Rachit Agarwal

3/1/2019

Analyzing Gross Metropolitan Products

**Introduction**

After the civil war, the United States took part in what was known as the second industrial revolution. During this time, almost eleven million people migrated from rural to urban areas. Today the population of our metropolitan areas is getting denser as even more people are migrating to the cities. It is necessary that the U.S. economy ensures financial stability and growth during this migration. This paper uses records taken from the U.S Bureau of Economic Analysis to figure out how the gross metropolitan product (GMP) is related to other aspects of our economy. The Bureau gives us estimates of population from the census, shares of GMP from four sectors: finance, information, communication and technology (ICT), professional and technical services (Prof.Tech) and management. Metropolitan statistical areas (MSA)’s from 2006 are also given to us. We will use this data to explore and test the supra-linear scaling law. Some academicians clam that his law states that GMP is a supra-linear function of population size. That is that as population of a city increases, GMP increases supra-linearly. Another goal is to test alternative models for the four sectors of the economy.

**Methods**

**Handling Missing Data**

The data was provided to us by the U.S Bureau of Economic Analysis and this data has missing values in some entries of the last four columns which are the shares of the GMP (given as a percentage) in the four sectors of the economy. The Bureau has determined that releasing the data would reveal sensitive information about individual companies. Companies do not to release any indication of financial performance to other companies. This would be competitive disadvantage to companies who share this confidential information. In order to handle this missing data, we created an analysis sample. Missing data was removed from the analysis sample for finance ,prof.tech and ict since these are the variables we are most interested in. The following table indicates the number of missing values for each variable as well as the number of missing values in the entire dataset.

*Table 1:*

|  |  |
| --- | --- |
| Variable | Number of Missing Values |
| Finance | 9 |
| Prof.Tech | 80 |
| ICT | 41 |
| Management | 109 |
| All Data | 239 |

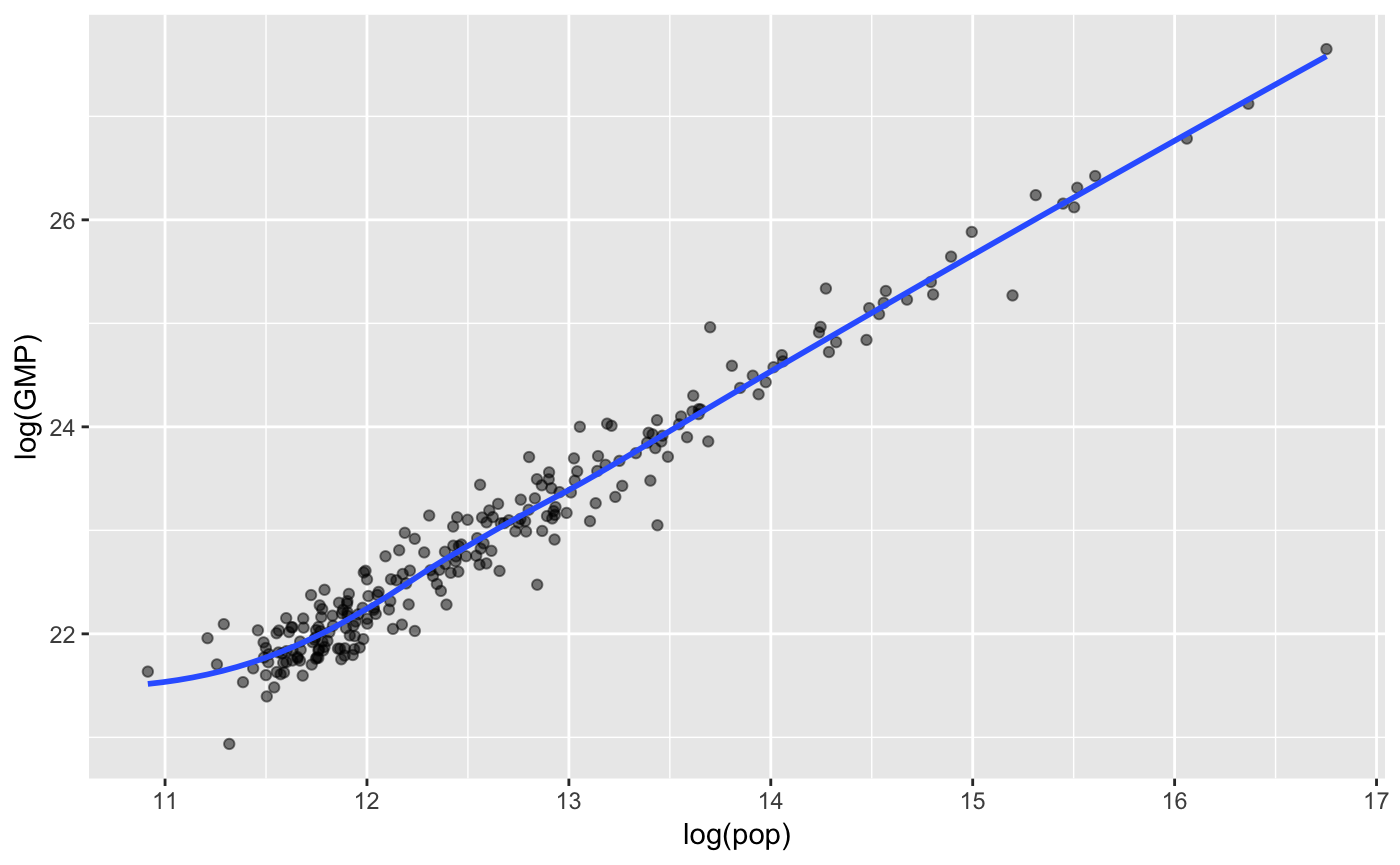
We can see from the data that if we were to remove all the rows with missing values, we would remove most our rows. We can also see that finance has the least number of missing rows. This shows us that finance accounts for a big proportion of the economy for many dense metropolitan areas. 2

