36402 Data Exam 1

Joong Ho Choi

TOTAL POINTS

97 / 100

QUESTION 1

1 Introduction and Format 20 / 20

✓ **- 0 pts** **All items are excellent (Contents, Figures & Tables, Questions & Objectives, Final Findings).**

Overall Requirements: Contents

- 5 pts R codes/outputs are included

- 10 pts Not a report. Simply a numbered or bulleted list of responses to the questions.

- 1 pts Includes mathematical derivations

- 1 pts Very divergent format (different line spacing, margins, and/or font size).

Overall Requirements: Format

- 2 pts Unlabeled/mistakenly labeled figures/tables.

- 2 pts Figures/tables unclear.

- 2 pts Figures/tables without captions.

Introduction: Questions & Objectives

- 1 pts All goals mentioned but are vague.

- 2 pts Not all objectives from the instructions are explicitly mentioned upfront

- 4 pts Questions and objectives not addressed

Introduction: Final Findings

- 1 pts Final findings unclear/not sufficient.

- 3 pts Final findings not reported

QUESTION 2

2 Exploratory Data Analysis 15 / 15

✓ **- 0 pts** **Correct**

Data Description

- 0 pts Correct

- 1 pts Missing meaning of relevant variables given with units

Response Variable

- 0 pts Correct

- 3 pts Missing

- 2 pts Missing distribution of the response variable

Univariate EDA

- 0 pts Correct

- 3 pts Missing

- 1 pts Missing interpretation

- 2 pts Missing graphs

- 2 pts Missing results on variables other than PCGMP

- 2 pts Some results (e.g. average, sd, shape) are not connected to the modeling decisions or results

- 1 pts It's better to show the figure "log y has a distribution resembling a normal distribution"

Multivariate EDA

- 0 pts Correct

- 3 pts Missing

- 1 pts It's better to say something here, such as what transformations they might need or what variables appear to be the strongest predictors

- 2 pts No graphs about the predictors and the response variable.

- 0.5 pts Scatterplots are a little small/unclear

- 1 pts Scatterplots may provide more information than correlation plots since we can observe nonlinear associations from scatterplots, which probably would help us build our models.

- 1.5 pts No graphs about the predictors and the response variable other than population

Trends and Feature Identification

- 0 pts Correct

- 3 pts Missing

- 2 pts Wrong analysis on the trend

Relevant Plots

- 1 pts More data visualization needed

- **3 pts** No Plots

QUESTION 3

3 Modeling & Diagnostics **28 / 30**

Variable Selection

✓ **+ 1 pts** **Goal (i).** The power-law-scaling model has $$\log Y$$ as the outcome and have $$\log N$$ as the predictor.

✓ **+ 1 pts** **Goal (iii).** The nonparametric model has $$\log Y$$ as the outcome and the other economic variables (`finance`, `prof.tech`, `ict`, and `management`) as predictors.

✓ **+ 1 pts** **Goal (iv).** The linear model for the residuals uses $$\log N$$ as its covariate.

Model Construction & Explanation

✓ **+ 3 pts** **Goal (i).** The power-law-scaling model is described clearly. (Either a statement of the form, "We fit a linear model of the log of...," or a formula will suffice.)

+ 2 pts **Goal (iii).** A nonparametric model (e.g. a kernel smoother with normal kernel) was fitted and described clearly. (Bandwidths may be chosen by less expensive procedures than cross-validation, but the procedure used should be clearly described.)

✓ **+ 2 pts** **Goal (iv).** A linear model was fitted to the residuals of the nonparametric model and it is described clearly. (In particular, it is clarified that the residuals are regressed on the log of the population.)

- 0.5 pts Minor issue with explanation

Cross Validation

✓ **+ 1 pts** Cross validation was performed on model 1.

✓ **+ 2 pts** Described (or at least reported) the CV errors **and** those measurements are on the original scale, not the log-transformed scale. (Error estimates should be on the order of $$8\times 10^7$$.)

✓ **+ 2 pts** Described (or at least reported) the estimated standard error. (The standard error should be on the order of $$1\times 10^7$$.)

- 0.5 pts CV errors are not summarized in this section, or they are summarized as a list.

- 1 pts Calculated CV errors are far off of their expectation. (See roughly the order of the error and its standard deviation in the above rubric items.)

