Brendon Hahm

Professor Xuan

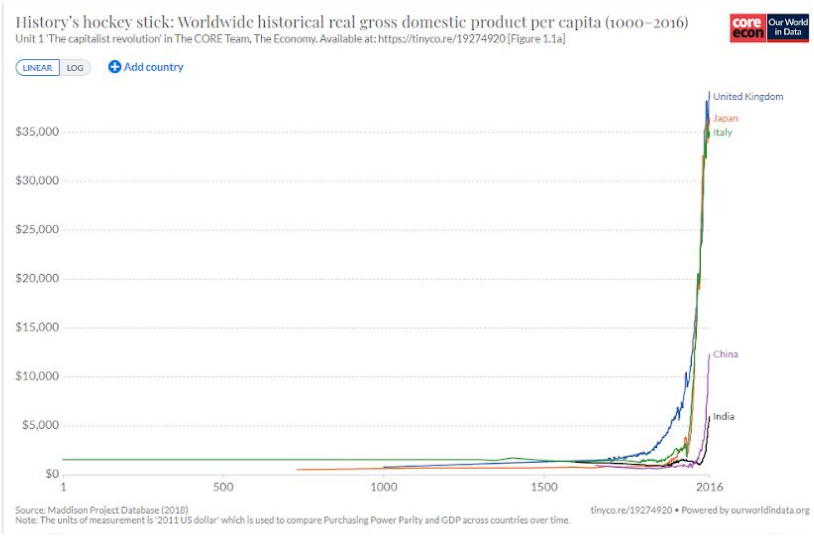
CSS 100

18 December 2020

TFP Regression Modeling and Macroeconomic Data Analysis of the Solow-Romer Model

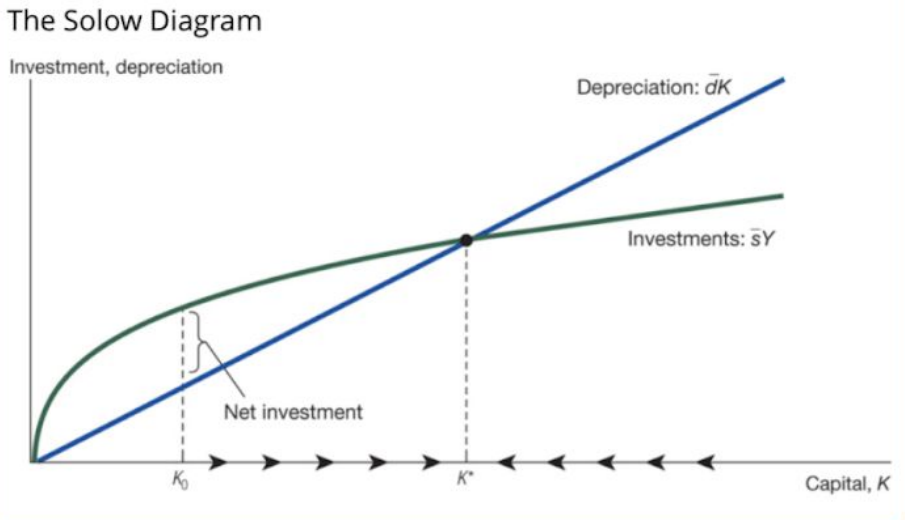
Introduction to the Literature of Endogenous Economic Growth

For the vast majority of human history, our civilizations and societies have experienced little to no economic growth. This observation has famously been called "history's hockey stick" due to the shape of GDP per capita plots on time.



It wasn't until the Second Industrial Revolution where countries like the United Kingdom, Germany, and the United States industrialized and began to see exponential growth in per capita GDP. This exponential growth has continued for many countries, especially in the western world, where for other countries growth had initial increases and then stagnated. This event is commonly referred to as the Great Divergence and it showed the per capita GDP growth is not a mere inevitability. It is both of these economic phenomena that puzzled economists. Robert Solow developed the first theory of economic growth that understood growth in per capita GDP as being exogenous.