**Methods of Statistical Models**

The models investigated to predict GMP are the supra-linear scaling law, simple linear regression and multiple linear regression. The multiple linear regression model uses a combination of finance, prof.tech and ict to determine GMP. The supra-linear scaling law ,where Y is the GMP and N is the population for some c>0 and b>1, states that if we encourage people living in less densely populated areas to move to already concentrated areas, economic productivity would improve. This super-linear scaling law was introduced by Bettencourt and our goal was to analyze and validate this for a sample of data taken from 2006. From our calculations we were able to validate this scaling law and determine the coefficients c and b for our data. The figure below shows us log(GMP) vs log(Population). Notice how significantly linear this relationship is.

*Figure 1:*

Log(GMP) vs Log(Population)



While the supra-linear scaling law provides us with a great estimate for estimating GMP using population size, we observe alternative hypothesis that may correlate to a similar but stronger representation of GMP by observing the shares of the economy in different sectors. We hypothesize that finance, ict and prof.tech rather than population size are a more promising representation of the GMP. This is because the companies and industries in a city may have more of a supra-linear relationship to economic success than the population size. From this hypothesis, we observe four alternative models: finance, ict, finance+ict, and finance+ict+prof.tech, which are the predictors of GMP. We select these four models because we believe that finance and communication technology are the major factors that correlate to the success of the economy of a city. Two of the biggest and most well-known cities in the United States: San Francisco and New York City are hubs for technology and finance. The two simple linear regression models were chosen to test weather just finance or just ict make a difference when trying to predict GMP. With these simple models, we can see if see if finance or ict is a bigger contributor to GMP. With a combination of finance and ict we test weather metropolitan cities with a large percentage of both these economic sectors correlate to a greater GMP. Additionally by adding prof.tech to finance and ict, we get a more diverse outlook on GMP by including more information.

**Assessment of Models**

In order to assess the performance of each of our models we use three methods, mean squared error (MSE) and 5-fold cross validation, and R squared. MSE allows us to assess the accuracy of our model by comparing the predicted values with the actual values in our dataset. Ideally, we would want our model to have a very low MSE. This allows us to conclude which model minimizes error. Additionally, 5 fold cross validation was used to test our models. For this method we split the data randomly into 5 group and use each group as a test dataset one at a time. This technique minimizes bias and reduces variability insuring that conclusions from are analysis are correct. Finally, we use the adjusted R-squared to make an assessment for our models. The coefficient of determination R-squared is a statistical measure of how close the data points are to the fitted regression line. We use the adjusted R squared because it takes into account the number of variables. R squared increases as the number of variables increase. However, by using adjusted R squared we solve this issue. We would ideally want a high R-squared value because R-squared is the percentage of the variation that is explained by the linear model. After testing our model using 5-fold cross validation, we used another model assessment technique. We held back a randomly selected 1/3 of the original MSA data and now use this holdout data as our new test data. We then used this holdout data to compute mean squared errors and again test our models. This was very useful because we have an entirely different set of data points for our test dataset. We can ensure that our model is not biased now since we can test on new data.

