Exploration into the Prediction of National R&D Expenditures

University of Texas at Dallas

Glen Cooper

for

Dr. Ho - Spring 2022 EPPS 6323

Introduction

There are 195 countries in the world today (Kershner, 2020). Knowing which counties will spend on research and development (R&D) and how much they will spend can help investors, world leaders, and innovation economists know where to focus their attention.

New inventions and ideas drive economic development in today's dynamic and innovation-driven world. R&D expenditures drive innovation and can be a proxy for where innovation is occurring. Considering the significant impact innovations have had on all facets of the modern world and on individual countries' economies, predicting where it is likely to occur could be an important step to understanding the nature of the modern world. This paper will explore the ability to predict a country's relative R&D expenditures from a country's level of democratic policies, encouragement of entrepreneurship, business friendliness, market dynamics, and legal structure as drivers of R&D expenditures.

This paper will be divided into the following sections: 1. Research question, this section will provide the specific statement of the paper's purpose; 2. Methodological approach and data analysis, this section will discuss the approach pursued and provide a review of the data used; 3. Model specifications, this section will provide the specifications for the predictive modeling; 4. Model fit and comparisons/estimations, this section will review the results of the models and the estimations developed; 5. Conclusion, this last section will discuss the overall conclusions reached.

Research Question

How much will a country spend, as a percent of GDP, on R&D each year?

A set of 285 observations covering 34 predictors and 1 response variable was developed to investigate this question. The response variable is the amount a country spends annually on R&D as a percentage of GDP. Short descriptions of the 34 predictor variables are provided in the table below:

Table of Predictor Variables

Degree of Democracy	
Democracy Index	
Entrepreneurial Cultural	
Business Services Sector	High Job Creation Expectation
Cultural and social norms	High Status to Successful Entrepreneurs
Entrepreneurial intentions	Perceived Entrepreneurial capabilities
Entrepreneurship as a Good Career Choice	Perceived Entrepreneurial opportunities
Fear of failure rate	Total early-stage Entrepreneurial Activity
Female/Male in Entrepreneurship	
Governmental Entrepreneurial Support	
Financing for entrepreneurs	Governmental support and policies for entrepreneurs
Governmental programs for entrepreneurs	Governmental R&D transfer
Infrastructure	
Fixed telephone subscriptions (per 100 people)	Percentage of Individuals using the Internet
Getting electricity Score	Quality of physical and services infrastructure
Legal environment	
Commercial and professional infrastructure	Score-Ease of shareholder suits index
Ease of doing business score	Score-Extent of director liability index
Extent of director liability index	Score-Protecting minority investors
Internal market openness	Score-Starting a business
Ease of shareholder suits index	Strength of legal rights index
Market Dynamics	
Internal market dynamics	Trade (% of GDP)
Other	
Patents Office & Patents Families	Proportion of seats held by women in national parliaments

Data Collection and Transformation

Data Collection

All data were collected from six sources: United Nations Educational, Scientific and Cultural Organization (UNESCO) / Institute for Statistics (UIS), The Economist Intelligence Unit (EIU), Global Entrepreneurship Monitor (GEM), United Nations' (UN), International Telecommunication Union (ITU), World Bank, and Organization for Economic Cooperation and Development (OECD) (see Appendix 1 for details on sources and Appendix 2 for detail variable descriptions).

Data Transformation

The original data set consisted of over 450 observations, 25 countries, and 55 parameters. Some countries had no response variable data and had to be removed upon detailed review. In addition, some of the remaining countries did not have sufficient yearly data to keep all years for that country. Below is a list of countries and year ranges for the remaining data set:

Country	Beginning Year	Ending Year
Argentina	1999	2017
Brazil	2000	2017
Canada	1999	2018
China	2002	2017
Colombia	2006	2018
Estonia	2012	2017
France	1999	2017
Germany	1999	2017
India	1999	2018
Indonesia	2013	2018
Italy	1999	2017
Japan	1999	2017

Country	Beginning Year	Ending Year
Malaysia	2006	2016
Russia	2006	2017
Saudi Arał	2009	2010
Singapore	1999	2014
South Afri	2001	2016
South Kor	1999	2017
Turkey	2006	2017
UAE	2011	2018
UK	1999	2017
US	1999	2017
Vietnam	2013	2017

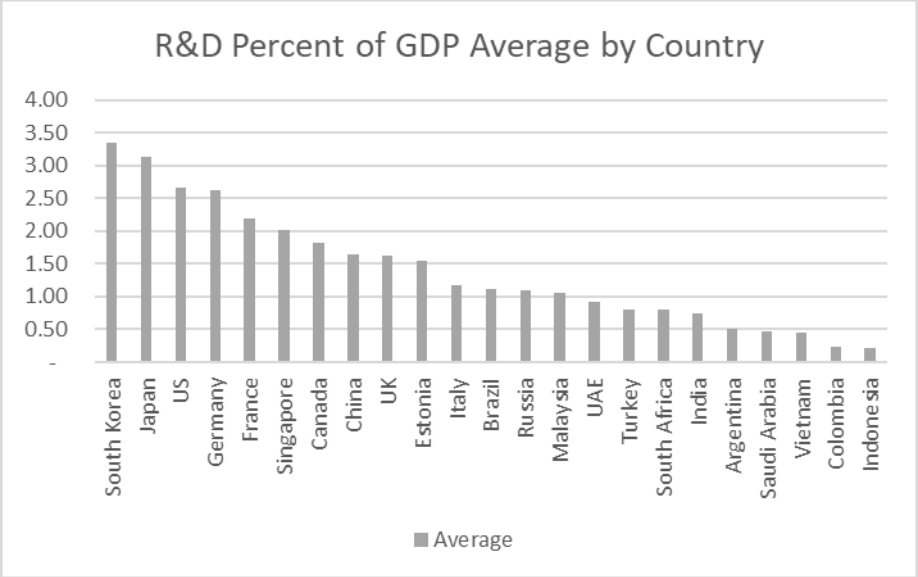
From the potential parameters, those with insufficient data to support inclusion were removed. Certain predictor variables did not have sufficient data to cover the full range of available years. However, because these predictors were generally consistent within a country but had reasonable variance between countries, the average value for each country was calculated. That value was applied to all years within the country. This average country value transformation was applied to the following variables:

Variable Name	Short Description
FXTELSUB	Fixed telephone subscriptions (per 100 people)
GETELEC	Getting electricity (DB10-15 & 20 methodology) - Score
EASEBUS	Ease of doing business score (DB14, 15 & 20 methodology)
MINEXTDIRIN	Extent of director liability index (0-10)
MINSUITIND	Protecting minority investors: Ease of shareholder suits index
MINSUITSCR	Score-Ease of shareholder suits index (0-10) (DB06-14 methodology)
MINEXTDIRSCR	Score-Extent of director liability index (0-10)
MINRTS	Score-Protecting minority investors (DB06-14 & 20 methodology)
STARTBUSSCR	Score-Starting a business
STRGLEG	Strength of legal rights index (0=weak to 12=strong)

All other variables had sufficient values to recommend inclusion. However, missing values did appear within the data set in certain cases. In cases where the data was missing at the beginning or end of a country series, the first or last values were repeated; generally, no more than two periods required this “repeat” procedure. When gaps appeared within the data set and between country years, the average of the values between the periods was used to extrapolate the missing data.

Exploratory Data Analysis

Below is a graphic of the by country average R&D as a percent of GDP.



As is apparent, there is considerable variation in these values. Among all the countries, the R&D response variable ranges from a minimum of 0.07 to a maximum of 4.30, with the average value being 1.58. For a summary of predictor minimum, maximum, and average values, see Appendix 3.

