University of Virginia
Fall 2022
STAT 3220

Carson Crenshaw
Jack Penney
Rithika Nistala

THE LINGERING EFFECT OF COLONIALISM:

AN ANALYSIS OF CONTEMPORARY ECONOMIC DEVELOPMENT IN THE INTERNATIONAL WORLD

INTRODUCTION

BACKGROUND

Globalization through international trade, which emerged in the 17th century, created immense but uneven economic opportunities. Much of the expansion of international trade was done through colonialism, when a nation takes over another, often for economic gain.

Colonialism had extreme and long lasting economic effects. Most colonizing countries became extremely wealthy, while many colonized states were exploited and were not able to form robust economies, which has hindered them to the present day.

We hope to analyze the factors that influence a country's economic prosperity with an eye toward the effects of past colonialism. We will do this with a multiple linear regression model with GDP per capita as the response variable.

RESEARCH QUESTIONS

Does the effect of colonialism exist in the modern day as reflected in a country's economic performance?

Do countries that gained independence after 1930 have a lower GDP per capita?

Do countries which exist in regions of the world that contained higher numbers of colonies (experienced higher amounts of colonialism) have a lower GDP per capita?

Do countries allied with democratic states (adopted the popular western mode of government) have a higher GDP per capita?

DATA SUMMARY

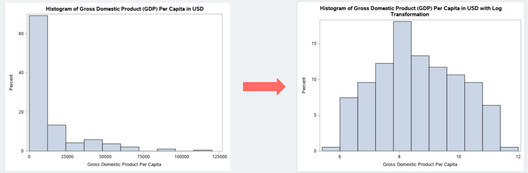
Name (Name in Code)	Description	Units/Levels
GDP per capita (GDP)	Gross domestic product is a valuation of the total goods and services produced by a country in a single year (2020) divided by the country's population.	USD ($)
World Region (WR)	Categorical geographical locations of countries in the world.	Asia and the Pacific, Europe, Africa, Americas, Western Asia
Total Population (Pop_2020)	The total number of people residing in a country (2020).	People
Average Population Age (APA)	Average age of a citizen of each country (2020).	Years
Life Expectancy (LE)	Based on national mortality rates, average lifespan of a person born in each country (2020).	Years
Independence before 1930 (IB_1930)	Whether each country was self-governing before 1930.	Y: independent before 1930 N: not independent before 1930
Membership in NATO (NATO)	Whether each country is a member of the North American Trade Organization (2020).	Y: NATO member N: NATO non-member
Global State of Democracy Index (GSoD_Index)	Generalized description of the form of government for each nation.	Authoritarian Regime, Democracy, Hybrid
Major Importer or Exporter (MIE)	Whether the value of each country's imports in 2020 is greater than its exports, or vice versa.	Major Exporter: more exports than imports Major importer: The country imports more than it exports
Unemployment Rate (UR)	Percentage of the population that is unemployed as of 2020, standard measurement of unemployment.	Percent of total population

Observation data for countries sourced from the most recent full year the data was available, 2020. Data constructed around the population of 158 countries as recognized by the CIA. These countries contain all the necessary data and do not contain any missing values.

RESEARCH

EXPLORATORY DATA ANALYSIS

Log Transformation of Response Variable



The original GDP histogram depicts a right skewed curve that is not continuous or symmetric or uniform where a majority of our observations fall within the lower GDP range (>50000). In order to correct a violation of normality, the natural log of the response variable was taken (GDP → lgdp).