In the Solow model, given fixed exogenous variables, we can only understand per capita GDP as moving towards a steady state or equilibrium determined by savings and depreciation of capital.



steady state equation for per capita GDP:
$y_{ss} = A^{3/2}(s/d)^{1/2}$
where s = savings rate and
d = depreciation rate

Finally economists had some way to explain per capita GDP growth. However, the exogenous nature of per capita GDP growth in the Solow model only allowed us to see the effects of savings and depreciation rates of capital on per capita GDP. Given all of these variables are exogenous, we can't properly model or rigorously understand the growth rate of per capita GDP.

Paul Romer introduced a model that would finally give us a way to endogenize TFP that would allow us to discuss per capita GDP growth within the model. Let's take a Cobbs-Douglas production function with the form:

$$Y = AK^{\alpha}L^{1-\alpha}$$

where Y = total output, A = Total Factor Productivity, K = capital, L = Labor, α = output elasticity of capital. If you take α to be equal to .5 you can derive a new equation of:

$$y = Ak^{1/2}$$

where y = output per capita, A = Total Factor Productivity, and k = capital per capita. So we can see that GDP per capita is determined by TFP and capital per capita. In the Solow model, capital

is endogenous and TFP exogenous leading to the aforementioned result of GDP per capita steady states. In the Romer model, TFP is endogenized and can be intuitively understood as access to the global stock of ideas. The endogeneity of TFP allows us to understand its growth rate and in the model, we can use the following equation to measure the growth rate of TFP. Some particularly useful insights from the Romer model can be algebraically expressed as the following:

(a) $g_A \stackrel{\text{def}}{=} \Delta A_{t+1}/A_t = zlL$

(b) $y_t = A_o(1-l)(1+g_A)^t$

(c) $(y_{t+1} - y_t)/y_t = g_A = zlL$

(d) $g_{yt} = g_{A_t} + (1/3)g_{k_t} + (2/3)g_{l_{yt}}$

where g_x = growth rate of x, z = productivity of research, and l = share of labor in research. (a) gives us a way to measure the growth rate of TFP, (b) tells us that per capita GDP can be determined by a geometric series of a constant growth rate, (c) tells us that the percent change of per capita GDP is equal to the growth rate of TFP, and finally that the growth rate of GDP per capita is equal to a weighted sum of the growth rates of the variables in the production function. We can see that under the Romer model, we now have an endogenous model of economic growth that can help explain the sustained exponential growth seen in the hockey stick visualization.

With the Solow-Romer model, we have a very powerful theoretical tool to understand sustained macroeconomic growth. However, how well does our macroeconomic data support the model? For most, it would seem counterintuitive to think that most of the sustained macroeconomic growth is a product of idea creation. The first objective of this project will be to explore how well the data fits our relatively recent understanding of macroeconomic growth. The

second objective will be to see how strong of a predictive model we can build to determine TFP. TFP is a latent variable that is unobservable. Recall that the production function is:

$$Y = AK^{\alpha}L^{1-\alpha}$$

TFP can, and is, calculated as:

$$A = Y/(K^{\alpha}L^{1-\alpha})$$

While this formulation for calculating TFP may be necessarily true as defined by the model, this project will explore other macroeconomic parameters that ought to explain TFP via predictive models outside of the parameters of the neoclassical production function.

Data

All data was compiled from the Federal Reserve Economic Data (FRED). The project uses time series data of macroeconomic features from the years 1960 - 2017. The features used are:

"Capital Stock at Constant National Prices for China", "Capital Stock at Constant National Prices for United States", "Civilian Labor Force Level", "Current-Cost Depreciation of Fixed Assets", "Current-Cost Depreciation of Fixed Assets: Private: Intellectual property products", "Current-Cost Depreciation of Private and government fixed assets: Nonresidential: Intellectual property products", Government Gross Investment: Federal: Gross Investment: Intellectual Property Products: Research and Development", "Gross Domestic Product", "Gross Domestic Product Excluding Research", "Gross Domestic Product for China", "Gross Domestic Product: Research and Development", "Gross Private Domestic Investment: Fixed Investment: Nonresidential: Intellectual Property Products: Research and Development", "Private Fixed Investment in Intellectual Property Products: Research and development: Business:

Manufacturing: Pharmaceutical and medicine manufacturing", "Total Factor Productivity at Constant National Prices for China", "Total Factor Productivity at Constant National Prices for United States".