Generally, correlations greater than 70% to 80% can create problems with regression analysis. Therefore, the below correlation matrix was developed to identify correlations above 70%:

Greater than 70% Correlation

	EFGSP	*EFRD*	*EFBSS*	*EASEBUS*	*MINRTS*	*MINSUITIND*	*MINEXTDIRIN*
EFGP	71%						
EFIMO		77%					
MINSUITIND					78%		
MINSUITSCR					78%	100%	
MINEXTDIRIN					80%		
MINEXTDIRSCR					80%		100%
STARTBUSSCR				82%			
FXTELSUB			77%	73%			

The correlations appear reasonable. Note, however, that there are two perfectly correlated variables. One of the variables in of these perfectly correlated variables will have to be dropped in certain cases because a perfect linear correlation will stop some regression models from calculating. For example, forward and backward stepwise regressions can encounter this difficulty. Other regression methodologies, such as Ridge and Lasso regressions, will not be impacted by perfectly correlated variables but will feature them as part of the regression algorithm. The final dataset included

Methodological Approach

Introduction

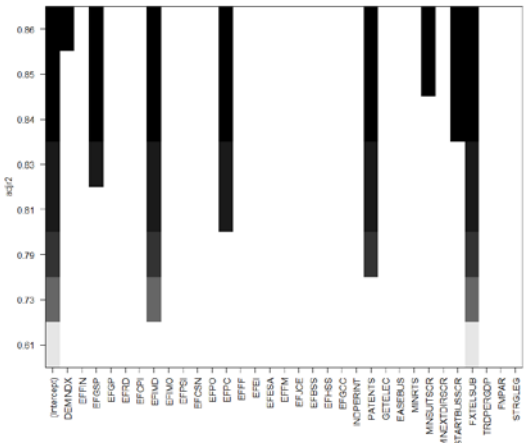
To predict the response variable, R&D as a percent of GDP, the seven modeling techniques outlined by James, Witten, Hastie & Tibshirani (2013) were applied to the predictor variables. These techniques included three subset selection models (i.e., Forward stepwise regression, Backward stepwise regression, & Best subset regression), two shrinkage methods (i.e., Ridge regression & Lasso regression), and two dimension reduction methods (i.e., Principal components regression & Partial least

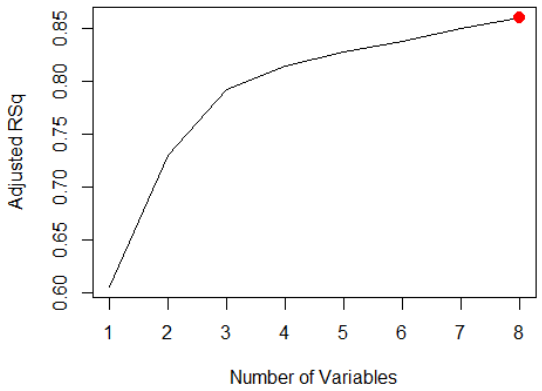
squares). Below are the results of these analyzes (see supporting R code in Appendix 4).

Forward Stepwise Regression

Forward stepwise selection is an efficient regression algorithm that begins with no predictors, only the intercept. Then, it adds one predictor to the model, incorporating only those predictors that improve the model (James, Witten, Hastie & Tibshirani, 2013, p 207).

The adjusted r-square statistic measures model fit that penalizes for increasing the number of model predictors (James, Witten, Hastie & Tibshirani, 2013, p 212). Generally, the adjusted r-square statistic represents the percentage of variation in the response variable that is explained by the predictor variables. Using the forward stepwise regression approach, the dark line bars over the variable names in the bar chart indicate those predictors that had the greatest impact on the adjusted r-squared statistic, and the line graph demonstrates that the best-adjusted r-squared is obtained using eight predictors.





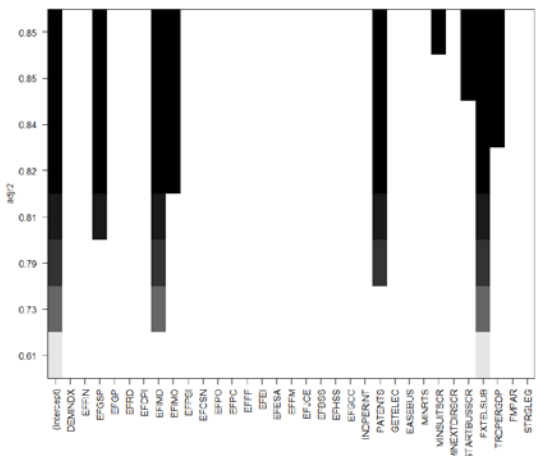
Below are the coefficients of the final model produced:

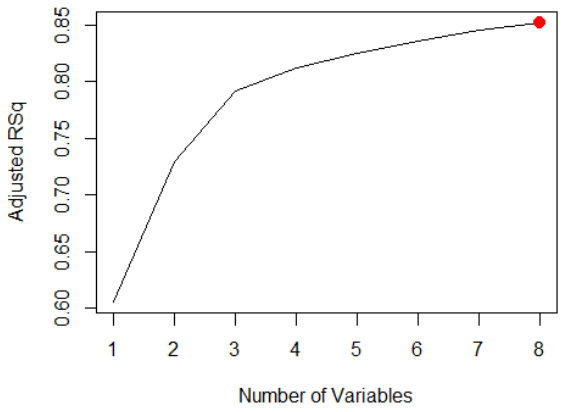
Variable	Coefficient
(Intercept)	0.97
DEMINDX	-0.07
EFGSP	0.31
EFIMD	0.37
EFPC	-0.01
PATENTS	0.00
MINSUITSCR	0.01
STARTBUSSCR	-0.03
FXTELSUB	0.04

The r-squared statistic measures model performance that does not penalize for increasing the number of predictors but is often cited as an indicator of what percentage a model explains the variation in the response variable. Here the r-squared value for this model was: 86%.

Backward Stepwise Regression

The backward stepwise regression is similar to the forward stepwise regression, except it begins with the full least squares model containing all the predictors and then removes the least useful predictors one by one (James, Witten, Hastie & Tibshirani, 2013, p 209). Like the forward stepwise regression, the backward stepwise regression identified eight predictor variables, each with a similar impact on the adjusted r-squared statistic as indicated in the below graphics:





Below are the coefficients of the final model produced:

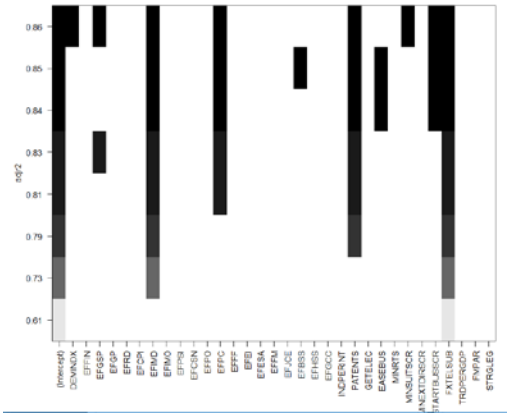
Variable	Coefficient
(Intercept)	-0.23
EFGSP	0.31
EFIMD	0.48
EFIMO	-0.27
PATENTS	0.00
MINSUITSCR	0.01
STARTBUSSCR	-0.02
FXTELSUB	0.04
TRDPERGDP	0.00

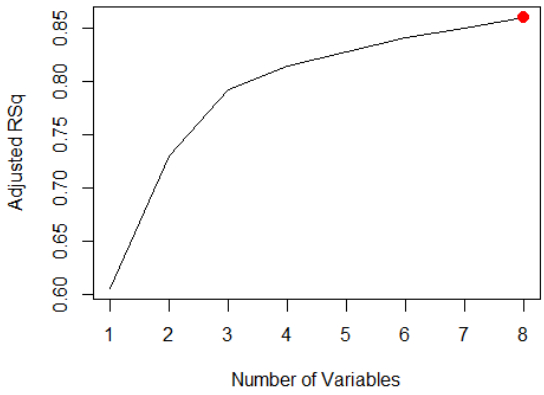
Note that the forward and backward regressions did not identify the same predictors for inclusion. The r-squared value for this model was: 86%.

