Winter 18

Investigating Gross Metropolitan Product Theories with Statistical Modeling

Jack Schaffer

STATS 485 – Capstone Seminar

Table of Contents

Introduction

Around the world and within the United States, population is an ever-growing factor with respect to how we function economically. Regions in the US have become denser and as population grows, it is necessary that our economy becomes more efficient and effective to ensure financial stability and growth. In order to gauge GMP, statisticians continually define models to infer how GMP is related to underlining characteristics of our economy. This paper investigates records from the U.S. Bureau of Economic Analysis, which includes estimates of gross metropolitan product (GMP), and the shares of GMP in each of three sectors of economic activity: finance, information, communication and technology (ICT), and professional and technical services (Prof.Tech), along with the U.S. Census Bureau's estimates of population. This data is used to explore the supra-linear scaling law, as it relates to GMP and population size of metropolitan statistical areas, and examine various alternative models that describe GMP.

Methodology

Handling of Missing Data

The data provided from the U.S. Bureau of Economic Analysis, which provided the shares of GMP from different sectors of the economy used in this study's analysis, had instances of missing data for various metropolitan statistical areas. The reason for the missing information from the U.S. Bureau of Economic Analysis was not due to randomness; in fact, the rational is straightforward. The U.S. Bureau of Economic is subject to ensuring privacy to corporations and business by not releasing information that would reveal any indication financial performance

This suggests that data is rescinded from the dataset because there may be a limited amount of organizations within a specific economy sector in a metropolitan statistical area would enable viewers to infer the estimated share of GMP that a specific company contributes in the region up to a dollar amount if the share of the economy was released in this dataset.

The instances in which data was missing for the share of the GMP from Finance, ICT, and Prof.Tech were removed from the analysis sample. The following table indicates the difference in rows and mean of each variable utilized in our analysis:

Variable	Row Difference	Mean Difference
Population Size	113747.9	123
Share of Finance	114	0.0015544
Share of ICT	82	-0.0045088
Share of Prof.Tech	43	0.0012284

Table 1: Difference in number of rows and means for population, finance, ict, and prof.tech from all complete cases and the analysis sample subset

As you can see from the data, while population size mean varies, the means of the shares of finance, ICT, and Prof.Tech are quite small. This table allows us to justify using the smaller analysis sample because they are similar and representative of the complete cases.

Modeling Methods

The methods used to determine GMP were the supra-linear scaling law and additional simple and multiple linear regression models using the shares of the economy from Finance, ICT, and Prof.Tech. The supra-linear scaling law, $Y \approx cN^b$, has been used in the past to "present empirical evidence indicating that the processes relating urbanization to economic development…are very general, being shared by all cities belonging to the same urban system" (Bettencourt 1). In *Growth,*

innovation, scaling, and the pace of life in cities by Bettencourt, Lobo, Helbing, Kühnert, and West, the statisticians state "quantities revolving around wealth creation, [or GMP in this study's case], are shown to be power law functions of population size with scaling exponents." where quantities reflecting wealth creation, such as GMP, are raised to the power of about 1.2 (Bettencourt 1).

The supra-linear scaling law, introduced by Bettencourt, was validated where b is approximately 1.2. In order to do this, we had to solve for coefficient c and b, from the supra-linear scaling power law (Appendix - Conceptual Preliminaries). From our calculations, we were able to determine that our data is fit where b is 1.13. To get an understanding of what this fit looks like observe *figure 1:*