- 2 pts Other minor error

Model diagnostics

✓ **+ 1.5 pts** Plotted appropriate diagnostic plot(s)

✓ **+ 1.5 pts** Plot has good/okay format (ie one of: properly labeled axis, no code, easily viewable, etc)

+ 0.5 pts comment on/describe the plots [with minor error (ie explicitly said that ____ assumption is true, ie need to explain/describe more)]

✓ **+ 1 pts** comment on/describe the plots

+ 0.5 pts Is there any violation of assumptions? if not, state it and summarize why. If so, suggest improvements if needed/addressed this [with minor error, ie suggested improvement doesn't really help or needs more explanation]

✓ **+ 1 pts** Is there any violation of assumptions? if not, state it and summarize why. if so, suggest improvements if needed/addressed this

+ 0 pts Missing

+ 0 pts Figures should have captions. Or plots have bad axis names (Removed this rubric item because Intro already penalized for missing caption)

Model Comparison

✓ **+ 1.5 pts** Presented the standard errors for the CV errors.

✓ **+ 2.5 pts** Compared the difference between the CV errors against their standard errors.

✓ **+ 1 pts** Explained why you chose this 'best' model.

+ 0 pts Missing

Bootstrap

✓ **+ 1 pts** Proper argument using residual diagnostics. (Note that linear relationship is not necessary for parametric or nonparametric bootstrap by resampling residuals.)

✓ **+ 1 pts** Clearly addressed the assumptions for (at least the chosen) bootstrap methods

✓ **+ 2 pts** Sufficient residual diagnostic plots for the argument.

✓ **+ 1 pts** **Choice of a valid bootstrap method.**

+ 0 pts No credible effort

- 1 pts Pages not selected

- 1 pts Formatting problem (e.g. one long page, font too small)

QUESTION 4

4 Results 20 / 20

CV estimate of Model 1 prediction error and the bias

✓ **- 0 pts** **Correct**

- 2 pts Incorrect/Missing estimated prediction error

- 2 pts Incorrect/Missing estimated bias

- 2 pts Incorrect/Missing comment on the suitability of Model 1

CV estimate of Model 2 prediction error

✓ **- 0 pts** **Correct**

- 2 pts Incorrect/Missing estimated Model 2 prediction error

- 2 pts Incorrect/Missing contrast of Model 1 and 2

Estimate of a scaling exponent and its CI

✓ **- 0 pts** **Correct**

- 2 pts Incorrect/Missing estimate of a scaling exponent

- 2 pts Incorrect/Missing CI for a scaling exponent

- 2 pts Poor/Missing justification for the bootstrap choice (which bootstrap to use and why).

Interpreting a CI for scaling exponent

✓ **- 0 pts** **Correct.**

1. A negative exponent implies that the relationship between `pcgmp` and `pop` appears to be the reverse of what was described in the problem.

2. The CI, however, gets very close to 0 or may overlap 0.

3. Economic variables account for the majority of `pcgmp`, and `pop` is a weak predictor.

- 2 pts Poor/Incorrect interpretation of the relationship between `pcgmp` and `pop`.

- 2 pts Missing/Incorrect statistical (in)significance of estimate for a scaling exponent.

QUESTION 5

5 Conclusions 14 / 15

- 0 pts Correct

Overall Requirements: Writing & Style

✓ **- 0 pts** **Awesome**

- 1 pts Too much stats jargon.

- 3 pts Is hard to understand for non-statisticians

Conclusions: Main Findings

- 0 pts Perfect.

- 2 pts This is not strong support for the power law scaling model.

✓ **- 1 pts** **You should not conclude this proves Urban hierarchy hypothesis.**

- 3 pts Partially missing

- 5 pts Missing

Conclusions: Discussion & Limitations

✓ **- 0 pts** **Excelent.**

- 2 pts You should mention that this was not a causal analysis.

- 2 pts not enough discussion on limitations

- 5 pts Missing

- 1 pts Do not write out a list!

- 1 pts No labels for answers

- 15 pts Missing

36-402 DA Exam 1

Joong Ho Choi (joonghoc)

March 25, 2022

1 Introduction

For cities to improve their economies and subsequently their standards of living, it is essential they understand what drives their economic growths. urban hierarchy hypothesis claims economically productive companies tend to be in large cities, and they are responsible for cities' large GMPs. On the other hand, power law scaling hypothesis states that larger cities are more economically productive than small ones due to larger production capabilities brought through larger population. Is city's economic growth driven by simply having larger population or having businesses with skilled workers? This knowledge is also paramount to companies in general, as they can use this knowledge to benefit from requesting tax breaks when they move into new cities. Our client is one of those companies; our client is a huge technology company with thousands of highly skilled employees. The company is considering moving its headquarters to a large city. If the urban hierarchy hypothesis is true, they can claim that moving to the city will improve its economy, and so the city should give them huge tax breaks to convince them to move there. To enjoy this benefit, our client is interested in testing which of the two hypothesis is true. Given a dataset that contains information about the economies of 133 cities in the United States, we will investigate if the urban hierarchy hypothesis is true or not **(1)** and if power law scaling hypothesis is true or not **(2)**.