Best Subset Regression

The best subset regression algorithm fits least square regressions for all possible combinations of predictors. However, this can be computationally burdensome as a total of 2^p, where p is the number of predictors, models must be calculated (James, Witten, Hastie & Tibshirani, 2013, p 205). Here, after excluding the two perfectly linear variables, that is 2^{32}or over 4 billion models.

This model again identified eight predictors, as noted in the graphics below:





Below are the coefficients of the final model produced:

Variable	Coefficient
(Intercept)	0.97
DEMINDX	-0.07
EFGSP	0.31
EFIMD	0.37
EFPC	-0.01
PATENTS	0.00
MINSUITSCR	0.01
STARTBUSSCI	-0.03
FXTELSUB	0.04

Again the best subset regression selected a couple of different predictors than either the forward or backward regressions. However, like both of those regressions, the best subset regression returned an r-squared value of 86%.

Ridge Regression

Ridge regression is a class of models known as shrinkage models. Ridge regression is similar to least square regression, but coefficients are estimated by not only the residual sum of squares values but also a tuned sum of the model's coefficients squared (James, Witten, Hastie & Tibshirani, 2013, p 215). Using this approach, coefficients,

except for the intercept, are “encouraged” to shrink to a zero value. The tuning can be selected through a cross-validation process. Using this method, a tuning parameter of log value -4.6 was specified. Here are the coefficients identified using this approach:

Variable	Coefficient	Variable	Coefficien
(Intercept)	3.59	GETELEC	-0.01
EFCPI	-0.27	EFFF	-0.01
EFIMO	-0.16	EFESA	-0.01
EFFM	-0.13	EFJCE	-0.01
DEMINDX	-0.11	EFPC	-0.01
STRGLEG	-0.06	EFBSS	-0.01
EFGP	-0.05	FMPAR	0.00
STARTBUSSCR	-0.04	PATENTS	0.00
MINRTS	-0.03	INDPERIN	0.00
EFGCC	-0.02	TRDPERGL	0.00
MINSUITSCR	0.02	MINEXTDI	0.01
EASEBUS	0.03	EFEI	0.01
FXTELSUB	0.04	EFPO	0.01
EFFIN	0.05	EFHSS	0.01
EFCSN	0.11		
EFRD	0.12		
EFIMD	0.12		
EFPSI	0.12		
EFGSP	0.21		

Notice how many of the coefficients, separated on the right-hand side, are at or near zero. To evaluate this model’s performance, a cross-validation approach was used. That approach requires the data set to be divided between a “training” set that builds the model coefficient values and a “test” set that is used to compare the prediction that the “trained” model provides against the “test” response values. Using this approach, an r-squared value for this model was 90%.

Lasso Regression

Lasso regression is similar to ridge regression, but it can allow predictor coefficients to actually become zero. Something that the ridge regression will not do. The Lasso regression does this by applying a coefficient penalty of the absolute value of the

coefficient adjusted for a tuning parameter rather than the coefficient squared adjustment used in ridge regression (James, Witten, Hastie & Tibshirani, 2013, p 219). Like the ridge regression, the Lasso regression produced a log tuning parameter of -4.6. Here are the coefficients identified using this approach:

Lasso Regression

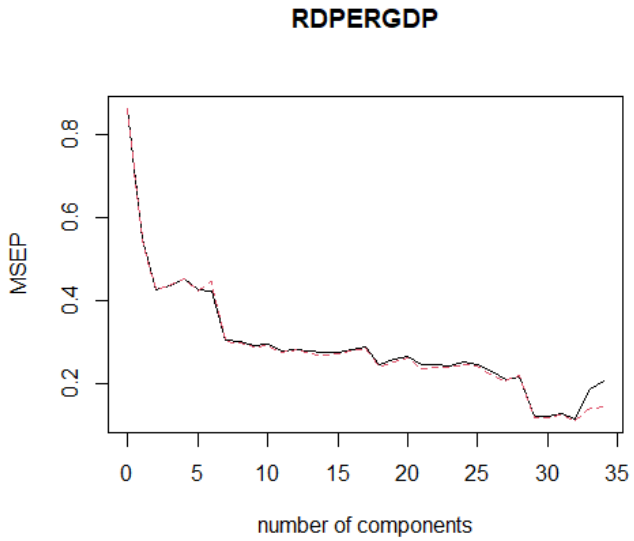
Variable	Coefficient	Variable	Coefficient
(Intercept	2.14	MINEXTDIRSCR	0.00
EFIMD	0.22	EFPO	0.00
EFGSP	0.21	EASEBUS	0.00
EFPSI	0.10	TRDPERGDP	0.00
FXTELSUB	0.04	FMPAR	0.00
MINSUITS	0.01	INDPERINT	0.00
EFHSS	0.01	PATENTS	0.00
EFFF	-0.01	EFJCE	0.00
EFBSS	-0.01	EFPC	0.00
STRGLEG	-0.01	EFFIN	-
EFGCC	-0.01	EFGP	-
MINRTS	-0.01	EFRD	-
STARTBUS	-0.02	EFCSN	-
DEMINDX	-0.08	EFEI	-
EFIMO	-0.08	EFESA	-
EFCPI	-0.18	EFFM	-
		GETELEC	-

Note here that many coefficients were near zero; some were at zero. Using this approach, an r-squared value for this model was 88%.

Principal Components Analysis (PCA)

PCA is a dimension reduction method. This method aims to reduce predictors used by the model by combining like predictors into a single predictor variable. (James, Witten, Hastie & Tibshirani, 2013, p 228). The idea behind PCR is to reduce the predictor variables into a small number of components, M, and then fix the linear regression model on those M components (James, Witten, Hastie & Tibshirani, 2013, p 223). Using this method did not cause the number of predictors to drop. As noted in the

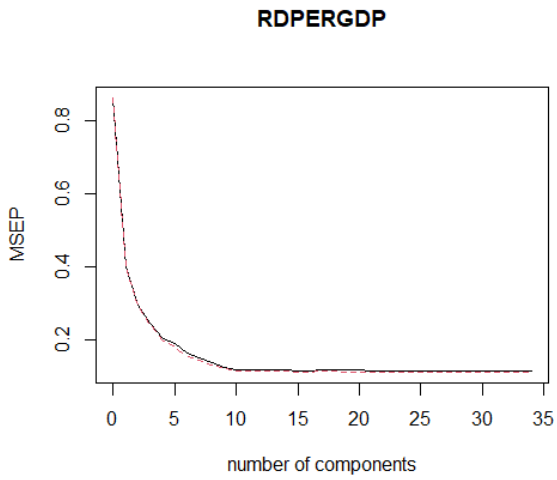
below graphic, the number of components remained near 30 as the mean square error declined:



This model produced an r-squared value of 89%.