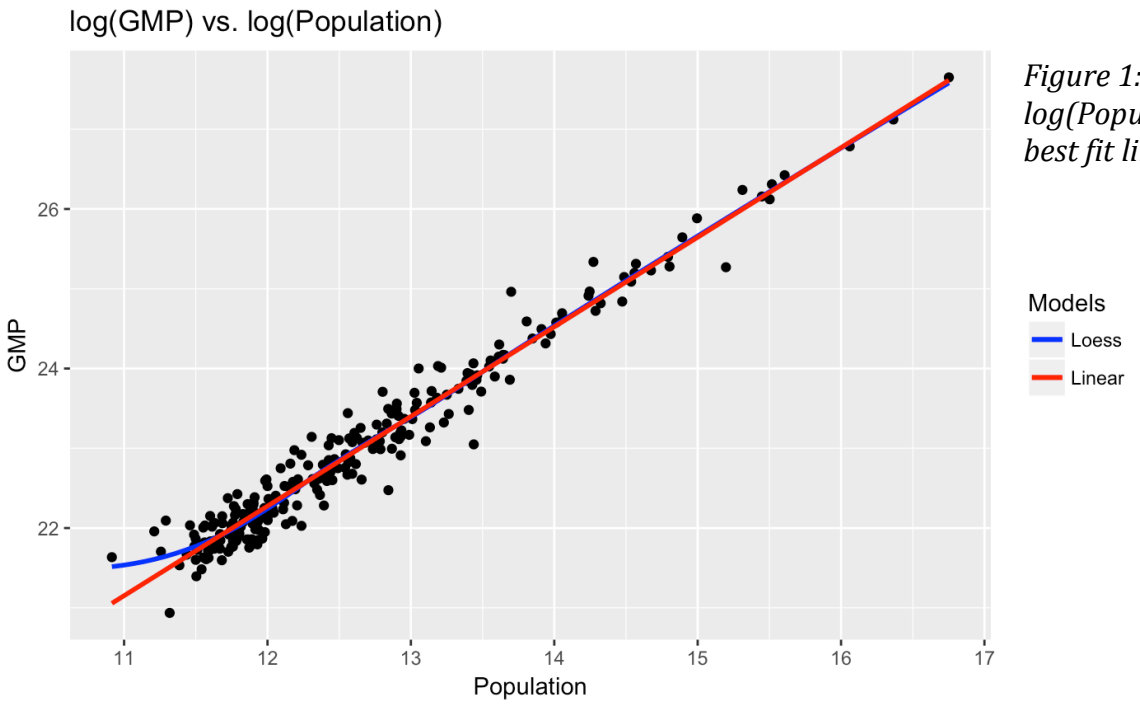

Figure 1: log(GMP) plotted against log(Population) with Leoss and Linear best fit lines.

While the supra-linear scaling law is representative of GMP, we observe alternative hypotheses that may correlate to a similar, more promising representation of GMP by observing the shares of the economy for three sector, Finance, ICT, and Prof.Tech rather than population size. From these three characteristics we construct four

alternative models: finance, ict,, finance + ict, and finance + ict + prof.tech. The reasoning behind the selection of these models is the distribution of a metropolitan statistical area's greatest economic sector, which is shown in *figure 2*:

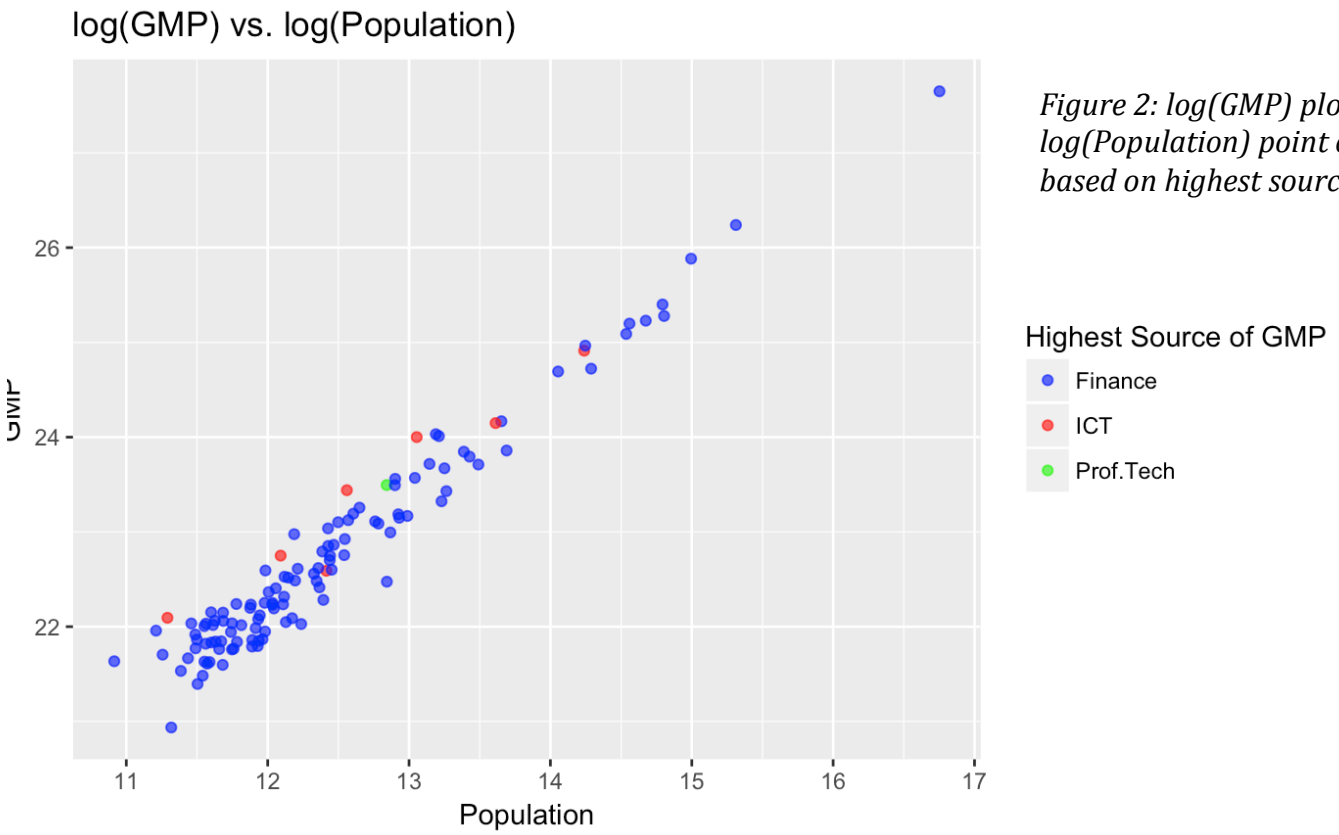

Figure 2: log(GMP) plotted against log(Population) point color varying based on highest source of GMP

The initial two simple linear models that utilize solely finance and ICT were chosen because they are the larger players with respect to GMP in our data set, while Prof.Tech is the highest source of GMP in only one metropolitan statistical area. Additionally, by adding ICT, and then Prof.Tech to finance, we allow a diverse outlook on GMP by including more information regarding the economic ecosystem of each metropolitan statistical area.

Model Assessment

In order to assess the performance of each model, we observe their mean squared error (MSE), which allows us to infer the effectiveness of our predictive model. Utilizing MSE, we can conclude which models minimize error. Additionally,

5-fold cross validation was used. This technique is common in practice as it minimizes bias and reduces variability. By utilizing 5-fold cross validation we are able to avoid issues of over fitting and ensure that the conclusions from our analysis are concrete. Lastly, the ANOVA Test was used to compare our models to each other. This test is used to determine if the reduction of MSE, compared to different models, is significant. If it is significant, we can infer that a model performs better than another.

Results

The supra-linear power scaling law and alternative models were verified through in-sample loss calculation and 5-fold cross validation. The performance of these models are conveyed in *table 2:*

Model	In-Sample Loss	5-Fold Cross Validation
Supra-Linear	0.064	0.064
Finance	0.069	0.072
ICT	0.066	0.069
Finance + ICT	0.056	0.061
Finance + ICT + Prof.Tech	0.053	0.057

Table 2: Model In-Sample Loss and 5-Fold Cross Validation

According to our table, we can see that GMP is best estimated from the linear regression model, which observes shares of the economy from the finance, ICT, and Prof.Tech sectors. It has the lowest In-Sample Loss and 5-Fold Cross Validation output. Additionally, the p-value from ANOVA testing between Supra-Linear and Finance + ICT + Prof.Tech was 3.86e-05, which is extremely small. Thus we reject the null hypothesis that the supra-linear model is similar to the linear regression model. This indicates that the multiple linear regression model is a more optimal fit to our data.