**Results**

The alternate models and the supra-linear scaling law were assessed with Adjusted R-squared ,5-fold cross validation, and a holdout MSE. The performance of these models is summarized in Table 2 below:

*Table 2*

Model Assessment

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Adjusted R-squared | 5-Fold Cross Validation | Holdout Sample MSE |
| Supra-Linear | 0.190 | 0.064 | 0.061 |
| Finance | 0.113 | 0.070 | 0.070 |
| ICT | 0.153 | 0.067 | 0.069 |
| Finance+ICT | 0.274 | 0.056 | 0.057 |
| Finance+ICT+Prof.Tech | 0.307 | 0.053 | 0.053 |

In this table we can compare the Adjusted R squared and 5-fold cross validation error for each of the different models. If we look at the plausibility of the supra-linear power law graphically, we may assume that this model has a strong linear relationship. However, if we asses this model further using R-squared and MSE we see that it is not the best model. Although this model has a low R squared value, it has significant coefficients and a low MSE. Overall the supra-linear power law does a good job estimating the GMP from the population of the city. From our table we can see that GMP is best estimated from the shares of the economy coming from finance, ict, and prof.tech. This model has the lowest 5-fold cross validation output and the highest R-squared value. This means that the multiple linear regression model has the lowest mean squared error and accounts for the highest percentage of the variation of the actual dataset of metropolitan statistical areas. We can conclude that a multiple linear regression model taking into account shares of the economy deriving from finance, ict, and prof.tech is a better model for predicting GMP than the population size of a metropolitan city. Another MSE was also calculated using a holdout sample. These results are also consistent with the R-Squared and 5-fold cross validation. That is our the lowest MSE using our holdout sample is still the last multiple linear regression model. Finally, we did a two-way ANOVA test to assess the effects of population when predicting GMP with ict, finance, and prof.tech as predictor variables. This additional hypothesis testing was done to asses weather a model predicting GMP with ict, finance, and prof.tech is significantly different than a model predicting GMP with ict, finance, prof.tech and population. After interpreting the results of the ANOVA test, we see that population is a significant factor in predicting GMP.

**Discussion**

From our analysis of the supra-linear power scaling law, we were able to validate that population is a key factor and has a strong relationship with the GMP of each metropolitan city. We were able to calculate the scaling coefficient (c = 6607.76) and power degree (b = 1.23). This is important because it shows us how historically the industrial revolution and urbanization improved our economy. This provides hope for our economy has population rises in the metropolitan areas of the United States. Although the supra-linear law is strongly related to GMP, it is not the strongest linear regression model. The inclusion of shares of GMP form different sectors of the economy actually provide us with a closer estimate for per capita GMP and give us a stronger linear relationship. This information is useful for the government and economy of the United States as it shows us what makes a cities GMP strong. After we further analyze the GMP of metropolitan cities, this information can help the United States grow its economy as a whole.

### Appendix

### Overview:

This file gathers information on the msa data. We summarize the missing values from the data. We do exploratory analysis on the data. Finally, we implement and test linear regession models using many differnt methods for assesment.

### Apendex 1 Detail of Statistical Models:

Conceptual proofs

1. Find log(Y/N) = B0 + B1log(N) from Y = c(N^b) where c>0 and b>1.

Y = cN^b

Y/ N = cN^(b-1)

log(Y/N) = log(cN^(b-1))

log(Y/N) = log(c) + (b-1)log(N)

since c>0 and b>1 we have B0 and B1>0

1. Find log(Y) = B0 + (1+B1)log(N) from log(Y/N) = B0 + B1log(N):

log(Y/N) = B0 + B1log(N)

log(Y) – log(N) = B0 + B1log(N)

log(Y) – log(N) – B1log(N) = B0

log(Y) – (1+B1)log(N) = B0

log(Y) = B0 + (1+B1)log(N)

For our first hypothesis, when taking into account shares of the economy deriving from finance and ict and how they relate to GMP it produces a better linear model than the relationship between GMP and population size.