Partial Least Squares (PLS)

The PLS approach is similar to the PCA approach, except it uses the response variable in a supervising way to help improve the model components (James, Witten, Hastie & Tibshirani, 2013, p 237). Using this approach did show a dramatic drop in the number of components. Here ten components appeared sufficient to improve the MSE. See the graphic below:



Using this approach produced an r-squared value of 89%

Conclusion

All models produced a similar r-squared value:

Model Type	R-Square
Forward regression	86%
Backward regression	86%
Best subset regression	86%
Ridge regression	90%
Lasso regression	88%
PCR	89%
PLS	89%

Ridge regression had the best predictive power with a 90% r-squared value, followed closely by PCR, PLS, and Lasso. However, the basic linear regressions (forward, backward, and best subset) models also produced a reasonable level of predictive power. In conclusion, I believe that the simplicity of the basic linear regression models argues for their use in this situation.

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Kershner, E. (2020, August 26). How many countries are there in the world? WorldAtlas. Retrieved May 6, 2022, from https://www.worldatlas.com/articles/how-many-countries-are-in-the-world.html

Appendix 1 - Data Sources

Below is the table of data sources:

Source	Variable Class	Variable Name	Short Description	Source Link
UNESCO / UIS.Stat				
	Response	RDPERGDP	Total R&D as a percentage of GDP	https://en.unesco.org/
The Economist Intelligence Unit				
	Democracy Index	DEMINDX	Economist Intelligence Unit's Democracy Index	https://infographics.economist.com/2018/DemocracyIndex/
Global Entrepreneurship Monitor				
	Entrepreneurial Cultural	EFBSS	Business Services Sector	https://www.gemconsortium.org/data/key-aps
	Entrepreneurial Cultural	EFCSN	Cultural and social norms	https://www.gemconsortium.org/data/key-aps
	Entrepreneurial Cultural	EFEI	Entrepreneurial intentions	https://www.gemconsortium.org/data/key-aps
	Entrepreneurial Cultural	EFGCC	Entrepreneurship as a Good Career Choice	https://www.gemconsortium.org/data/key-aps
	Entrepreneurial Cultural	EFFF	Fear of failure rate	https://www.gemconsortium.org/data/key-aps
	Entrepreneurial Cultural	EFFM	Female/Male TEA	https://www.gemconsortium.org/data/key-aps
	Entrepreneurial Cultural	EFJCE	High Job Creation Expectation	https://www.gemconsortium.org/data/key-aps
	Entrepreneurial Cultural	EFHSS	High Status to Successful Entrepreneurs	https://www.gemconsortium.org/data/key-aps
	Entrepreneurial Cultural	EFPC	Perceived capabilities	https://www.gemconsortium.org/data/key-aps
	Entrepreneurial Cultural	EFPO	Perceived opportunities	https://www.gemconsortium.org/data/key-aps
	Entrepreneurial Cultural	EFESA	Total early-stage Entrepreneurial Activity (TEA)	https://www.gemconsortium.org/data/key-aps
	Governmental Entrepreneurial Support	EFFIN	Financing for entrepreneurs	https://www.gemconsortium.org/data/key-aps
	Governmental Entrepreneurial Support	EFGP	Governmental programs	https://www.gemconsortium.org/data/key-aps
	Governmental Entrepreneurial Support	EFGSP	Governmental support and policies	https://www.gemconsortium.org/data/key-aps
	Governmental Entrepreneurial Support	EFRD	R&D transfer	https://www.gemconsortium.org/data/key-aps
	Infastructure	EFPSI	Physical and services infrastructure	https://www.gemconsortium.org/data/key-aps
	Legal environment	EFCPI	Commercial and professional infrastructure	https://www.gemconsortium.org/data/key-aps
	Legal environment	EFIMO	Internal market openness	https://www.gemconsortium.org/data/key-aps
	Market Dynamics	EFIMD	Internal market dynamics	https://www.gemconsortium.org/data/key-aps
UN's International Telecommunication Union (ITU)				
	Infastructure	INDPERINT	Percentage of Individuals using the Internet	https://www.itu.int/en/ITU-D/Pages/default.aspx
World Bank				
	Infastructure	FXTELSUB	Fixed telephone subscriptions (per 100 people)	https://datacatalog.worldbank.org/public-licenses#cc-by
	Infastructure	GETELEC	Getting electricity (DB10-15 & 20 methodology) - Score	https://datacatalog.worldbank.org/public-licenses#cc-by
	Legal environment	EASEBUS	Ease of doing business score (DB14, 15 & 20 methodology)	https://datacatalog.worldbank.org/public-licenses#cc-by
	Legal environment	MINEXTDIRIN	Extent of director liability index (0-10)	https://datacatalog.worldbank.org/public-licenses#cc-by
	Legal environment	MINSUITIND	Protecting minority investors: Ease of shareholder suits index	https://datacatalog.worldbank.org/public-licenses#cc-by
	Legal environment	MINSUITSCR	Score-Ease of shareholder suits index (0-10) (DB06-14 methodology)	https://datacatalog.worldbank.org/public-licenses#cc-by
	Legal environment	MINEXTDIRSCR	Score-Extent of director liability index (0-10)	https://datacatalog.worldbank.org/public-licenses#cc-by
	Legal environment	MINRTS	Score-Protecting minority investors (DB06-14 & 20 methodology)	https://datacatalog.worldbank.org/public-licenses#cc-by
	Legal environment	STARTBUSSCR	Score-Starting a business	https://datacatalog.worldbank.org/public-licenses#cc-by
	Legal environment	STRGLEG	Strength of legal rights index (0=weak to 12=strong)	http://www.doingbusiness.org/
	Market Dynamics	TRDPERGDP	Trade (% of GDP)	https://datacatalog.worldbank.org/public-licenses#cc-by
	Other	FMPAR	Proportion of seats held by women in national parliaments (%)	www.ipu.org
Organisation for Economic Cooperation and Development (OECD)				
	Other	PATENTS	Patents Office & Patents Families - Triadic Patent Families	https://stats.oecd.org/

Appendix 2 – Detail Data Descriptions

Below are the tables providing detailed descriptions of the data:

Source	Short Description	Detailed Description
UNESCO / UIS.Stat		
	Total R&D as a percentage of GDP	R&D expenditures as a percent of GDP
The Economist Intelligence Unit		
	Economist Intelligence Unit's Democracy Index	Index of democratic policies
Global Entrepreneurship Monitor		
	Business Services Sector	Percentage of those involved in TEA in the Business Services sector Information and Communication, Financial Intermediation and Real Estate, Professional Services or Administrative Services, as defined by the ISIC 4.0 Business Type Codebook
	Cultural and social norms	The extent to which social and cultural norms encourage or allow actions leading to new business methods or activities that can potentially increase personal wealth and income
	Entrepreneurial intentions	Percentage of 18-64 population (individuals involved in any stage of entrepreneurial activity excluded) who are latent entrepreneurs and who intend to start a business within three years
	Entrepreneurship as a Good Career Choice	Percentage of 18-64 population who agree with the statement that in their country, most people consider starting a business as a desirable career choice
	Fear of failure rate	Percentage of 18-64 population perceiving good opportunities to start a business who indicate that fear of failure would prevent them from setting up a business
	Female/Male TEA	Percentage of female 18-64 population who are either a nascent entrepreneur or owner-manager of a new business, divided by the equivalent percentage for their male counterparts
	High Job Creation Expectation	Percentage of those involved in TEA who expect to create 6 or more jobs in 5 years
	High Status to Successful Entrepreneurs	Percentage of 18-64 population who agree with the statement that in their country, successful entrepreneurs receive high status
	Perceived capabilities	Percentage of 18-64 population who believe they have the required skills and knowledge to start a business
	Perceived opportunities	Percentage of 18-64 population who see good opportunities to start a firm in the area where they live
	Total early-stage Entrepreneurial Activity (TEA)	Percentage of 18-64 population who are either a nascent entrepreneur or owner-manager of a new business
	Financing for entrepreneurs	The availability of financial resourcesâ - "equity and debt" - for small and medium enterprises (SMEs) (including grants and subsidies)
	Governmental programs	The presence and quality of programs directly assisting SMEs at all levels of government (national, regional, municipal)
	Governmental support and policies	The extent to which public policies support entrepreneurship - entrepreneurship as a relevant economic issue
	R&D transfer	The extent to which national research and development will lead to new commercial opportunities and is available to SMEs
	Physical and services infrastructure	Ease of access to physical resources, communication, utilities, transportation, land or space at a price that does not discriminate against SMEs
	Commercial and professional infrastructure	The presence of property rights, commercial, accounting and other legal and assessment services and institutions that support or promote SMEs
	Internal market openness	The extent to which new firms are free to enter existing markets
	Internal market dynamics	The level of change in markets from year to year
UN's International Telecommunication Union (ITU)		
	Percentage of Individuals using the Internet	Percentage of country individuals using the internet