Discussion

From our analysis and exploration of the supra-linear power scaling law, we were able to validate the work done in *Growth, Innovation, Scaling, and the pace of life in cities* by calculating the scaling coefficient (c = 6425.18) and power degree (b = 1.126). Additionally, our results confirmed that quantities representing wealth creation is correlated to population size with statistical significance. This is noteworthy as it provides hope for our economy as population rises. Although, while the supra-power scaling law models our data with significance, the inclusion of different aspects of the economy such as shares of GMP from different economic sectors provides a closer estimate for actual GMP. This information is useful as it allows economist to formulate ideologies about sector diversity and generate economic portfolios for metropolitan statistical areas, which can be used to ensure expected growth in our economy.

Revision Comments

After revisiting this paper following the peer-review and discussion with my GSI, I made a few changes to improve upon this document. I added the model assessment section under methods, which discusses mean squared error, 5-fold cross validation, and ANOVA testing. This section informs the reader on why I utilized these methods for assessment, which provides a solid foundation for the reader to follow and enables them to read along more succinctly with the subsequent sections, results and discussion.

Works Cited

Bettencourt, Luís M. A., et al. "Growth, Innovation, Scaling, and the Pace of Life in Cities." *Proceedings of the National Academy of Sciences*, National Academy of Sciences, 24 Apr. 2007, www.pnas.org/content/104/17/7301.

Appendix

The following code is required in order to carry out the computations of this paper.

```
# Load Dependencies
library(ggplot2)
library(knitr)
```

Overview

The computations in this document assesses the power-law scaling proposal, develop alternate models to predict GMP as linear but no supra-linear function of population size, as well as additional variables describing the nature of its local economy, and compare model fits utilizing loss functions. In order to carry out this analysis, various data structures and functions were used to organize our data and compute effectively in addition to the generation of visuals and figures to represent data.

Conceptual Preliminaries

If $Y \approx cN^b$ with c>0 and b>1, then $log(Y/N) \approx \beta_0 + \beta_1 log(N)$, for some β_0 and some $\beta_1 > 0$, and also that $log(Y) \approx \beta_0 + (1 + \beta_1) log(N)$.

We can show this with the following algebra.

Here we will find $log(Y/N) \approx \beta_0 + \beta_1 log(N)$ from $Y \approx cN^b$:

$$Y \approx cN^b$$

$$\frac{Y}{N} \approx cN^{b-1}$$

$$\log(Y/N) \approx \log(cN^{b-1})$$

$$\log(Y/N) \approx \log(c) + (b-1)\log(N)$$

And because c > 0 we have some β_0 and because b > 1, we have some $\beta_1 > 0$.

Now to show $log(Y) \approx \beta_0 + (1 + \beta_1) log(N)$ from $\log(Y/N) \approx \beta_0 + \beta_1 \log(N)$:

$$\log(Y/N) \approx \beta_0 + \beta_1 \log(N)$$

$$\log(Y) - \log(N) \approx \beta_0 + \beta_1 \log(N)$$

$$\log(Y) - \log(N) - \beta_1 \log(N) \approx \beta_0$$

$$\log(Y) - \log(N)(1 + \beta_1) \approx \beta_0$$

$$\log(Y) \approx \beta_0 + (1 + \beta_1)\log(N)$$

As for our hypotheses, the appearance of supra-linear scaling is present when taking into account shares of the economy deriving from finance, however when the share of the economy from finance and ict firms are taken into account, the appearance of supra-linear scaling disappears.

Additionally, the appearance of supra-linear scaling is present when taking into account shares of the economy deriving from finance, however when the share of the economy from finance, ict, prof.tech firms are taken into account, the appearance of supra-linear scaling disappears.