Additionally, when taking into account shares of the economy deriving from finance, ict and prof.tech and how theyrelate to GMP it produces a better linear model than the relationship between GMP and population size.

### External Requirments

library(tidyverse)

## ── Attaching packages ───────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0 ✔ purrr 0.3.0  
## ✔ tibble 2.0.1 ✔ dplyr 0.7.8  
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ──────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)  
library(knitr)  
library(dplyr)

### Apendex 2 Exploratory Analyis:

# read in missing data  
msadata=read.csv("http://dept.stat.lsa.umich.edu/~bbh/s485/data/gmp-2006.csv")  
## read in holdout sample  
holdout=read.csv("http://dept.stat.lsa.umich.edu/~bbh/s485/data/gmp-2006-holdout.csv")  
## add all neccary variables to the data  
GMP = as.numeric(msadata$pop)\*as.numeric(msadata$pcgmp)  
msadata=data.frame(msadata,GMP)  
  
#number of missing values   
sum(is.na(msadata$finance))

## [1] 9

sum(is.na(msadata$prof.tech))

## [1] 80

sum(is.na(msadata$ict))

## [1] 41

sum(is.na(msadata$management))

## [1] 109

sum(is.na(msadata))

## [1] 239

## filter complete cases for the four sectors  
finance <- mean(complete.cases(dplyr::select(msadata, finance)))  
prof.tech <- mean(complete.cases(dplyr::select(msadata, prof.tech)))  
ict <- mean(complete.cases(dplyr::select(msadata, ict)))  
management <- mean(complete.cases(dplyr::select(msadata, management)))  
#filter complete cases by pairs  
mean(complete.cases(dplyr::select(msadata, finance, prof.tech)))

## [1] 0.6557377

mean(complete.cases(dplyr::select(msadata, finance, ict)))

## [1] 0.795082

mean(complete.cases(dplyr::select(msadata, finance, management)))

## [1] 0.545082

mean(complete.cases(dplyr::select(msadata, prof.tech, ict)))

## [1] 0.5122951

mean(complete.cases(dplyr::select(msadata, prof.tech, management)))

## [1] 0.5368852

mean(complete.cases(dplyr::select(msadata, ict, management)))

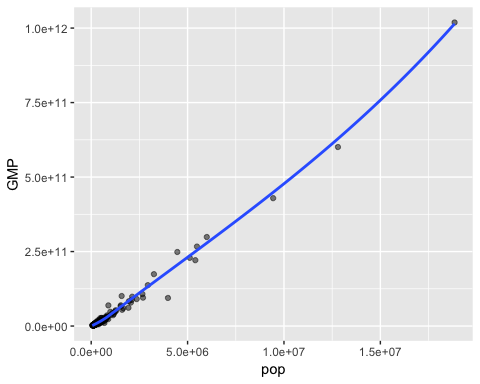
## [1] 0.397541

#complete cases for entire dataset  
mean(complete.cases(dplyr::select(msadata,finance,prof.tech,ict,management)))

## [1] 0.3729508

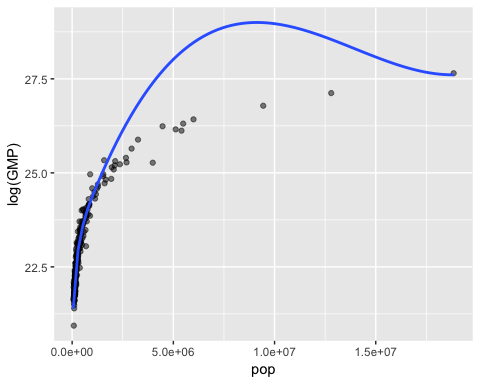
##final dataset#  
  
##scaterplots  
ggplot(data=msadata,mapping = aes(x=pop,y=GMP)) + geom\_point(alpha=0.5)+geom\_smooth(se=FALSE)

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



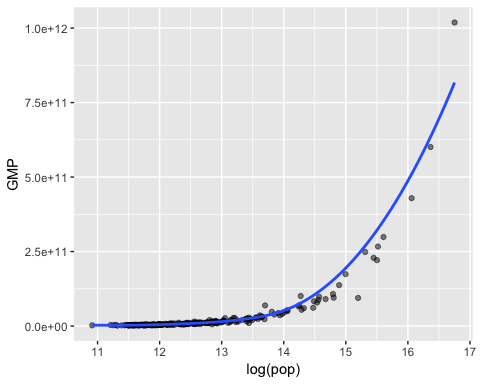
ggplot(data=msadata,mapping = aes(x=pop,y=log(GMP))) + geom\_point(alpha=0.5)+geom\_smooth(se=FALSE)