Source	Short Description	Detailed Description
World Bank		
	Fixed telephone subscriptions (per 100 people)	Fixed telephone subscriptions refers to the sum of active number of analogue fixed telephone lines, voice-over-IP (VoIP) subscriptions, fixed wireless local loop (WLL) subscriptions, ISDN voice-channel equivalents and fixed public payphones.
	Getting electricity (DB10-15 & 20 methodology) - Score	The score for getting electricity is the simple average of the scores for each of the component indicators: the procedures, time, cost for a business to obtain a permanent electricity connection and supply for a standardized warehouse, as well as the reliability of supply and transparency of tariffs index. The score is computed based on the methodology in the DB10-15 studies.
	Ease of doing business score (DB14, 15 & 20 methodology)	The ease of doing business score is the simple average of the scores for each of the Doing Business topics: starting a business, dealing with construction permits, getting electricity, registering property, getting credit, protecting minority investors, paying taxes, trading across borders, enforcing contracts and resolving insolvency. The score is computed based on the methodology in the DB10-14 studies for topics that underwent methodology updates.
	Extent of director liability index (0-10)	The extent of director liability index measures when board members can be held liable for harm caused by related-party transactions and what sanctions are available. It has seven components: (i) whether shareholders can sue directly or derivatively for the damage the transaction causes to the company; (ii) whether a shareholder plaintiff can hold Mr. James liable for the damage the Buyer-Seller transaction causes to the company; (iii) whether a shareholder plaintiff can hold other executives and directors (the CEO, members of the board of directors or members of the supervisory board) liable for the damage the transaction causes to the company; (iv) whether Mr. James pays damages for the harm caused to the company upon a successful claim by the shareholder plaintiff; (v) whether Mr. James repays profits made from the transaction upon a successful claim by the shareholder plaintiff; (vi) whether Mr. James is disqualified upon a successful claim by the shareholder plaintiff; and (vii) whether a court can void the trans-action upon a successful claim by a shareholder plaintiff.
	Protecting minority investors: Ease of shareholder suits index (0-10) (DB06-14 methodology) - Score	The ease of shareholder suits index measures how likely plaintiffs are to access internal corporate evidence. It has six components: (i) whether shareholders owning 10% of the company's share capital have the right to inspect the Buyer-Seller transaction documents before filing a suit; (ii) whether shareholders owning 10% of the company's share capital can request that a government inspector investigate the Buyer-Seller transaction without filing a suit; (iii) what range of documents is available to the shareholder plaintiff from the defendant and witnesses during trial; (iv) whether the plaintiff can obtain cate-gories of relevant documents from the defendant without identifying each document specifically; (v) whether the plaintiff can directly examine the defendant and witnesses during trial (0-2); and (vi) whether the standard of proof for civil suits is lower than that for criminal cases. The index is computed based on the methodology in the DB06-14 studies.
	Score-Ease of shareholder suits index (0-10) (DB06-14 methodology)	The score for ease of shareholder suits index benchmarks economies with respect to the regulatory best practice on the indicator. The score is indicated on a scale from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance, and is computed based on the methodology in the DB06-14 studies.
	Score-Extent of director liability index (0-10)	The score for extent of director liability index benchmarks economies with respect to the regulatory best practice on the indicator. The score is indicated on a scale from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance.
	Score-Protecting minority investors (DB06-14 & 20 methodology)	The score for protecting minority investors benchmarks economies with respect to the regulatory best practice on the indicator set. The score is indicated on a scale from 0 to 100, where 0 represents the worst regulatory performance and 100 the best regulatory performance, and is computed based on the methodology in the DB06-14 studies.
	Score-Starting a business	The score for starting a business is the simple average of the scores for each of the component indicators: the procedures, time and cost for an entrepreneur to start and formally operate a business, as well as the paid-in minimum capital requirement.
	Strength of legal rights index (0=weak to 12=strong)	Strength of legal rights index measures the degree to which collateral and bankruptcy laws protect the rights of borrowers and lenders and thus facilitate lending. The index ranges from 0 to 12, with higher scores indicating that these laws are better designed to expand access to credit.
	Trade (% of GDP)	Trade is the sum of exports and imports of goods and services measured as a share of gross domestic product.
	Proportion of seats held by women in national parliaments (%)	Women in parliaments are the percentage of parliamentary seats in a single or lower chamber held by women.
Organisation for Economic Cooperation and Development (OECD)		
	Patents Office & Patents Families - Triadic Patent Families	Number of patents granted each year by a country

Appendix 3 – Predictor Values Quantitative Summary

The below table summarizes the predictor variable minimum, maximum, and average values:

Variable Name	Short Description	Min	Max	Average
DEMINDX	Democracy Index	1.86	9.09	6.83
EFFIN	Financing for entrepreneurs	1.57	4.30	2.70
EFGSP	Governmental support and policies	1.37	3.79	2.74
EFGP	Governmental programs	1.42	3.71	2.67
EFRD	R&D transfer	1.68	3.83	2.55
EFCPI	Commercial and professional infrastructure	1.94	4.21	3.00
EFIMD	Internal market dynamics	1.85	4.40	3.15
EFIMO	Internal market openness	1.86	4.56	2.71
EFPSI	Physical and services infrastructure	1.91	4.79	3.66
EFCSN	Cultural and social norms	2.01	4.59	2.95
EFPO	Perceived opportunities	5.25	75.84	34.56
EFPC	Perceived capabilities	8.65	72.98	42.59
EFFF	Fear of failure rate	13.91	65.32	35.40
EFEI	Entrepreneurial intentions	0.75	62.56	15.62
EFESA	Total early-stage Entrepreneurial Activity (TEA)	1.48	27.35	9.78
EFFM	Female/Male TEA	0.09	1.34	0.63
EFJCE	High Job Creation Expectation	1.62	71.51	22.86
EFBSS	Business Services Sector	1.30	42.64	19.70
EFHSS	High Status to Successful Entrepreneurs	31.47	92.26	69.05
EFGCC	Entrepreneurship as a Good Career Choice	24.27	92.45	62.41
INDPERINT	Percentage of Individuals using the Internet	0.53	98.45	48.97
PATENTS	Patents Office & Patents Families - Triadic Patent Families	-	107.60	2,788.51
GETELEC	Getting electricity - Score	54.95	99.85	79.98
EASEBUS	Ease of doing business score	55.82	88.25	72.09
MINRTS	Score-Protecting minority investors	44.44	91.87	66.72
MINSUITIND	Ease of shareholder suits index	1.33	9.00	6.43
MINSUITSCR	Score-Ease of shareholder suits index (0-10)	13.33	90.00	64.32
MINEXTDIRIN	Extent of director liability index (0-10)	1.00	9.00	5.54
MINEXTDIRSCR	Extent of director liability index (0-10) score	10.00	90.00	55.38
STARTBUSSCR	Starting a business score	64.27	97.22	82.92
FXTELSUB	Fixed telephone subscriptions (per 100 people)	3.12	63.24	36.28
TRDPERGDP	Trade (% of GDP)	18.35	30.39	71.48
FMPAR	% of seats held by women in national parliaments	-	44.50	19.48
STRGLEG	Strength of legal rights index (0=weak to 12=strong)	2.00	1.00	5.57