Pearson Correlation Coefficient

The CORR Procedure

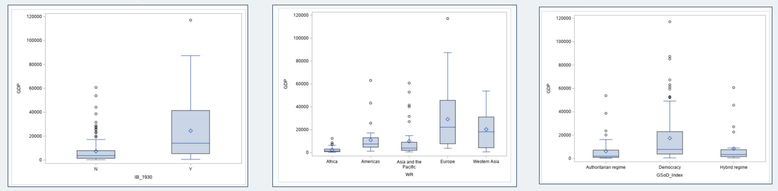
1 With Variables:	GDP
4 Variables:	Pop_2020 APA LE UR

Pearson Correlation Coefficients
Number of Observations

	Pop_2020	APA	LE	UR
GDP	-0.00698 186	0.61772 182	0.64323 180	-0.12909 168

The relationship between the quantitative variables and GDP is explored above. Life expectancy (LE) and average population age (APA) are illustrated to have strong positive relationships to a country's economic performance (GDP), while population density (Pop_2020) and unemployment rate (UR) have a neutral/low correlation. Correlation is a good initial indicator of the variables which will stay in the model.

Boxplots of Variables Relevant to Research Questions



MULTIPLE LINEAR REGRESSION

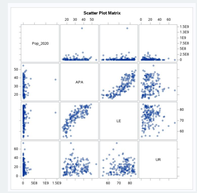
MODEL BUILDING

MULTICOLLINEARITY (with all quant. variables)

Average VIF = 2.0875 (moderate concern)
No individual VIF is greater than 10

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.67448	0.73185	-0.92	0.3582	0
APA	1	0.06516	0.01021	6.38	< 0001	3.18374
LE	1	0.09994	0.01298	7.70	< 0001	3.17117
Pop_2020	1	2.54234E-10	1.038783E-9	0.24	0.8070	1.01861
UR	1	-0.00225	0.00431	-0.52	0.6022	1.02660



STAGE 1: QUANTITATIVE (using stepwise reg.)

Initial: $lgdp = B_0 + B_1APA + B_2LE + B_3Pop\_2020 + B_4UR$
Final: $lgdp = B_0 + B_1APA + B_2LE$

STAGE 2: QUALITATIVE (t-tests and nested-tests)

Initial: $lgdp = B_0 + B_1APA + B_2LE + B_3dummyind + B_4dummyexport + B_5dummyauthor + B_6dummyhybrid + B_7dummynato + B_8dummyasia + B_9dummyafrica + B_{10}dummyameri + B_{11}dummyeuro$

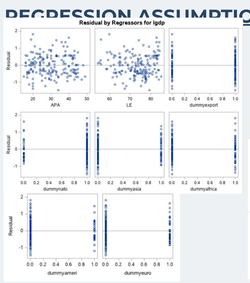
Final: $lgdp = B_0 + B_1APA + B_2LE + B_3dummyexport + B_4dummynato + B_5dummyasia + B_6dummyafrica + B_7dummyameri + B_8dummyeuro$

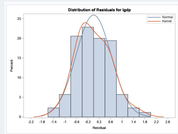
*where dummyexport = {1 if MIE = 'Major Exporter', 0 if otherwise}
*where dummynato={1 if NATO = 'N', 0 if otherwise}
*where dummyasia = {1 if WR = 'Asia and the Pacific', 0 if otherwise}
*where dummyafrica = {1 if WR = 'Africa', 0 if otherwise}
*where dummyameri = {1 if WR = 'Americas', 0 if otherwise}
*where dummyeuro = {1 if WR = 'Europe', 0 if otherwise}

STAGE 3: INTERACTIONS

No interactions were tested as none were needed.

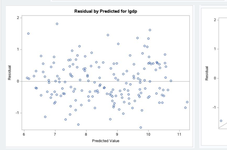
REGRESSION ASSUMPTIONS

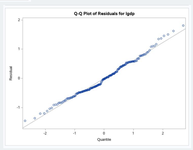




The REG Procedure
Model: MODEL1
Dependent Variable: lgdp

Durbin-Watson D	2.228
Pr < DW	0.9344
Pr > DW	0.0656
Number of Observations	175
1st Order Autocorrelation	-0.116





ADDED TECHNIQUES

Variable Screening

- Stepwise Regression
  - Removed Pop_2020 and UR
  - Kept APA and LE
- SLentry and SLstay = 0.10

All variables left in the model are significant at the 0.1000 level.

