adjusted R-squared: 0.671

F-statistic: 107.7

Prob (F-statistic): $p(2.81 \cdot 10^{-49})$

Durbin-Watson: 1.441

The results in Table 13 shows that the $F$-statistic and $P_{or}$ (F-statistic) is good enough to deem the model significant. It can also be seen that the deletion of influential data points have decreased the $R^2$ value with 0.8% units compared to the model in section 4.3.2. The amount of deleted observations is 16, which is removal of approximately 7.1% of the observations. The p-values for all the regressors are less than 0.05, indicating that the regressors are statistically significant. Therefore, we can reject the null hypothesis of the model.

<table>
<thead>
<tr>
<th>Table 14: VIF Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor</td>
</tr>
<tr>
<td>HDI%</td>
</tr>
<tr>
<td>GDPpC%</td>
</tr>
<tr>
<td>Inflation change</td>
</tr>
<tr>
<td>$S_{t-1}$</td>
</tr>
</tbody>
</table>

In Table 14 there is still multicollinearity present for the variables HDI% and GDPpC%, however as in the model in section 4.3.2 these regressors can’t be excluded from the model because that would defeat the purpose of the research question.

In Figure 12 the QQ-plot suggests the residuals are normally distributed. In Tukey-Anscombe plot which is seen in Figure 13 shows that the residuals are symmetrically scattered around 0 and therefore the assumption that $E[e] = 0$ is not violated. The Figure 10 and 11 should be compared to Figure 6 and Figure 7. Comparing the results shows that the deletion of influential data points doesn’t improves the QQ and Tukey-Anscombe plot in a way that improves the model significantly.
### 4.4.3 Fourth model for Developing countries

Too see countries included see Appendix 7

<table>
<thead>
<tr>
<th>Table 15: Developing countries</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coefficient</strong></td>
</tr>
<tr>
<td>HDI%</td>
</tr>
<tr>
<td>GDPpC%</td>
</tr>
<tr>
<td>Inflation change</td>
</tr>
<tr>
<td>SI_{t-1}</td>
</tr>
</tbody>
</table>

Dependent Variable: $SI$
Model: $SI = c_1 \cdot HDI + c_2 \cdot GDPpC + c_3 \cdot IC + c_4 \cdot SI_{t-1}$
Method: Ordinary Least Squares
Sample: 238 observations
R-squared: 0.751
Adjusted R-squared: 0.747
F-statistic: 176.7
Prob (F-statistic): $p(1.75 \cdot 10^{-69})$
Durbin-Watson: 1.672

The results in Table 15 shows that the $F$-statistic and $P_{orb} (F$-statistic) is good enough to deem the model significant. It can also be seen that the deletion of influential data points have improved the $R^2$ value with 2.5% units compared to the model in section 4.3.1. The amount of deleted observations is 61, which is removal of approximately 20.4% of the observations. The $p$-value for the HDI% regressor is 0.461. This $p$-value suggests that there is insufficient evidence to reject the null hypothesis for the HDI% regressor, indicating that its explanation is not statistically significant in the model.

Table 16 shows the VIF test performed on this model and give the result that there is no multicollinearity between regressors.

The QQ-plot in Figure 14 indicates that the residuals follow a normal distribution,
Table 16: VIF Values

<table>
<thead>
<tr>
<th>Predictor</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI%</td>
<td>3.60</td>
</tr>
<tr>
<td>GDPpC%</td>
<td>4.09</td>
</tr>
<tr>
<td>Inflation change</td>
<td>1.01</td>
</tr>
<tr>
<td>SI_{t-1}</td>
<td>1.65</td>
</tr>
</tbody>
</table>

Figure 14: suggesting that the assumption of normality is met. Additionally, the Tukey-Anscombe plot in Figure 15 displays symmetric scatter around 0, indicating that the assumption of $\mathbb{E}[\epsilon] = 0$ is not violated. It is important to compare Figures 14 and 15 with Figures 8 and 9. The comparison reveals that removing influential data points has resulted in improved QQ and Tukey-Anscombe plots.
5 Discussion

5.1 Model Adequacy

The initial model had limitations as it had to drop a significant number of data points due to issues with the Gini index dataset. The second model was a simple model with far more observations but still underperformed. Therefore, we applied a variety of transformations, which resulted in the third model.

The third model's accuracy measures indicate that it is a mediocre model. With an R-squared value of 0.711, it falls short of confirming our initial hypothesis that the chosen predictors would significantly predict the response variable. As a result, we decided to split the model into two parts, with a focus on analyzing the results for developing nations. This approach should provide more relevant insights for an investment strategy built upon our research.

However, even after dividing the model, we found that it still underperformed. Both the models for developing and developed countries showed issues of multicollinearity, inadequate residual analysis, low R-squared values, and low p-values for the most relevant variables. In order to address these issues, we constructed the fourth model by removing all data points that had a DFFITS-value of above 2.9. While this approach could increase the model’s explanatory power, it also introduces a significant amount of bias, which raises concerns about the validity of the model.