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



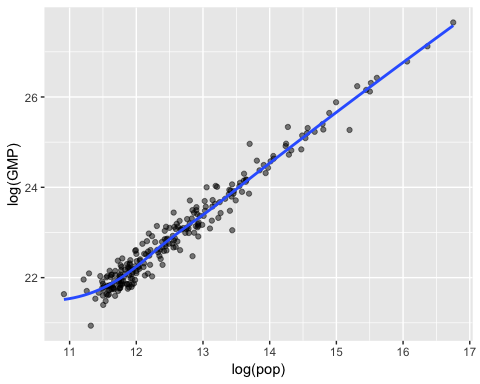
ggplot(data=msadata,mapping = aes(x=log(pop),y=GMP)) + geom\_point(alpha=0.5)+geom\_smooth(se=FALSE)

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



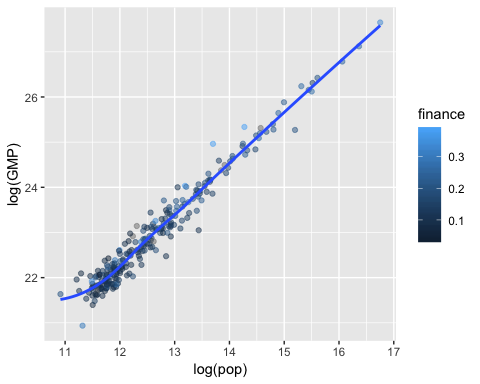
ggplot(data=msadata,mapping = aes(x=log(pop),y=log(GMP))) + geom\_point(alpha=0.5)+geom\_smooth(se=FALSE)

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



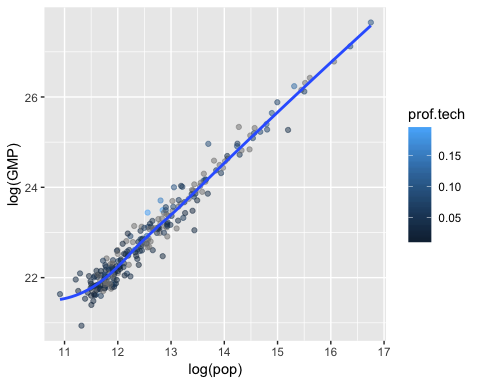
##scaterplots with colors corresponding to other variables  
ggplot(data=msadata,mapping = aes(x=log(pop),y=log(GMP),color=finance)) + geom\_point(alpha=0.5)+geom\_smooth(se=FALSE)

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



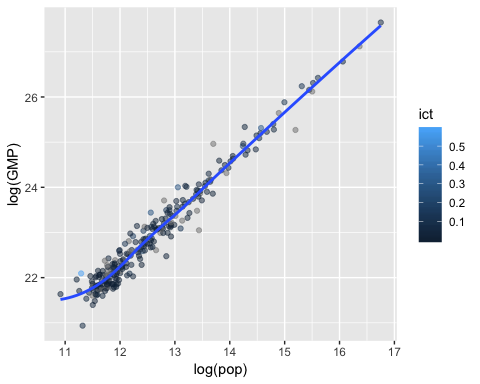
ggplot(data=msadata,mapping = aes(x=log(pop),y=log(GMP),color=prof.tech)) + geom\_point(alpha=0.5)+geom\_smooth(se=FALSE)

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



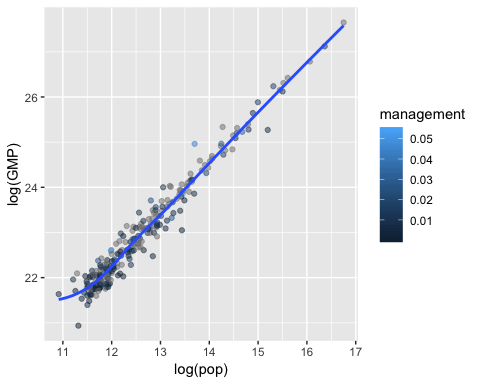
ggplot(data=msadata,mapping = aes(x=log(pop),y=log(GMP),color=ict)) + geom\_point(alpha=0.5)+geom\_smooth(se=FALSE)