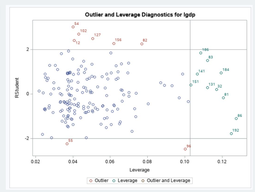
No other variable met the 0.1000 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	LE		1	0.7333	0.7333	39.9568	428.88	<.0001
2	APA		2	0.0558	0.7891	1.3706	41.02	<.0001

Influential Observation Analysis

- Removed observations:
  - 12, 54, 82, 96, 102, 127, and 156
  - Outliers
  - And also violate other jackknifing techniques: Cook's D, Deleted Studentized Residual, Student Residual



Weight: REWEIGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	272.71285	34.08911	110.03	<.0001
Error	159	49.26155	0.30982		
Corrected Total	167	321.97440			

Root MSE	0.55662	R-Square	0.8470
Dependent Mean	8.53169	Adj R-Sq	0.8393
Coeff Var	6.52410		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.84218	0.82278	1.02	0.3076
dummyexport		1	-0.45954	0.09161	-5.02	<.0001
dummynato		1	-0.38854	0.15879	-2.45	0.0155
dummyasia		1	-0.87978	0.20353	-4.32	<.0001
dummyafrica		1	-0.49785	0.22097	-2.25	0.0256
dummyameri		1	-0.72907	0.20488	-3.56	0.0005
dummyeuro		1	-1.11334	0.24925	-4.47	<.0001
Average_Population_Age	Average_Population_Age	1	0.08191	0.01161	7.05	<.0001
Life_Expectancy	Life_Expectancy	1	0.09008	0.01224	7.36	<.0001

Final Model Post-removal: Adj. R-squared goes up 0.01 (becomes 0.84) and all variables are significant at 0.05

ADDED TECHNIQUES CONT.

Model Validation

The GLMSELECT Procedure
Least Squares Model (No Selection)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	126.72496	15.84062	49.19	<.0001
Error	75	24.16045	0.32201		
Corrected Total	83	150.87540			

Root MSE	0.56746
Dependent Mean	8.49873
R-Square	0.8399
Adj R-Sq	0.8229
AIC	-0.70717
AICC	2.30653
SBC	-64.82982
ASE (Train)	0.28751
ASE (Test)	0.33462

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.115278	1.265729	-0.09	0.9277
dummyexport	1	-0.502988	0.138107	-3.64	0.0005
dummynato	1	-0.469745	0.225962	-2.08	0.0410
dummyasia	1	-0.553836	0.276878	-2.00	0.0491
dummyafrica	1	-0.220266	0.292628	-0.75	0.4540
dummyameri	1	-0.734952	0.275668	-2.67	0.0094
dummyeuro	1	-1.215434	0.345565	-3.52	0.0007
Average_Population_A	1	0.084059	0.017643	4.76	<.0001
Life_Expectancy	1	0.101783	0.019286	5.28	<.0001

The technique used for the purpose of this project was data-splitting or cross validation:

- Approximately 50% of the data was split into a training dataset and approximately 50% was split into a testing dataset. The data meets the qualification requirements in which the entire sample should consist of at least n = 2k + 25 observations (41), where k is the number of beta parameters in the model.
- Using cross validation, the relatively high predictive power of the model is confirmed using the testing and training datasets, concluding the analysis with an R-squared value of 0.84. This regression model reveals that 84% of the variability observed in gross domestic product is explained by the regression model.