Exploratory Analysis

The following code reads in the data

```
# Read gmp-2006.csv and store it as a dataframe variable named msadata`
``{r}
#msadata = read.csv("http://dept.stat.lsa.umich.edu/~bbh/s485/data/gmp-
2006.csv")
msadata = read.csv("msadata.csv")

# Set rownames
row.names(msadata) <- msadata$MSA

msadata[,"pop"] = as.double(msadata[,"pop"])
msadata[,"pcgmp"] = as.double(msadata[,"pcgmp"])

# Create sub samples
msadata_omit_finance = msadata[complete.cases(msadata[,"finance"]),]
msadata_omit_prof.tech = msadata[complete.cases(msadata[,"prof.tech"]),
]
msadata_omit_ict = msadata[complete.cases(msadata[,"ict"]),]
finance_ict = intersect(msadata_omit_finance, msadata_omit_ict)
finance_ict_prof.tech = intersect(finance_ict, msadata_omit_prof.tech)
```

In this chunk of code we create the variables gmp and primary:

```
gmp = msadata$pcgmp * msadata$pop
msadata = data.frame(msadata, gmp)

# Create sub samples
msadata_omit_finance = msadata[complete.cases(msadata[,"finance"]),]
msadata_omit_prof.tech = msadata[complete.cases(msadata[,"prof.tech"]),
]
msadata_omit_ict = msadata[complete.cases(msadata[,"ict"]),]
finance_ict = intersect(msadata_omit_finance, msadata_omit_ict)
finance_ict_prof.tech = intersect(finance_ict, msadata_omit_prof.tech)

primary = rep("", nrow(finance_ict_prof.tech))
```

```
for (i in 1:length(primary)){
  if (finance_ict_prof.tech[i,"finance"] > finance_ict_prof.tech[i,"ict
"] & finance_ict_prof.tech[i,"finance"] > finance_ict_prof.tech[i,"prof
.tech"]){
    primary[i] = "finance"
  }
  else if (finance_ict_prof.tech[i,"ict"] > finance_ict_prof.tech[i,"fi
nance"] & finance_ict_prof.tech[i,"ict"] > finance_ict_prof.tech[i,"pro
f.tech"]){
    primary[i] = "ict"
  }
  else if (finance_ict_prof.tech[i,"prof.tech"] > finance_ict_prof.tech
[i,"finance"] & finance_ict_prof.tech[i,"prof.tech"] > finance_ict_prof
.tech[i,"ict"]){
    primary[i] = "prof.tech"
  }
}

finance_ict_prof.tech = data.frame(finance_ict_prof.tech, primary)
```

As discussed in the paper, missing data was omitted from our sample during analysis. The following code does this for us:

```
msadata_omit_nan = msadata[complete.cases(msadata),]
msadata_omit_management = msadata[complete.cases(msadata[,"management"]
),]
finance_manage = intersect(msadata_omit_finance,msadata_omit_management
)
finance_prof.tech = intersect(msadata_omit_finance, msadata_omit_prof.t
ech)
prof.tech_ict = intersect(msadata_omit_prof.tech, msadata_omit_ict)
prof.tech_manage = intersect(msadata_omit_prof.tech, msadata_omit_manag
ement)
ict_manage = intersect(msadata_omit_ict, msadata_omit_management)
finance_ict_manage = intersect(finance_ict, msadata_omit_management)
ict_prof.tech_manage = intersect(ict_manage, msadata_omit_prof.tech)

DF_Subsets = list(msadata,msadata_omit_finance, msadata_omit_ict, msada
ta_omit_prof.tech, msadata_omit_management, finance_ict,finance_ict_pro
f.tech,finance_manage, prof.tech_ict,ict_manage,prof.tech_manage, finan
ce_ict_prof.tech, finance_ict_manage, ict_prof.tech_manage, msadata_omi
t_nan)

Subsets = c("All", "Finance", "ICT", "Prof.Tech", "Management","Finance
+ ICT", "Finance + Prof.Tech", "Finance + Management","ICT + Prof.Tech"
, "ICT + Management", "Prof.Tech + Management", "Finance + ICT + Prof.T
ech", "Finance + ICT + Management", "ICT + Prof.Tech + Management", "Fu
ll")
Rows = c(nrow(msadata),nrow(msadata_omit_finance), nrow(msadata_omit_ic
t), nrow(msadata_omit_prof.tech),nrow(msadata_omit_management), nrow(fi
nance_ict),nrow(finance_prof.tech), nrow(finance_manage), nrow(prof.tec
h_ict), nrow(ict_manage), nrow(prof.tech_manage),nrow(finance_ict_prof.
```

```
tech), nrow(finance_ict_manage), nrow(ict_prof.tech_manage), nrow(msada
ta_omit_nan))

Mean_Finance = c()
Mean_ICT = c()
Mean_Prof.Tech = c()
Mean_Management = c()

for (df in DF_Subsets){
  Mean_Finance = c(Mean_Finance,mean(df[,"finance"]))
  Mean_ICT = c(Mean_ICT, mean(df[,"ict"]))
  Mean_Prof.Tech = c(Mean_Prof.Tech, mean(df[,"prof.tech"]))
  Mean_Management = c(Mean_Management,mean(df[,"management"]))
}

table1 = data.frame(Subsets, Rows, Mean_Finance, Mean_ICT, Mean_Prof.Te
ch, Mean_Management)
row.names(table1) = table1$Subsets

kable(table1[order(-Rows),])
```