Appendix 4 – R Code

Below is the R code used in the development of this paper.

```
################################################################################
#Administration
#File Name: Predicting R&D.R
#Theme: Predict a country's R&D (as percent of GDP) using various predictors
#Date: 05/06/2022
#Version: 001.000
#Author: Glen Cooper
################################################################################
setwd("C:/Users/glenc/downloads")
rm(list = ls())   # Clear environment
oldpar <- par()   # save default graphical parameters
if (!is.null(dev.list()["RStudioGD"]))
  dev.off(dev.list()["RStudioGD"])   # Clear plot window
options(warn = 0) #Enable global warnings.  Note to disable use: options(warn = -1)
cat("\014")  # Clear the Console

#NAMING CONVENTIONS ###############################################################
#Dataframe & matrixes names: Begins with capital letter and separated by "_"  e.g., Data_name
#Models: Begins with capital letter and separated by "." e.g., Model.name
#Values & vectors: Begin with all capital letters and separated by "_" e.g.,  VARIABLE_name
#Functions: Begins with lowercase letter and separated by "." e.g., function.name
# Major blocks identified with three line block heading, minor blocks with one line heading
# After Major and Minor blocks the following subblock conventions used:
# "##Major subblock heading"
# "#Minor subsubblock heading"
#Ctrl+Shift+A used to reformat highlighted code
################################################################################

################################################################################
#                    Packages                                                  #
#                      and                                                     #
#                    Libraries                                                 #
################################################################################

library(readxl) # Read in Excel files
library(dplyr)  # Review data
library(stats)  # Regressions functions
library(ISLR2)  # Regression functions
library(leaps)  # Regression functions
library(glmnet) # Regression functions
library(pls)    # Regression functions
```

```
#############################################################################
#                        Load Data                                          #
#                          and                                              #
#                  Review / Clean / Normalize                               #
#############################################################################

##Load data, view, and make available
#Load data
RND <-
  read_excel(
    "C:/Users/glenc/Downloads/EPPS6323_Data4R.xlsx",
    col_names = TRUE,
    na = "",
    trim_ws = TRUE
  ) #Read the Excel R&D response & predictors file and load into matrix
#Review view data
View(RND) #View the matrix
#Determine if matrix has any na values
countna <-
  function(x) {
    sum(is.na(x))
  } #Define count the number of nas function
sapply(RND, countna) #Count number of nas within data
#RND <- na.omit(RND) #Use to remove nas within data if deemed necessary
#Make available
attach(RND) #Make all variables in dataset available
#Convert data to data frame and variables from character to factors
RND_df <-
  as.data.frame(RND) #Convert to dataframe
dim(RND)
summary(RND_df) #Summarize data components

#Normalize Features
#Temp <- scale(RND_df[,-1]) #Normalization not considered necessary as all variables are
      normalized
#RND_scale_df <- as.data.frame(cbind(RND$RDPERGDP, Temp))
#remove(Temp)


#############################################################################
#                    Parametric Linear Models                               #
#             Stepwise (forward/backward) & Best Subset Selection           #
#                          Regression                                       #
#############################################################################


#############################################################################
#                Linear Regression Forward Stepwise                           #
#############################################################################

##Drop predictors that are linear dependent (LD)
```

```
RND_df_02 <-
  RND_df[, -29] #Drop MINEXTDIRIN which in LD on MINEXTDIRSCR
RND_df_02 <-
  RND_df_02[, -27] #Drop MINSUITIND which in LD on MINSUITSCR


##Run Model and Create Summary
#Run model
RND.reg.step.frwd <-
  regsubsets(RDPERGDP ~ ., RND_df_02, method = "forward")
#Create summary of regression output features
reg.summary <- summary(RND.reg.step.frwd)

##Review Model Output
#Plot relative value of each predictor based on Adjusted R sqr and Cp
plot(RND.reg.step.frwd, scale = "adjr2") #Larger Adj R sqr better
plot(RND.reg.step.frwd, scale = "Cp") #Lower Cp value better
#Identify impact of number of predictors to performance measures
#  for Adjusted RSq
plot(reg.summary$adjr2,
     xlab = "Number of Variables",
     ylab = "Adjusted RSq",
     type = "l")
points(
  which.max(reg.summary$adjr2),
  reg.summary$adjr2[which.max(reg.summary$adjr2)],
  col = "red",
  cex = 2,
  pch = 20
)
#  for CP
plot(reg.summary$cp,
     xlab = "Number of Variables",
     ylab = "Cp",
     type = "l")
points(
  which.min(reg.summary$cp),
  reg.summary$cp[which.min(reg.summary$cp)],
  col = "red",
  cex = 2,
  pch = 20
)
#Model output
reg.summary # Predictors included in each model indicated with "*"
coef(RND.reg.step.frwd, 8) #Display predictor coefficients
reg.summary$adjr2 #Highest adj r sq best
reg.summary$cp #Least cp value best
reg.summary$rsq #Highest r sq best
LR_frwrd_Rsq <- max(reg.summary$rsq)
LR_frwrd_Rsq
```

```
##############################################################################
#                Linear Regression Backward Stepwise                         #
##############################################################################

##Drop predictors that are linear dependent (LD)
RND_df_02 <-
  RND_df[, -29] #Drop MINEXTDIRIN which in LD on MINEXTDIRSCR
RND_df_02 <-
  RND_df_02[, -27] #Drop MINSUITIND which in LD on MINSUITSCR

##Run Model and Create Summary
#Run model
RND.reg.step.bckwd <-
  regsubsets(RDPERGDP ~ ., RND_df_02, method = "backward")
#Create summary of regression output features
reg.summary <- summary(RND.reg.step.bckwd)

##Review Model Output
#Plot relative value of each predictor based on Adjusted R sqr and Cp
plot(RND.reg.step.bckwd, scale = "adjr2") #Larger Adj R sqr better
plot(RND.reg.step.bckwd, scale = "Cp") #Lower Cp value better
#Identify impact of number of predictors to performance measures
#  for Adjusted RSq
plot(reg.summary$adjr2,
     xlab = "Number of Variables",
     ylab = "Adjusted RSq",
     type = "l")
points(
  which.max(reg.summary$adjr2),
  reg.summary$adjr2[which.max(reg.summary$adjr2)],
  col = "red",
  cex = 2,
  pch = 20
)
#  for CP
plot(reg.summary$cp,
     xlab = "Number of Variables",
     ylab = "Cp",
     type = "l")
points(
  which.min(reg.summary$cp),
  reg.summary$cp[which.min(reg.summary$cp)],
  col = "red",
  cex = 2,
  pch = 20
)
#Model output