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



ggplot(data=msadata,mapping = aes(x=log(pop),y=log(GMP),color=management)) + geom\_point(alpha=0.5)+geom\_smooth(se=FALSE)

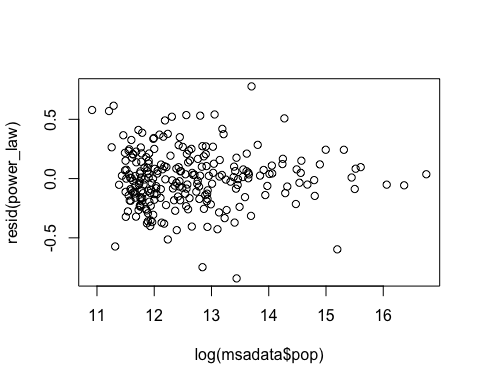
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

 ### Apendex 3 Fitting the power law model:

## fitting model  
power\_law<-lm(log(msadata$pcgmp)~log(msadata$pop))  
summary(power\_law)

##   
## Call:  
## lm(formula = log(msadata$pcgmp) ~ log(msadata$pop))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.84226 -0.13993 0.00157 0.12942 0.77779   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.79623 0.18350 47.936 < 2e-16 \*\*\*  
## log(msadata$pop) 0.12326 0.01449 8.509 1.86e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.238 on 242 degrees of freedom  
## Multiple R-squared: 0.2303, Adjusted R-squared: 0.2271   
## F-statistic: 72.4 on 1 and 242 DF, p-value: 1.86e-15

## verifying with residuals  
plot(log(msadata$pop),resid(power\_law))



## in sample loss  
in\_sample\_loss = mean(resid(power\_law)^2)  
in\_sample\_loss

## [1] 0.05619567

## cross validation  
cv.lm <- function(data, formulae, nfolds = 5)   
{  
 data <- na.omit(data)  
 formulae <- sapply(formulae, as.formula)  
 n <- nrow(data)  
 fold.labels <- sample(rep(1:nfolds, length.out = n))  
 mses <- matrix(NA, nrow = nfolds, ncol = length(formulae))  
 colnames <- as.character(formulae)  
 for (fold in 1:nfolds) {  
 test.rows <- which(fold.labels == fold)  
 train <- data[-test.rows, ]  
 test <- data[test.rows, ]  
 for (form in 1:length(formulae)) {  
 current.model <- lm(formula = formulae[[form]], data = train)  
 predictions <- predict(current.model, newdata = test)  
 test.responses <- eval(formulae[[form]][[2]], envir = test)  
 test.errors <- test.responses - predictions  
 mses[fold, form] <-mean(test.errors^2)  
 }  
 }  
 return(colMeans(mses))  
}  
cross\_valid<-"log(msadata$pcgmp)~log(msadata$pop)"  
cv = cv.lm(msadata,cross\_valid,5)

cv

## [1] 0.05619567

From the proofs derived above, we know that log(GMP) = log(c) + log(b-1)log(pop), therefore form the model summary of the power law we can determine what the values of c and b are. The coeficent for the intercept is 8.796 so 8.796 log(c) and c = 6607.76. The coeficent for log(pop) is 0.123 which is equal to b-1 so b = 1.123. This is compatible with the supra-linear power law because c>0 and b>1.

Accoridng to the residual plot, we see that there is no indication of a pattern and therefore there is a porper fit for the power law and we can use the summary function and beleive the standard eror it provides which is 0.238 on 242 degrees of freedom.

### Apendex 4 Fitting the assement of alternate models:

## now we have to delete missing values for finance,ict and prof.tech from our dataset ##  
analyis\_data = drop\_na(msadata,finance,ict,prof.tech)  
## alternate models  
original\_mod<-lm(log(analyis\_data$pcgmp)~log(analyis\_data$pop))  
summary(original\_mod)

##   
## Call:  
## lm(formula = log(analyis\_data$pcgmp) ~ log(analyis\_data$pop))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.74764 -0.17161 -0.00932 0.16367 0.61805   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.76798 0.29080 30.15 < 2e-16 \*\*\*  
## log(analyis\_data$pop) 0.12546 0.02328 5.39 3.62e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2541 on 119 degrees of freedom  
## Multiple R-squared: 0.1962, Adjusted R-squared: 0.1895   
## F-statistic: 29.05 on 1 and 119 DF, p-value: 3.62e-07

fin\_pop<-lm(log(analyis\_data$pcgmp)~analyis\_data$finance)  
summary(fin\_pop)