The subset we will be using as our analysis sample is Finance + ICT + Prof.Tech. This table shows the differences of the Rows and Means for each of the three variables where their data is available:

```
difference = matrix(ncol = 3, nrow = 3)
difference[1,1] = "Finance"
difference[1,2] = "114"
difference[1,3] = "0.0015544"
difference[2,1] = "ICT"
difference[2,2] = "82"
difference[2,3] = "-0.0045088"
difference[3,1] = "Prof.Tech"
difference[3,2] = "43"
difference[3,3] = "0.0012284"
table = data.frame(difference)

kable(difference, col = c("Economy Source", "Row Difference", "Mean Dif
ference"))
```

The following code generates various scatter plots:

```
gmp_pop_gp = ggplot(msadata,aes(x = pop, y = gmp)) + labs(title = "GMP
vs. Population", x = "Population", y = "GMP")
gmp_pop_gp + geom_point() + geom_smooth(method = "loess", se = FALSE, a
es(color = "blue")) + geom_smooth(method = "lm", se = FALSE, aes(color
= "red")) + scale_color_manual(name = "Models", labels = c("Loess", "Li
near"), values = c("blue","red"))

loggmp_pop_gp = ggplot(msadata,aes(x = pop, y = log(gmp))) + labs(title
= "log(GMP) vs. Population", x = "Population", y = "log(GMP)")
loggmp_pop_gp + geom_point() + geom_smooth(method = "loess", se = FALSE
```

```
, aes(color = "blue")) + geom_smooth(method = "lm", se = FALSE, aes(col
or = "red")) + scale_color_manual(name = "Models", labels = c("Loess",
"Linear"), values = c("blue","red"))

gmp_logpop_gp = ggplot(msadata,aes(x = log(pop), y = gmp)) + labs(title
= "GMP vs. log(Population)", x = "log(Population)", y = "GMP")
gmp_logpop_gp + geom_point() + geom_smooth(method = "loess", se = FALSE
, aes(color = "blue")) + geom_smooth(method = "lm", se = FALSE, aes(col
or = "red")) + scale_color_manual(name = "Models", labels = c("Loess",
"Linear"), values = c("blue","red"))

loggmp_logpop_gp = ggplot(msadata,aes(x = log(pop), y = log(gmp))) + la
bs(title = "log(GMP) vs. log(Population)", x = "Population", y = "GMP")
loggmp_logpop_gp + geom_point() + geom_smooth(method = "loess", se = FA
LSE, aes(color = "blue")) + geom_smooth(method = "lm", se = FALSE, aes(
color = "red")) + scale_color_manual(name = "Models", labels = c("Loess
", "Linear"), values = c("blue","red"))
```

The preferred plot from above is log(GMP) vs. log(Population) plot as the plots clearly depict an upward linear trend, with less variance than the other plots.