reg.summary # Predictors included in each model indicated with "*"
coef(RND.reg.step.bckwd, 8) #Display predictor coefficients
```

```
reg.summary$adjr2 #Highest adj r sq best
reg.summary$cp #Least cp value best
reg.summary$rsq #Highest r sq best
LR_Bkwrd_Rsq <- max(reg.summary$rsq)
LR_Bkwrd_Rsq

#############################################################################################
#             Linear Regression Best Subset Selection                                       #
#############################################################################################

##Drop predictors that are linear dependent (LD)
RND_df_02 <-
  RND_df[, -29] #Drop MINEXTDIRIN which in LD on MINEXTDIRSCR
RND_df_02 <-
  RND_df_02[, -27] #Drop MINSUITIND which in LD on MINSUITSCR

##Run Model and Create Summary
#Run model
RND.reg.step.exhast <-
  regsubsets(RDPERGDP ~ ., RND_df_02, method = "exhaustive")
#Create summary of regression output features
reg.summary <- summary(RND.reg.step.exhast)


##Review Model Output
#Plot relative value of each predictor based on Adjusted R sqr and Cp
plot(RND.reg.step.exhast, scale = "adjr2") #Larger Adj R sqr better
plot(RND.reg.step.exhast, scale = "Cp") #Lower Cp value better
#Identify impact of number of predictors to performance measures
#  for Adjusted RSq
plot(reg.summary$adjr2,
     xlab = "Number of Variables",
     ylab = "Adjusted RSq",
     type = "l")
points(
  which.max(reg.summary$adjr2),
  reg.summary$adjr2[which.max(reg.summary$adjr2)],
  col = "red",
  cex = 2,
  pch = 20
)
#  for CP
plot(reg.summary$cp,
     xlab = "Number of Variables",
     ylab = "Cp",
     type = "l")
points(
  which.min(reg.summary$cp),
  reg.summary$cp[which.min(reg.summary$cp)],
  col = "red",
  cex = 2,
```

```
  pch = 20
)
#Model output
reg.summary # Predictors included in each model indicated with "*"
coef(RND.reg.step.exhast, 8) #Display predictor coefficients
reg.summary$adjr2 #Highest adj r sq best
reg.summary$cp #Least cp value best
reg.summary$rsq #Highest r sq best
LR_BestSub_Rsq <- max(reg.summary$rsq)
LR_BestSub_Rsq

##############################################################################
#                    Shrinkage Models                                        #
#                     Ridge & LASSO                                          #
#                      Regression                                            #
##############################################################################

##############################################################################
#                    Ridge Regression                                        #
##############################################################################

##Drop predictors that are linear dependent (LD)
#Note in this Ridge Regression leaving the linear combinations did not
#cause the analysis from running but did worsen the MSE
RND_df_02 <-
  RND_df[, -29] #Drop MINEXTDIRIN which in LD on MINEXTDIRSCR
RND_df_02 <-
  RND_df_02[, -27] #Drop MINSUITIND which in LD on MINSUITSCR

##Split Data into Train and Test
set.seed(1)
#In subset function sample, the first parameter nrow(df) is the number of elements
#to choose from and second parameter nrow(df) is the number of items to choose
subset <-
  sample(nrow(RND_df_02), nrow(RND_df_02) * 0.5) #Splits data 50% / 50%
Train <-
  RND_df_02[subset, ] #Assigns to train all the selected sample rows from vector subset
Test <-
  RND_df_02[-subset, ] #Assigns to test all the rows (negative) not in vector subset
#Compare number of rows in train and test dataframes to verify split
nrow(Test) / nrow(Train) #Calculate percent test vs train


##Fit ridge regression model, find best lambda, cross validate, and show output
#Model Creation
set.seed(1)
RRTrain <-
  (model.matrix(Train$RDPERGDP ~ ., data = Train)) #Creates a model matrix using train data
RRTest <-
  (model.matrix(Test$RDPERGDP ~ ., data = Test)) #Creates a model matrix using test data
Grid <-
```

```
  10 ^ seq(10, -2, length = 100) #Creates Gride for lambda evaluation in glmnet function below
#glmnet function RRTrain is x matrix, Train$RDPERGDP is y response variable
#and alpha=0 defines Ridge regression model, thresh here and in cv.glment
#function brings penalty paths to equal (or at least closer)
#glmnet function default is to standardizes all variables
RRModel <-
  glmnet(
    RRTrain,
    Train$RDPERGDP,
    alpha = 0,
    lambda = Grid,
    thresh = 1e-20
  )
#Cross validate model
RRModelcv <-
  cv.glmnet(
    RRTrain,
    Train$RDPERGDP,
    alpha = 0,
    lambda = Grid,
    thresh = 1e-20
  ) #cv.glmnet does k-fold (default is 10) cross-validation for glmnet,
#Model output best lambda
RRbestlam <-
  RRModelcv$lambda.min #Get minimum lambda value from RRModelcv
RRbestlam #Display raw best lambda
log(RRbestlam) #Log of best lambda
#Use best lamda from cross validation
#In function predict RRModel is the model, newx is the new x matrix for
#prediction calculation and s is the lambda to be used
RRpred <- predict(RRModel, s = RRbestlam, newx = RRTest)
#Test error
RRtestMSE <- mean((Test$RDPERGDP - RRpred) ^ 2) #Calculate test MSE
RRtestMSE #Display MSE
#Display Model coefficients
RRpredcoeff <- predict(RRModel, s = RRbestlam, type = "coefficients")
RRpredcoeff
#Calculate Final R Squared on last test MSE
AVGRDOERGDPtest <- mean(Test$RDPERGDP)
AVGtestMSE <- mean((AVGRDOERGDPtest - Test$RDPERGDP) ^ 2)
RR_Rsq <- 1 - (RRtestMSE / AVGtestMSE)
RR_Rsq


#############################################################################
#                       Lasso Regression                                    #
#############################################################################

##Drop predictors that are linear dependent (LD)
#Note in this Lasso Regression leaving the linear combinations did not
#cause the analysis from running nor did it impact the MSE
```

```
RND_df_02 <-
  RND_df[, -29] #Drop MINEXTDIRIN which in LD on MINEXTDIRSCR
RND_df_02 <-
  RND_df_02[, -27] #Drop MINSUITIND which in LD on MINSUITSCR

##Split Data into Train and Test
set.seed(1)
#In subset function sample, the first parameter nrow(df) is the number of elements
#to choose from and second parameter nrow(df) is the number of items to choose
subset <-
  sample(nrow(RND_df_02), nrow(RND_df_02) * 0.5) #Splits data 50% / 50%
Train <-
  RND_df_02[subset, ] #Assigns to train all the selected sample rows from vector subset
Test <-
  RND_df_02[-subset, ] #Assigns to test all the rows (negative) not in vector subset
#Compare number of rows in train and test dataframes to verify split
nrow(Test) / nrow(Train) #Calculate percent test vs train

##Model
set.seed(1)
LASSOTrain <- (model.matrix(Train$RDPERGDP ~ ., data = Train))
LASSOTest <- (model.matrix(Test$RDPERGDP ~ ., data = Test))
grid <-
  10 ^ seq(10, -2, length = 100) #Creates Gride for lambda evaluation in glmnet function below
#glmnet function LASSOTrain is x matrix, train$Apps is y response
#and alpha=1 defines lasso model, thresh here and in cv.glment
#function brings penalty paths to equal (or at least closer)
#glmnet function default is to standardizes all variables
LASSOModel <-
  glmnet(
    LASSOTrain,
    Train$RDPERGDP,
    alpha = 1,
    lambda = grid,