CONCLUSION

INTERPRETATION

$$\widehat{log(GDP)} = 0.842 + 0.082APA + 0.090LE - 0.460Dummyexport - 0.389Dummynato - 0.880Dummyasia - 0.498Dummyafrica - 0.729Dummyameri - 1.113Dummyeuro$$

- The final model was found to be statistically significant with a p-value of <0.0001 and an adjusted R-squared of 0.839. This means that average population age, life expectancy, major importer or exporter, and regional location of the country are altogether significant predictors of GDP per capita.
- A higher average population age and life expectancy correspond with a higher GDP per capita; net importers also have higher GDP per capita.
- Membership in NATO and location in Asia, Africa, America, and Europe decrease GDP per capita. One should consider, however, that NATO members and European countries do not have low GDPs per capita, as the negativity of the coefficients may indicate at first glance. Instead, it implies that if there were two countries, with one in Europe and one in Africa with otherwise identical values, the African country is expected to have the higher GDP per capita.

Examples

Country Name	Model Prediction	Actual Value	Residual	Why this country?
Ghana	1785.404	2225.506	440.102	Relevance to research questions (Post-1930 independence, African region, Democracy)
Samoa	3928.135	4067.815	139.68	Randomly chosen to assess fit

FURTHER RESEARCH

- There are a multitude of other factors that may influence GDP per capita, especially variables that are related to past colonialism: mineral resources, coastline, average temperature, etc. These variables could be combined with the factors used in our current model.
- Different measurements of variables currently in our dataset may lead us to more accurate conclusions. For example, the power of each country's passport could be a more accurate measure of citizens' freedom than the nation's form of government.
- Our current model is limited in its scope to answer the research questions since the final model didn't include two of the three variables of interest (independence before 1930 and mode of government). Further analysis could be conducted with those variables in isolation.
- As a final limitation, a more accurate interpretation of the model would include the exponentiation of our coefficients to reverse the effect of the log transformation. This process was undertaken in the prediction examples to generate GDP instead of log(GDP) values, but involves a more complex process in the analysis of the model variables. The sign of the coefficients remains the same, but the amount of change will be altered through this process.

Works Cited

Acemoğlu, Daron, and James Robinson. “The Economic Impact of Colonialism.” *CEPR*, 30 Jan. 2017, https://new.cepr.org/voxeu/columns/economic-impact-colonialism.

“Data Set and Resources.” International IDEA, https://www.idea.int/gsod-indices/dataset-resources.

Databank, https://databank.worldbank.org/metadataglossary/statistical-capacity-indicators/series/5.51.01.10.gdp#:~:text=GDP%20per%20capita%20is%20the,divided%20by%20mid%2Dyear%20population.

“Finance & Development.” *Finance & Development | F&D*, https://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm.

“GDP per Capita (Current US$).” *Data*, https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2020&start=1960.

“Independence.” *Central Intelligence Agency*, Central Intelligence Agency, https://www.cia.gov/the-world-factbook/field/independence/.

“Life Expectancy at Birth, Total (Years).” *Data*, https://data.worldbank.org/indicator/SP.DYN.LE00.IN.

“Median Age.” *Central Intelligence Agency*, Central Intelligence Agency, https://www.cia.gov/the-world-factbook/field/median-age/country-comparison.

Nato. “Member Countries.” NATO, https://www.nato.int/cps/en/natohq/nato_countries.htm.

“Population, Total.” *Data*, https://data.worldbank.org/indicator/SP.POP.TOTL.

Sen, Kunal. *Unu-Wider : Blog : Why Should I Care about Economic Growth?* Jan. 2021, https://www.wider.unu.edu/publication/why-should-i-care-about-economic-growth.

“Top Exporters and Importers.” *Top Exporters and Importers by Country and Region 2020 | WITS Data*, https://wits.worldbank.org/CountryProfile/en/Country/WLD/Year/LTST/TradeFlow/EXPIMP.

“Unemployment, Youth Ages 15-24.” *Central Intelligence Agency*, Central Intelligence Agency, https://www.cia.gov/the-world-factbook/field/unemployment-youth-ages-15-24.

“What Is Globalization?” *PIIE*, 17 Aug. 2022, https://www.piie.com/microsites/globalization/what-is-globalization.