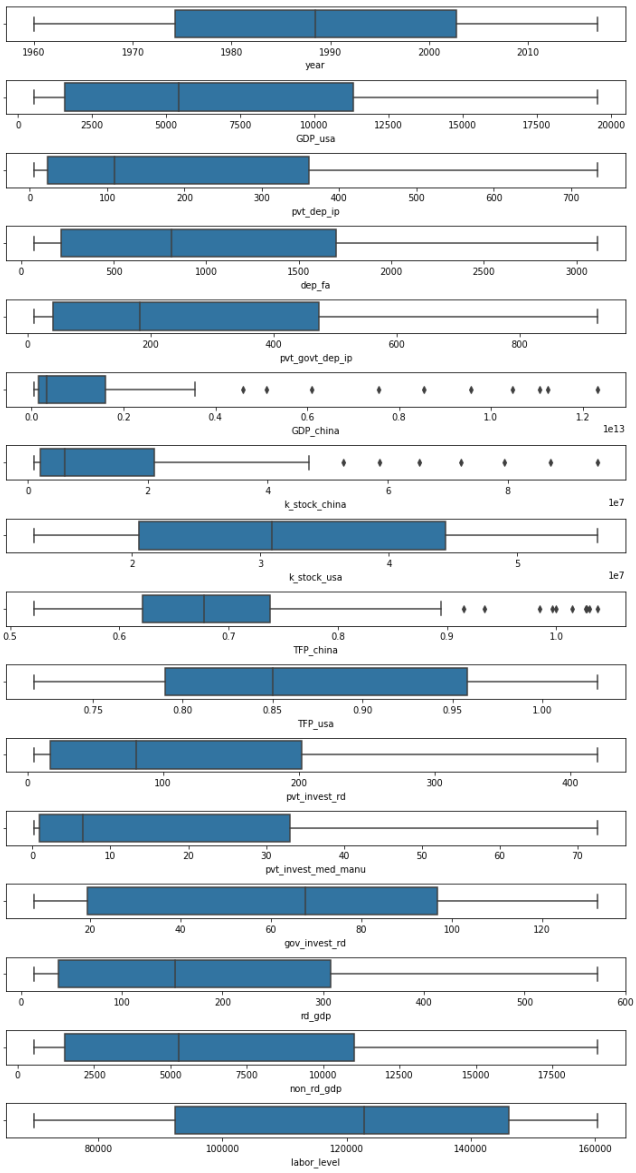
All values were measured at nominal price levels and recorded at annual frequencies except for the labor force levels. The features can be categorized into the following groups for analysis:

Analysis Group 1: year, GDP_usa, TFP_usa, k_stock_usa

Analysis Group 2: year, GDP_china, TFP_china, k_stock_china

Analysis Group 3: year, TFP_usa, rd_gdp, non_rd_gdp, dep_fa, pvt_invest_med_manu, gov_invest_rd, labor_level, pvt_invest_rd, pvt_dep_ip, pvt_govt_dep_ip

Distribution of Features

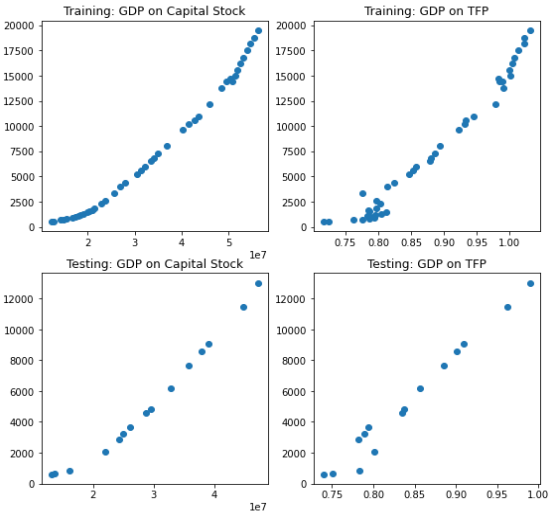


(Although there are outliers in the data, it is acceptable given the relationship of growth rate of TFP with growth rate of GDP. This relationship with the phenomena of History's Hockey Stick justifies the inclusion of these outliers.)