The fourth model we developed had to remove around 20% of the dataset as outliers, which may have introduced a significant amount of bias and resulted in inaccurate predictions. Therefore, interpreting the results from this model requires caution. It is important to note that our third model, despite its limitations, could provide a more reliable representation of the relationship between the predictor variables and the response variable, since it did not require the removal of a significant portion of the data.

As such, we recommend prioritizing the interpretation of our third model over the fourth model. However, we acknowledge the limitations of both models and approach their results with caution. Future studies could address these limitations by exploring alternative data sets or collecting additional data to strengthen the analysis.

5.2 Limitations

While this aimed to provide valuable insights into the relationship between financial development and human development, it is important to acknowledge its limitations. One of the most significant limitations was the Gini index dataset, which posed several challenges. Firstly, the dataset had limitations in terms of the number of data points it contained. This resulted in the first model being inadequate, as it dropped too many data points due to the missing data in the Gini index dataset.

Another limitation of this project was the limited number of years of HDI data. While HDI is a valuable measure of human development it has only been measured since 1990. This places a constraint on the number of years over which the relationship between financial development and human development can be explored. Consequently, the scope of conclusions that can be drawn from this research regarding the longitudinal relationship between these variables is limited by this restriction.

In addition, the dataset used in this project only included a limited number of indices from developing nations. This limits the generalizability of the findings to
other developing nations not included in the study. A data set with more developing
countries would have been preferable and could alter the conclusion of this study.

5.3 Further Research

One potential avenue for further research could be exploring alternative datasets to the
Gini index. The limitations of the Gini index dataset presented significant challenges
in the development of the model. Exploring alternative measures to income inequality could provide a more comprehensive understanding of the relationship between
financial development and human development.

Another area for further research could be expanding the scope of the study to
include a broader range of developing nations. The limited number of indices from
developing nations included in the dataset used in this study may have impacted the
validity of the findings. By including more developing nations in the analysis, a more
comprehensive understanding of the relationship between financial development and
human development in developing nations could be achieved.

Additionally, further research could focus on exploring the relationship between
financial development and human development over a longer period of time. The lim-
ited number of years of HDI data available for this study restricted the ability to draw
conclusions about the longitudinal relationship between these variables. By collecting
additional data or exploring alternative data sources, the scope of conclusions that
can be drawn from this research could be expanded.

Finally, future studies could explore alternative statistical models to address the
limitations and challenges faced in this study. The use of multiple regression models
is one approach, but alternative models such as time series analysis or panel data
analysis may provide more accurate and reliable results. By exploring different statis-
tical models and approaches, a more comprehensive understanding of the relationship
between stock markets and societal development could be achieved.
6 Conclusion

The conclusion is that none of the 6 final models, which were created using OLS multiple linear regression, seen in section 4.3 and 4.4 can answer the research question:

Does a country’s societal development affect the domestic stock market?

Despite the limitations and challenges faced in developing the models, our analysis did reveal some indication of a positive relationship between societal development and the domestic stock market. However, this relationship was not clearly established due to the issues with multicollinearity, lack of statistical significance for regressors, and the need to remove influential data points.

Nevertheless, we believe that this relationship is worth exploring further through alternative modeling approaches or problem-solving strategies. By addressing the limitations of our current models, future studies could potentially provide a more accurate and reliable understanding of the relationship between societal development and the domestic stock market. Overall, this highlights the importance of continued research in this area and the need for more advanced modeling techniques to better understand the underlying dynamics of this relationship.
7 Appendix

The index data was sourced from either Wall Street Journal (WSJ)[19] or investing.com[7], see Table 17 below. The countries included in the developed nations are the ones from Switzerland down to Italy, and the developing countries are the ones from Türkiye down to Pakistan (They are separated by the thicker line, between Italy and Türkiye)