The follow code generates plots that display other characteristics of the data such as shares of the economy from finance, ict, and prof.tech:

```
loggmp_logpop_gp + geom_point(alpha = 2/3, aes(colour = c(finance))) +
scale_colour_continuous(low = "red", name = "Share of Finance", na.valu
e = "green")

loggmp_logpop_gp + geom_point(alpha = 2/3, aes(colour = c(ict))) + scal
e_colour_continuous(low = "red", name = "Share of ICT", na.value = "gre
en")

loggmp_logpop_gp + geom_point(alpha = 2/3, aes(colour = c(prof.tech)))
+ scale_colour_continuous(low = "red", name = "Share of Prof.Tech", na.
value = "green")

loggmp_logpop_gp = ggplot(finance_ict_prof.tech,aes(x = log(pop), y = l
og(gmp))) + labs(title = "log(GMP) vs. log(Population)", x = "Populatio
n", y = "GMP")
loggmp_logpop_gp + geom_point(alpha = 2/3, aes(colour = primary)) + sca
le_color_manual(name = "Highest Source of GMP", labels = c("Finance", "
ICT", "Prof.Tech"), values = c("blue","red", "green"))
```

Interestingly from the primary variable we created to showcase which source of the economy is the highest contributor to an areas GMP, finance is overwhelmingly the most frequent, followed by ICT, and, lastly, Prof.Tech only having a higher share of the economy in a single datum. This may not be as useful as previously thought before this exploration.

```
loggmp_logpop_gp = ggplot(msadata,aes(x = log(pop), y = log(gmp))) + la
bs(title = "log(GMP) vs. log(Population)", x = "Population", y = "GMP")
loggmp_logpop_gp + geom_point() + geom_smooth(method = "loess", se = FA
```

```
LSE, aes(color = "blue")) + geom_smooth(method = "lm", se = FALSE, aes(
color = "red")) + scale_color_manual(name = "Models", labels = c("Loess
", "Linear"), values = c("blue","red"))
```

Now, we will linearly regress log GMPon the log of sample size:

```
mod1 = lm(log(pcgmp) ~ log(pop), data = msadata)
summary(mod1)
```

From the proofs we derived above, we known that $\log(pcgmp) = \log(c) + (b - 1)\log(pop)$, thus from the model summary we can determine what c and b are in the power law scaling formula.

The coefficient for the intercept is 8.7623, thus:

$$8.7623 = \log(c)$$

$$c = 6388.788$$

The coefficient for log(pop) is 0.12326, thus:

$$0.12326 = (b - 1)$$

$$b = 1.12326$$

These findings are compatible with the supra-linear power-law scaling hypothesis because it follows the requirements that c > 0 and b > 1.

Now we will observe the residual of the linear model to determine the credibility of the model:

```
plot(residuals(mod1), xlab = "Fitted Values", ylab = "Residual")
abline(0,0)
```

According to the residual plot, we can see that there is no indication of a pattern. This suggests that the fit is proper. If there was an indication of a pattern we would consider other variables, transformations of current variables, or increased degrees of variables. Additionally, the summary output indicated a standard error of 0.238 on 242 degrees of error and the plot shows that we should believe the standard errors that the summary provides.

The in-sample loss can be calculated using the squared-error loss on the log scale, $L(Z, \theta) = (\log_{10}(Y) - \mu_\theta(N))^2$.

The following calculation gives us the in-sample loss:

```
in_sample_loss = 0
predictions = as.double(predict(mod1))

for (i in 1:nrow(msadata)){
  in_sample_loss = in_sample_loss + (log(msadata[i,"pcgmp"]) - predicti
ons[i])^2
}
in_sample_loss
```