Preprocessing of Data

The various times series data on each feature spanned different years but they all covered the time period from 1960 - 2017. With that in mind, I transformed labor force levels from a monthly frequency to annual by averaging the values for each year. Dropping all NaN values led to the range of years from 1960 - 2017.

Analysis of Groups 1 and 2:



First we want to create a linear regression to model the predictive ability of TFP_usa and k_stock_usa on GDP_usa. To do so, we used sklearn's LinearRegression model with a simple holdout method for cross validation. Given the nature of time series data, we also ran the a regression model without cross validation using the statsmodels module. For the sklearn regression with holdout, we were able to achieve R^2 values of .978 for training set, .956 for testing set, and .987 for adjusted statsmodels.

OLS Regression Results

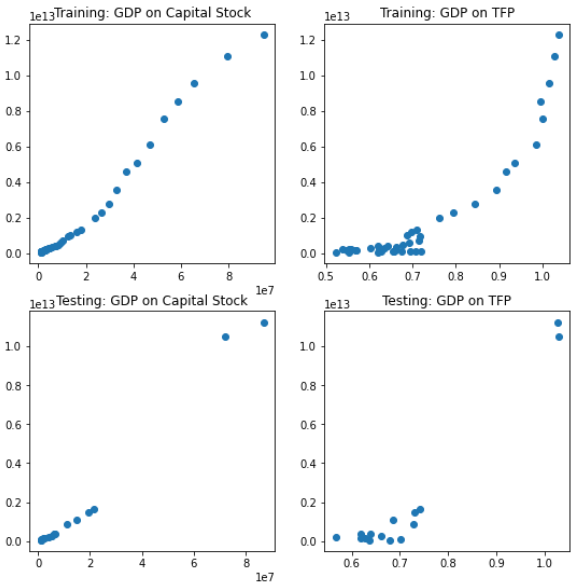
Dep. Variable:	GDP_usa	R-squared (uncentered):	0.988
Model:	OLS	Adj. R-squared (uncentered):	0.987
Method:	Least Squares	F-statistic:	2227.
Date:	Fri, 18 Dec 2020	Prob (F-statistic):	4.27e-54
Time:	08:50:48	Log-Likelihood:	-483.75
No. Observations:	58	AIC:	971.5
Df Residuals:	56	BIC:	975.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
TFP_usa	-1.019e+04	535.894	-19.018	0.000	-1.13e+04	-9117.923
k_stock_usa	0.0005	1.33e-05	36.718	0.000	0.000	0.001

Omnibus:	7.085	Durbin-Watson:	0.046
Prob(Omnibus):	0.029	Jarque-Bera (JB):	6.996
Skew:	0.800	Prob(JB):	0.0303
Kurtosis:	2.419	Cond. No.	1.39e+08

As expected from the model, extremely strong R^2 values with both coefficients being statistically significant. However, the multicollinearity of k-stock and TFP indicates that the coefficients will be inaccurate but the model's predictability can still be inferred. When regressed on either TFP_usa or k_stock_usa individually, the coefficients of each for their own univariate regression models are 8643.7116 for TFP_usa and .0002 for k_stock_usa. So we can see that these two features produce strong predictive power in predicting GDP for the United States.

To check if this data is consistent with production functions in general, and not just in the United States, I also ran a linear regression with the equivalent features for China. For the



sklearn regression with holdout, we were able to achieve R^2 values of .981 for training set, .982 for testing set, and .986 for adjusted statsmodels.