<table>
<thead>
<tr>
<th>Country</th>
<th>Earliest Ticker Year</th>
<th>HDI (2021)</th>
<th>Stock Index Ticker Name</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switzerland</td>
<td>2008</td>
<td>0.962</td>
<td>SMI</td>
<td>WSJ</td>
</tr>
<tr>
<td>Australia</td>
<td>1992</td>
<td>0.951</td>
<td>S&amp;P/ASX200</td>
<td>WSJ</td>
</tr>
<tr>
<td>Sweden</td>
<td>2008</td>
<td>0.947</td>
<td>OMXSPI</td>
<td>WSJ</td>
</tr>
<tr>
<td>Germany</td>
<td>1990</td>
<td>0.942</td>
<td>DAX</td>
<td>WSJ</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>1990</td>
<td>0.929</td>
<td>FTSE</td>
<td>WSJ</td>
</tr>
<tr>
<td>Japan</td>
<td>1990</td>
<td>0.925</td>
<td>NSE225</td>
<td>WSJ</td>
</tr>
<tr>
<td>South Korea</td>
<td>1990</td>
<td>0.925</td>
<td>KOSEI</td>
<td>WSJ</td>
</tr>
<tr>
<td>United States</td>
<td>1990</td>
<td>0.921</td>
<td>SP500</td>
<td>WSJ</td>
</tr>
<tr>
<td>France</td>
<td>1990</td>
<td>0.903</td>
<td>CAC40</td>
<td>WSJ</td>
</tr>
<tr>
<td>Italy</td>
<td>1997</td>
<td>0.895</td>
<td>FTSE MIB</td>
<td>WSJ</td>
</tr>
<tr>
<td>Türkiye</td>
<td>2008</td>
<td>0.838</td>
<td>XU100</td>
<td>WSJ</td>
</tr>
<tr>
<td>Malaysia</td>
<td>1990</td>
<td>0.803</td>
<td>FBMKLCI</td>
<td>WSJ</td>
</tr>
<tr>
<td>Mauritius</td>
<td>1990</td>
<td>0.802</td>
<td>SEMDEX</td>
<td>Investing.com</td>
</tr>
<tr>
<td>Thailand</td>
<td>2008</td>
<td>0.800</td>
<td>SET50</td>
<td>WSJ</td>
</tr>
<tr>
<td>Mexico</td>
<td>1990</td>
<td>0.758</td>
<td>IPC</td>
<td>WSJ</td>
</tr>
<tr>
<td>Brazil</td>
<td>1991</td>
<td>0.754</td>
<td>BOVESPA</td>
<td>WSJ</td>
</tr>
<tr>
<td>Egypt</td>
<td>1998</td>
<td>0.731</td>
<td>EGX30</td>
<td>Investing.com</td>
</tr>
<tr>
<td>South Africa</td>
<td>1995</td>
<td>0.713</td>
<td>FTSE/JSE</td>
<td>WSJ</td>
</tr>
<tr>
<td>Indonesia</td>
<td>1990</td>
<td>0.705</td>
<td>JKDIIX</td>
<td>WSJ</td>
</tr>
<tr>
<td>Phillipines</td>
<td>1990</td>
<td>0.6999</td>
<td>PSEE</td>
<td>WSJ</td>
</tr>
<tr>
<td>Botswana</td>
<td>2001</td>
<td>0.693</td>
<td>DCIBT</td>
<td>Investing.com</td>
</tr>
<tr>
<td>Morocco</td>
<td>2002</td>
<td>0.683</td>
<td>MASe</td>
<td>Investing.com</td>
</tr>
<tr>
<td>India</td>
<td>1990</td>
<td>0.633</td>
<td>BSESN</td>
<td>WSJ</td>
</tr>
<tr>
<td>Kenya</td>
<td>2001</td>
<td>0.575</td>
<td>NSE20</td>
<td>Investing.com</td>
</tr>
<tr>
<td>Pakistan</td>
<td>1994</td>
<td>0.544</td>
<td>KSE</td>
<td>Investing.com</td>
</tr>
</tbody>
</table>
Human development index[20]:

The HDI (Human Development Index) is a composite measure of human development that takes into account indicators such as life expectancy, education, and income. It is calculated using the following equation:

\[ HDI = \sqrt[3]{LEI \times EI \times II} \]

- **HDI** represents the Human Development Index, which provides a summary measure of human development in a country. - **LEI** refers to the Life Expectancy Index, which measures the average life expectancy at birth. - **EI** represents the Education Index, which takes into account indicators of adult literacy and combined primary, secondary, and tertiary gross enrollment ratio. - **II** denotes the Income Index, which considers the logarithm of gross national income per capita.

The HDI combines these three indices using the geometric mean (cube root) to provide a comprehensive measure of human development. It ranges from 0 to 1, where a higher value indicates higher levels of human development.

Gini Index[4]:

\[ G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}} \]

G represents the Gini Index, which quantifies income inequality. It ranges from 0 to 1, where 0 indicates perfect equality (everyone has the same income) and 1 indicates maximum inequality (one person has all the income). \( n \) refers to the number of individuals or households in the population being measured. \( x_i \) and \( x_j \) denote the income or wealth of individuals or households \( i \) and \( j \) in the population. The double summation \( \sum_{i=1}^{n} \sum_{j=1}^{n} \) computes the sum of the absolute differences between all pairs of incomes. \( |x_i - x_j| \) calculates the absolute difference between incomes \( x_i \) and \( x_j \). \( \bar{x} \) represents the mean income or wealth in the population, computed as the sum of all incomes divided by the number of individuals or households (\( \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \)).

Developed nations model countries: ['United States' 'Australia' 'United Kingdom' 'Sweden' 'France' 'Germany' 'Italy' 'Japan' 'Switzerland' 'South Korea'] Antal: 10

Developing nations model countries: ['Botswana' 'Brazil' 'Egypt' 'India' 'Indonesia' 'Kenya' 'Malaysia' 'Mauritius' 'Mexico' 'Morocco' 'Pakistan' 'Philippines' 'South Africa' 'Thailand' 'Turkey'] Antal: 15

References