##   
## Call:  
## lm(formula = log(analyis\_data$pcgmp) ~ analyis\_data$finance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.92702 -0.15444 -0.01086 0.17275 0.67258   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.11740 0.05806 174.253 < 2e-16 \*\*\*  
## analyis\_data$finance 1.43221 0.35494 4.035 9.69e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2659 on 119 degrees of freedom  
## Multiple R-squared: 0.1204, Adjusted R-squared: 0.113   
## F-statistic: 16.28 on 1 and 119 DF, p-value: 9.686e-05

tech\_pop<-lm(log(analyis\_data$pcgmp)~analyis\_data$ict)  
summary(tech\_pop)

##   
## Call:  
## lm(formula = log(analyis\_data$pcgmp) ~ analyis\_data$ict)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73920 -0.16850 0.00021 0.17923 0.55992   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.26212 0.02764 371.253 < 2e-16 \*\*\*  
## analyis\_data$ict 1.57725 0.33155 4.757 5.56e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2598 on 119 degrees of freedom  
## Multiple R-squared: 0.1598, Adjusted R-squared: 0.1527   
## F-statistic: 22.63 on 1 and 119 DF, p-value: 5.565e-06

mod4<-lm(log(analyis\_data$pcgmp)~analyis\_data$ict+analyis\_data$finance)  
summary(mod4)

##   
## Call:  
## lm(formula = log(analyis\_data$pcgmp) ~ analyis\_data$ict + analyis\_data$finance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.96007 -0.12766 0.00014 0.16772 0.60414   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.04258 0.05444 184.477 < 2e-16 \*\*\*  
## analyis\_data$ict 1.60666 0.30698 5.234 7.31e-07 \*\*\*  
## analyis\_data$finance 1.46746 0.32118 4.569 1.21e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2405 on 118 degrees of freedom  
## Multiple R-squared: 0.2861, Adjusted R-squared: 0.274   
## F-statistic: 23.64 on 2 and 118 DF, p-value: 2.318e-09

mod5<-lm(log(analyis\_data$pcgmp)~analyis\_data$ict+analyis\_data$finance+analyis\_data$prof.tech)  
summary(mod5)

##   
## Call:  
## lm(formula = log(analyis\_data$pcgmp) ~ analyis\_data$ict + analyis\_data$finance +   
## analyis\_data$prof.tech)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8836 -0.1160 -0.0098 0.1510 0.6356   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.99635 0.05613 178.090 < 2e-16 \*\*\*  
## analyis\_data$ict 1.17329 0.34389 3.412 0.000887 \*\*\*  
## analyis\_data$finance 1.18952 0.33182 3.585 0.000493 \*\*\*  
## analyis\_data$prof.tech 2.22432 0.86367 2.575 0.011257 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.235 on 117 degrees of freedom  
## Multiple R-squared: 0.3244, Adjusted R-squared: 0.3071   
## F-statistic: 18.73 on 3 and 117 DF, p-value: 5.484e-10

##in-sample loss   
original\_in\_sample\_loss = mean(resid(original\_mod)^2)  
fin\_in\_sample\_loss = mean(resid(fin\_pop)^2)  
tech\_in\_sample\_loss = mean(resid(tech\_pop)^2)  
mod4\_in\_sample\_loss = mean(resid(mod4)^2)  
mod5\_in\_sample\_loss = mean(resid(mod5)^2)  
  
original\_in\_sample\_loss

## [1] 0.06351577

fin\_in\_sample\_loss

## [1] 0.06951125

fin\_in\_sample\_loss

## [1] 0.06951125

tech\_in\_sample\_loss

## [1] 0.06639525

mod4\_in\_sample\_loss

## [1] 0.05641494

mod5\_in\_sample\_loss

## [1] 0.05338831

## cross validataon ##  
cross\_valid1<-"log(analyis\_data$pcgmp)~log(analyis\_data$pop)"  
cross\_valid2<-"log(analyis\_data$pcgmp)~analyis\_data$finance"  
cross\_valid3<-"log(analyis\_data$pcgmp)~analyis\_data$ict"  
cross\_valid4<-"log(analyis\_data$pcgmp)~analyis\_data$ict+analyis\_data$finance"  
cross\_valid5<-"log(analyis\_data$pcgmp)~analyis\_data$ict+analyis\_data$finance+analyis\_data$prof.tech"  
cv1 = cv.lm(analyis\_data,cross\_valid1,5)