    thresh = 1e-20
  )
#Cross validate model
#cv.glmnet does k-fold (default is 10) cross-validation for glmnet,
LASSOModelcv <-
  cv.glmnet(
    LASSOTrain,
    Train$RDPERGDP,
    alpha = 0,
    lambda = grid,
    thresh = 1e-20
  )
##Model output best lambda
LASSObestlam <- LASSOModelcv$lambda.min
LASSObestlam #Raw best lambda
log(LASSObestlam) #Log of best lambda
#Use best lambda from cross validation
```

```
#In function LASSOModel is the model, newx is the new x matrix for
#prediction calculation and s is the lambda to be used
LASSOpred <- predict(LASSOModel, s = LASSObestlam,
                     newx = LASSOTest)
#Test error
LASSOtestMSE <-
  mean((Test$RDPERGDP - LASSOpred) ^ 2) #Calculate test MSE
LASSOtestMSE
#Display nonzero coefficient estimates
LASSOnozerocoeff <-
  predict(LASSOModel, s = LASSObestlam, type = "coefficients")
LASSOnozerocoeff
#Calculate Final R Squared on last test MSE
AVGRDOERGDPtest <- mean(Test$RDPERGDP)
AVGtestMSE <- mean((AVGRDOERGDPtest - Test$RDPERGDP) ^ 2)
LASSO_Rsq <- 1 - (LASSOtestMSE / AVGtestMSE)
LASSO_Rsq


################################################################################
#                  Dimension Reduction Models                                  #
#              Principal Components Analysis (PCA) &                           #
#                  Partial Least Squares (PLS)                                 #
################################################################################

################################################################################
#                 Principal Components Analysis (PCA)                          #
################################################################################

##Drop predictors that are linear dependent (LD)
#In PCA it is not necessary to drop the two linear components:
#MINEXTDIRIN which in LD on MINEXTDIRSCR and
#MINSUITIND which in LD on MINSUITSCR because the
#model will account for the exclusion of those components by showing a worsening
#MSE when they are included
#Therefore, using PCA the entire RND_df can be used

##Split Data into Train and Test
set.seed(1)
#In subset function sample, the first parameter nrow(df) is the number of elements
#to choose from and second parameter nrow(df) is the number of items to choose
subset <- sample(nrow(RND_df), nrow(RND_df) * 0.5) #Splits data 50% / 50%
Train <-
  RND_df[subset, ] #Assigns to train all the selected sample rows from vector subset
Test <-
  RND_df[-subset, ] #Assigns to test all the rows (negative) not in vector subset
#Compare number of rows in train and test dataframes to verify split
nrow(Test) / nrow(Train) #Calculate percent test vs train


##Model
```

```
#In pcr function, RDPERGDP is predicted variable, data is the predictor
#dataframe, scale=TRUE (is default) and scales the predictor variables to
#a unit variance, validation="CV" performs a 10 fold cross-validation
PCRmodel <- pcr(RDPERGDP ~ .,
                data = Train,
                scale = TRUE,
                validation = "CV")

##Model output
summary(PCRmodel) #Provides cross-validation values for each possible number of components
#Plot
#In validationplot function plots Mean square error (MSE) for each possible number of components
validationplot(PCRmodel, val.type = "MSEP")

##MSR by Number of Components
PCRpredictnc <- predict(PCRmodel, Test, ncomp = 6)
PCRtestncMSE <- mean((Test$RDPERGDP - PCRpredictnc) ^ 2)
PCRtestncMSE #MSE with 6 components
PCRpredictnc <- predict(PCRmodel, Test, ncomp = 29)
PCRtestncMSE <- mean((Test$RDPERGDP - PCRpredictnc) ^ 2)
PCRtestncMSE #MSE with 29 components
PCRpredictnc <- predict(PCRmodel, Test, ncomp = 32)
PCRtestncMSE <- mean((Test$RDPERGDP - PCRpredictnc) ^ 2)
PCRtestncMSE #MSE with 32 components
#Calculate Final R Squared on last test MSE
AVGRDOERGDPtest <- mean(Test$RDPERGDP)
AVGtestMSE <- mean((AVGRDOERGDPtest - Test$RDPERGDP) ^ 2)
PCRnc32_Rsq <- 1 - (PCRtestncMSE / AVGtestMSE)
PCRnc32_Rsq


#############################################################################
#                 Partial Least Squares (PLS)                               #
#############################################################################

##Drop predictors that are linear dependent (LD)
#In PLS it is not necessary to drop the two linear components:
#MINEXTDIRIN which in LD on MINEXTDIRSCR and
#MINSUITIND which in LD on MINSUITSCR because the
#model will account for the exclusion of those components by showing a worsening
#MSE when they are included
#Therefore, using PLS the entire RND_df can be used

##Split Data into Train and Test
set.seed(1)
#In subset function sample, the first parameter nrow(df) is the number of elements
#to choose from and second parameter nrow(df) is the number of items to choose
subset <- sample(nrow(RND_df), nrow(RND_df) * 0.5) #Splits data 50% / 50%
Train <-
  RND_df[subset, ] #Assigns to train all the selected sample rows from vector subset
Test <-
```

```
  RND_df[-subset, ] #Assigns to test all the rows (negative) not in vector subset
#Compare number of rows in train and test dataframes to verify split
nrow(Test) / nrow(Train) #Calculate percent test vs train

##Model
#In plsr function, RDPERGDP is predicted variable, data is the predictor
#dataframe, scale=TRUE (is default) and scales the predictor variables to
#a unit variance, validation="CV" performs a 10 fold cross-validation
PLSmodel <- plsr(RDPERGDP ~ .,
                 data = Train,
                 scale = TRUE,
                 validation = "CV")

##Model output
summary(PLSmodel) #Provides cross-validation values for each possible number of components
#Plot
#Function vialidationplot plots Mean square error (MSEP) for each possible number of components
validationplot(PLSmodel, val.type = "MSEP")
#MSR by Number of Components
PLSpredictnc <- predict(PLSmodel, Test, ncomp = 5)
PLStestncMSE <- mean((Test$RDPERGDP - PLSpredictnc) ^ 2)
PLStestncMSE #MSR with 5 components
PLSpredictnc <- predict(PLSmodel, Test, ncomp = 10)
PLStestncMSE <-
  mean((Test$RDPERGDP - PLSpredictnc) ^ 2)
PLStestncMSE #MSR with 10 components
PLSpredictnc <- predict(PLSmodel, Test, ncomp = 30)
PLStestncMSE <-
  mean((Test$RDPERGDP - PLSpredictnc) ^ 2)
PLStestncMSE #MSR with 30 components
#Calculate Final R Squared on last test MSE
AVGRDOERGDPtest <- mean(Test$RDPERGDP)
AVGtestMSE <- mean((AVGRDOERGDPtest - Test$RDPERGDP) ^ 2)
PLSnc30_Rsq <- 1 - (PLStestncMSE / AVGtestMSE)
PLSnc30_Rsq


#############################################################################
#                    Compare Models                                         #
#                     Using R Square                                        #
#                       Evaluation                                          #
#############################################################################


#Plot & List R2 by model
Allnames <- c("LRF", "LRB", "LRB", "Ridge", "Lasso", "PCR", "PLS")
AllR2 <-
  c(
    LR_frwrd_Rsq,
    LR_Bkwrd_Rsq,
    LR_BestSub_Rsq,
```

```
    RR_Rsq,
    LASSO_Rsq,
    PCRnc32_Rsq,
    PLSnc30_Rsq
  )
barplot(
  AllR2,
  xlab = "Models",
  ylab = "R2",
  ylim = c(0, 1),
  names = Allnames
)
Allnames
AllR2
```