As we can see, there are similar results here with the predictive model for the production function of the United States. Given the same issues with multicollinearity here, the coefficients under separate univariate regressions are 3.473e+12 for TFP_china and and 1.298e+05 for k_stock_china.

OLS Regression Results

Dep. Variable:	GDP_china	R-squared (uncentered):	0.986
Model:	OLS	Adj. R-squared (uncentered):	0.986
Method:	Least Squares	F-statistic:	2000.
Date:	Fri, 18 Dec 2020	Prob (F-statistic):	8.37e-53
Time:	08:51:18	Log-Likelihood:	-1638.9
No. Observations:	58	AIC:	3282.
Df Residuals:	56	BIC:	3286.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
TFP_china	-6.783e+11	1.21e+11	-5.624	0.000	-9.2e+11	-4.37e+11
k_stock_china	1.421e+05	3006.435	47.270	0.000	1.36e+05	1.48e+05

Omnibus:	0.764	Durbin-Watson:	0.181
Prob(Omnibus):	0.682	Jarque-Bera (JB):	0.710
Skew:	-0.257	Prob(JB):	0.701
Kurtosis:	2.828	Cond. No.	5.85e+07

Analysis of Group 3

Given the ability of our two models for predicting GDP in both the United States and China being strongly predictive, it becomes a question of how well do observable economic parameters relating to the concept of TFP predict TFP? For reference, the Solow-Romer model tells us that productivity of research, labor levels, and research are determinants of growth in TFP. Given that TFP is a latent variable that is determined by $A = Y/(K^{\alpha}L^{1-\alpha})$, let us see how well other macroeconomic parameters can predict TFP levels using time series data of the United States.

First, I created a multivariate regression of all the features in the Analysis 3 grouping. Using a simple holdout method for cross validation with sklearn's LinearRegression model and statsmodels's model without cross validation, I was able to achieve R^2 values of .991 for training set, .972 for testing set, and .998 for adjusted statsmodels.

OLS Regression Results

Dep. Variable:	TFP_usa	R-squared (uncentered):	0.999
Model:	OLS	Adj. R-squared (uncentered):	0.998
Method:	Least Squares	F-statistic:	3805.
Date:	Fri, 18 Dec 2020	Prob (F-statistic):	1.67e-66
Time:	08:51:48	Log-Likelihood:	115.51
No. Observations:	58	AIC:	-213.0
Df Residuals:	49	BIC:	-194.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
rd_gdp	0.0175	0.032	0.541	0.591	-0.048	0.083
non_rd_gdp	-4.897e-05	2.37e-05	-2.068	0.044	-9.65e-05	-1.39e-06
dep_fa	-0.0011	0.000	-4.604	0.000	-0.002	-0.001
pvt_invest_med_manu	0.0105	0.004	2.951	0.005	0.003	0.018
gov_invest_rd	-0.0161	0.033	-0.492	0.625	-0.082	0.050
labor_level	1.073e-05	1.84e-07	58.306	0.000	1.04e-05	1.11e-05
pvt_invest_rd	-0.0171	0.033	-0.526	0.601	-0.082	0.048
pvt_dep_ip	0.0008	0.001	0.592	0.556	-0.002	0.004
pvt_govt_dep_ip	0.0018	0.001	1.248	0.218	-0.001	0.005

Omnibus:	0.055	Durbin-Watson:	0.515
Prob(Omnibus):	0.973	Jarque-Bera (JB):	0.059
Skew:	-0.048	Prob(JB):	0.971
Kurtosis:	2.875	Cond. No.	1.47e+06

As we can see, we have another strong model that can almost completely explain the variance of TFP. Again we are faced with the issue of multicollinearity. To reduce unnecessary multicollinearity, I removed features that would clearly have strict linear relationships with each other. This left me with rd_gdp, pvt_invest_rd, pvt_govt_dep_ip. These three features left me with a model with R^2 values of .955 for training set, .956 for testing set, and .866 for adjusted statsmodels. This model gives us a low enough condition number to minimize multicollinearity. I would also like to note that in this model, pvt_invest_rd has a p-value that indicates it is not statistically significant at $\alpha = .05$.