cv1

## [1] 0.06351577

cv2 = cv.lm(analyis\_data,cross\_valid2,5)

cv2

## [1] 0.06951125

cv3 = cv.lm(analyis\_data,cross\_valid3,5)

cv3

## [1] 0.06639525

cv4 = cv.lm(analyis\_data,cross\_valid4,5)

cv4

## [1] 0.05641494

cv5 = cv.lm(analyis\_data,cross\_valid5,5)

cv5

## [1] 0.05338831

### Apendex A Nested alternative model:

## nested model for best altetrnative model  
nested\_model<- lm(log(analyis\_data$pcgmp)~analyis\_data$ict+analyis\_data$finance+analyis\_data$prof.tech+log(analyis\_data$pop))  
summary(nested\_model)

##   
## Call:  
## lm(formula = log(analyis\_data$pcgmp) ~ analyis\_data$ict + analyis\_data$finance +   
## analyis\_data$prof.tech + log(analyis\_data$pop))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.76786 -0.13412 0.00007 0.12591 0.60118   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.29432 0.30054 30.925 < 2e-16 \*\*\*  
## analyis\_data$ict 1.19248 0.33735 3.535 0.000587 \*\*\*  
## analyis\_data$finance 0.83030 0.35882 2.314 0.022430 \*   
## analyis\_data$prof.tech 1.37578 0.91922 1.497 0.137191   
## log(analyis\_data$pop) 0.06385 0.02687 2.376 0.019135 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2304 on 116 degrees of freedom  
## Multiple R-squared: 0.3557, Adjusted R-squared: 0.3335   
## F-statistic: 16.01 on 4 and 116 DF, p-value: 1.822e-10

## anova test comparing nestred model with alternative model   
anova(nested\_model,mod5)

## Analysis of Variance Table  
##   
## Model 1: log(analyis\_data$pcgmp) ~ analyis\_data$ict + analyis\_data$finance +   
## analyis\_data$prof.tech + log(analyis\_data$pop)  
## Model 2: log(analyis\_data$pcgmp) ~ analyis\_data$ict + analyis\_data$finance +   
## analyis\_data$prof.tech  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 116 6.1602   
## 2 117 6.4600 -1 -0.29982 5.6459 0.01913 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Apendex B Holdout sample:

# B #  
## drop missing values from hold out dataset  
hold\_data = drop\_na(msadata,finance,ict,prof.tech)  
# add yi or orginal value to holdout dataset  
hold\_data = data.frame(hold\_data,log(hold\_data$pcgmp))  
  
  
#predictions on holdout data   
predictions\_original<-predict(original\_mod, newdata = hold\_data)  
predictions\_fin<-predict(fin\_pop, newdata = hold\_data)  
predictions\_tech<-predict(tech\_pop, newdata = hold\_data)  
predictions\_mod4<-predict(mod4, newdata = hold\_data)  
predictions\_mod5<-predict(mod5, newdata = hold\_data)  
  
## calculate test errors and MSE for each model   
holdout\_errors\_original = hold\_data$log.hold\_data.pcgmp - predictions\_original  
MSE\_original<-mean((holdout\_errors\_original)^2)  
MSE\_original

## [1] 0.06351577

holdout\_errors\_fin = hold\_data$log.hold\_data.pcgmp - predictions\_fin  
MSE\_fin<-mean((holdout\_errors\_fin)^2)  
MSE\_fin

## [1] 0.06951125

holdout\_errors\_tech = hold\_data$log.hold\_data.pcgmp - predictions\_tech  
MSE\_tech<-mean((holdout\_errors\_tech)^2)  
MSE\_tech

## [1] 0.06639525

holdout\_errors\_mod4 = hold\_data$log.hold\_data.pcgmp - predictions\_mod4  
MSE\_mod4<-mean((holdout\_errors\_mod4)^2)  
MSE\_mod4

## [1] 0.05641494

holdout\_errors\_mod5 = hold\_data$log.hold\_data.pcgmp - predictions\_mod5  
MSE\_mod5<-mean((holdout\_errors\_mod5)^2)  
MSE\_mod5

## [1] 0.05338831