The in-sample loss is: 13.71174

We will now use 5-Fold Cross Validation to determine out-of sample error:

```
cv.lm <- function(data, formula, nfolds = 5) {
  formula <- sapply(formula, as.formula)
  n <- nrow(data)
  fold.labels <- sample(rep(1:nfolds, length.out = n))
  mses <- matrix(NA, nrow = nfolds, ncol = length(formula))
  colnames <- as.character(formula)
  for (fold in 1:nfolds){
    test.rows <- which(fold.labels == fold)
    train <- data[-test.rows, ]
    test <- data[test.rows, ]
    for (form in 1:length(formula)){
      current.model <- lm(formula = formula[[form]], data = train)
      predictions <- predict(current.model, newdata = test)
      test.responses <- eval(formula[[form]][[2]], envir = test)
      test.errors <- test.responses - predictions
      mses[fold,form] <- mean(test.errors^2)
    }
  }
  return(colMeans(mses))
}

formula = list(log(pcgmp)~log(pop))
cv.lm(msadata,formula)
```

The 5-fold cross validation mean squared error is 0.057.

Now that we have done the calculations for the full sample for population size, we must reduce our sample to an analysis sample in order to account for missing values of our data for the variables that are taken into account with our alternative models. Our analysis sample will omit nan values for the following column vectors: finance, ict, and prof.tech.

The alternative models are:

log(pcgmp) ~ finance log(pcgmp) ~ ict log(pcgmp) ~ finance + ict log(pcgmp) ~ finance + ict + prof.tech

For consistency with our analysis, we will repeat the above calculations for the model with formula: log($pcgmp$) log(pop):

```
analysisSample = finance_ict_prof.tech

mod1 <- lm(log(pcgmp)~log(pop), data = analysisSample)
summary(mod1)
```

According to the summary output above, the coefficents only differ to the thousandth decimal and the difference in residual standard error is very minimal.

```r
mod2 <- lm(log(pcgmp)~finance, data = analysisSample)
summary(mod2)
mod3 <- lm(log(pcgmp)~ict, data = analysisSample)
summary(mod3)
mod4 <- lm(log(pcgmp)~finance + ict, data = analysisSample)
summary(mod4)
mod5 <- lm(log(pcgmp)~finance + ict + prof.tech, data = analysisSample)
summary(mod5)

mod1_in_sample_loss = 0
mod1_predictions = as.double(predict(mod1))
mod2_in_sample_loss = 0
mod2_predictions = as.double(predict(mod2))
mod3_in_sample_loss = 0
mod3_predictions = as.double(predict(mod3))
mod4_in_sample_loss = 0
mod4_predictions = as.double(predict(mod4))
mod5_in_sample_loss = 0
mod5_predictions = as.double(predict(mod5))

for (i in 1:nrow(analysisSample)){
  mod1_in_sample_loss = mod1_in_sample_loss + (log(analysisSample[i,"pc
gmp"]) - mod1_predictions[i])^2
  mod2_in_sample_loss = mod2_in_sample_loss + (log(analysisSample[i,"pc
gmp"]) - mod2_predictions[i])^2
  mod3_in_sample_loss = mod3_in_sample_loss + (log(analysisSample[i,"pc
gmp"]) - mod3_predictions[i])^2
  mod4_in_sample_loss = mod4_in_sample_loss + (log(analysisSample[i,"pc
gmp"]) - mod4_predictions[i])^2
  mod5_in_sample_loss = mod5_in_sample_loss + (log(analysisSample[i,"pc
gmp"]) - mod5_predictions[i])^2
}

in_sample_loss = matrix(c(mod1_in_sample_loss, mod2_in_sample_loss, mod
3_in_sample_loss, mod4_in_sample_loss, mod5_in_sample_loss))
models = matrix(c("log(pcgmp)~log(pop)", "log(pcgmp)~finance", "log(pcg
mp)~ict", "log(pcgmp)~finance+ict", "log(pcgmp)~finance+ict+prof.tech")
)
df = data.frame(models, in_sample_loss)

formula = list(log(pcgmp)~log(pop), log(pcgmp)~finance, log(pcgmp)~ict,
log(pcgmp)~finance+ict, log(pcgmp)~finance+ict+prof.tech)
cv = cv.lm(analysisSample,formula)
df = data.frame(df, cv)

in_sample_lossTbl <- kable(df, col.names = c("Model", "In-Sample Loss",
"5-Fold Cross Validation"))
print(in_sample_lossTbl)
```