OLS Regression Results

Dep. Variable:	TFP_usa	R-squared (uncentered):	0.873
Model:	OLS	Adj. R-squared (uncentered):	0.866
Method:	Least Squares	F-statistic:	126.3
Date:	Fri, 18 Dec 2020	Prob (F-statistic):	1.21e-24
Time:	09:03:43	Log-Likelihood:	-14.569
No. Observations:	58	AIC:	35.14
Df Residuals:	55	BIC:	41.32
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
rd_gdp	0.0234	0.002	10.633	0.000	0.019	0.028
pvt_invest_rd	-0.0114	0.007	-1.729	0.089	-0.025	0.002
pvt_govt_dep_ip	-0.0084	0.003	-2.794	0.007	-0.014	-0.002

Omnibus:	22.536	Durbin-Watson:	0.046
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4.161
Skew:	0.085	Prob(JB):	0.125
Kurtosis:	1.699	Cond. No.	84.9

Remember that in the Solow-Romer model, we can understand percent change of GDP and the growth rate of TFP as: $(y_{t+1} - y_t)/y_t = g_A = zlL$. The final model I will build for this project is to observe the features that closely relate to the right hand side of this equation, namely: rd_gdp and labor_level. Creating a multivariate regression with these features created a model where R^2 values of .963 for training set, .957 for testing set, and .990 for adjusted statsmodels. In this model, both coefficients are statistically significant at $\alpha = .05$.

OLS Regression Results

Dep. Variable:	TFP_usa	R-squared (uncentered):	0.990
Model:	OLS	Adj. R-squared (uncentered):	0.990
Method:	Least Squares	F-statistic:	2860.
Date:	Fri, 18 Dec 2020	Prob (F-statistic):	4.19e-57
Time:	09:02:12	Log-Likelihood:	59.987
No. Observations:	58	AIC:	-116.0
Df Residuals:	56	BIC:	-111.9
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
rd_gdp	-0.0008	9.44e-05	-8.198	0.000	-0.001	-0.001
labor_level	8.496e-06	1.98e-07	43.005	0.000	8.1e-06	8.89e-06

Omnibus:	22.795	Durbin-Watson:	0.024
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5.990
Skew:	0.459	Prob(JB):	0.0500
Kurtosis:	1.720	Cond. No.	1.00e+03

Discussion of Methodology and Conclusion

Linear regressions are relatively simple but powerful tools for measuring the predictive ability of certain features for others. Throughout this project, I used various regression models to either predict the feature in question, or observe the relationship and significance of features relating to the Solow-Romer growth model. For the former kinds of regressions, I excluded values of low statistical significance at $\alpha = .05$ and refined choices in features to minimize multicollinearity. For the latter kinds of regressions, I focused on the inclusion features relevant to the Solow-Romer growth model even at the expense of unnecessary multicollinearity and inclusion of statistically insignificant features.

Building models to predict GDP and TFP was easily achieved as it seemed there were many features that strongly predicted both. However, given the concerns of multicollinearity throughout the various models built, it would seem that many of these features actually explain the same variance and are statistically similar enough to only marginally help our predictive power. I would say with regards to having access to features that predict GDP and TFP, was a strong success. However, given the strong multicollinearity in some of our models, I wasn't able

to either confidently determine unique coefficients in multivariate models, or recognize any statistically unique predictive power of each of the features. Because of these shortcomings, I can not infer causal inference from the features alone. However, as far as predictions are concerned, it would seem that the Solow-Romer model is strongly supported by the data reflecting its variables.

Works Cited

Solow, Robert M. “Technical Change and the Aggregate Production Function.” *The Review of Economics and Statistics*, vol. 39, no. 3, 1957, pp. 312–320. *JSTOR*, www.jstor.org/stable/1926047. Accessed 19 Dec. 2020.

Romer, Paul M. 1994. "The Origins of Endogenous Growth." *Journal of Economic Perspectives*, 8 (1): 3-22.