Our analysis led us to the conclusion that gross metropolitan product(GMP), which is the total economic output of a metropolitan statistical area, increases with more high-value businesses. Basically, the urban hierarchy hypothesis is true. On the other hand, with the given dataset, we found power law scaling hypothesis to be false. After accounting for the economic variables, we saw that log(pcgmp) and log(population) are negatively correlated. Considering how log transformation is monotonic transformation, we conclude

that larger cities being more economically productive than small ones does not result from being larger. TIn conclusion, the economies of cities improve with size, because they attract more business that rely on highly skilled workers to provide expensive services.

2 Exploratory Data Analysis

Our dataset contains data about the economies of 133 metropolitan statistical areas, which include cities and the areas surrounding them, in the United States. Relevant variables for our investigation are the followings:

-MSA: Name of the metropolitan statistical area

-pcgmp:The per-capita gross metropolitan product Y, in dollars

-pop: The MSA's population

-finance: The share of the MSA's economy that is in the financial industry (as a fraction between 0 and 1)

-prof.tech: The share of the MSA's economy that is in professional and technical services

-ict: The share of the MSA's economy that is in information, communication, and technology

-management: The share of the MSA's economy that is in corporate management

We began by examining the univariate distributions of the key variables(population as predictor and pcgmp as response) for power-law hypothesis. While pcgmp ranged from 15040 to 77260 with the mean of 32003.31, the population ranged from 54980 to 4487000 with the mean of 450485.60. Visually, it was clear to see that the histograms for both variables seem to be skewed to the right. Thus, to make them less skewed and give them more normality, we applied log transformation on the two variables.The desired effect was achieved, as the distribution became somewhat more symmetrical than before.

1 Introduction and Format **20 / 20**

✓ **- 0 pts All items are excellent (Contents, Figures & Tables, Questions & Objectives, Final Findings).**

Overall Requirements: Contents

- **- 5 pts** R codes/outputs are included
- **- 10 pts** Not a report. Simply a numbered or bulleted list of responses to the questions.
- **- 1 pts** Includes mathematical derivations
- **- 1 pts** Very divergent format (different line spacing, margins, and/or font size).

Overall Requirements: Format

- **- 2 pts** Unlabeled/mistakenly labeled figures/tables.
- **- 2 pts** Figures/tables unclear.
- **- 2 pts** Figures/tables without captions.

Introduction: Questions & Objectives

- **- 1 pts** All goals mentioned but are vague.
- **- 2 pts** Not all objectives from the instructions are explicitly mentioned upfront
- **- 4 pts** Questions and objectives not addressed

Introduction: Final Findings

- **- 1 pts** Final findings unclear/not sufficient.
- **- 3 pts** Final findings not reported

that larger cities being more economically productive than small ones does not result from being larger. TIn conclusion, the economies of cities improve with size, because they attract more business that rely on highly skilled workers to provide expensive services.

2 Exploratory Data Analysis

Our dataset contains data about the economies of 133 metropolitan statistical areas, which include cities and the areas surrounding them, in the United States. Relevant variables for our investigation are the followings:

-MSA: Name of the metropolitan statistical area

-pcgmp:The per-capita gross metropolitan product Y, in dollars

-pop: The MSA's population

-finance: The share of the MSA's economy that is in the financial industry (as a fraction between 0 and 1)

-prof.tech: The share of the MSA's economy that is in professional and technical services

-ict: The share of the MSA's economy that is in information, communication, and technology

-management: The share of the MSA's economy that is in corporate management

We began by examining the univariate distributions of the key variables(population as predictor and pcgmp as response) for power-law hypothesis. While pcgmp ranged from 15040 to 77260 with the mean of 32003.31, the population ranged from 54980 to 4487000 with the mean of 450485.60. Visually, it was clear to see that the histograms for both variables seem to be skewed to the right. Thus, to make them less skewed and give them more normality, we applied log transformation on the two variables.The desired effect was achieved, as the distribution became somewhat more symmetrical than before.

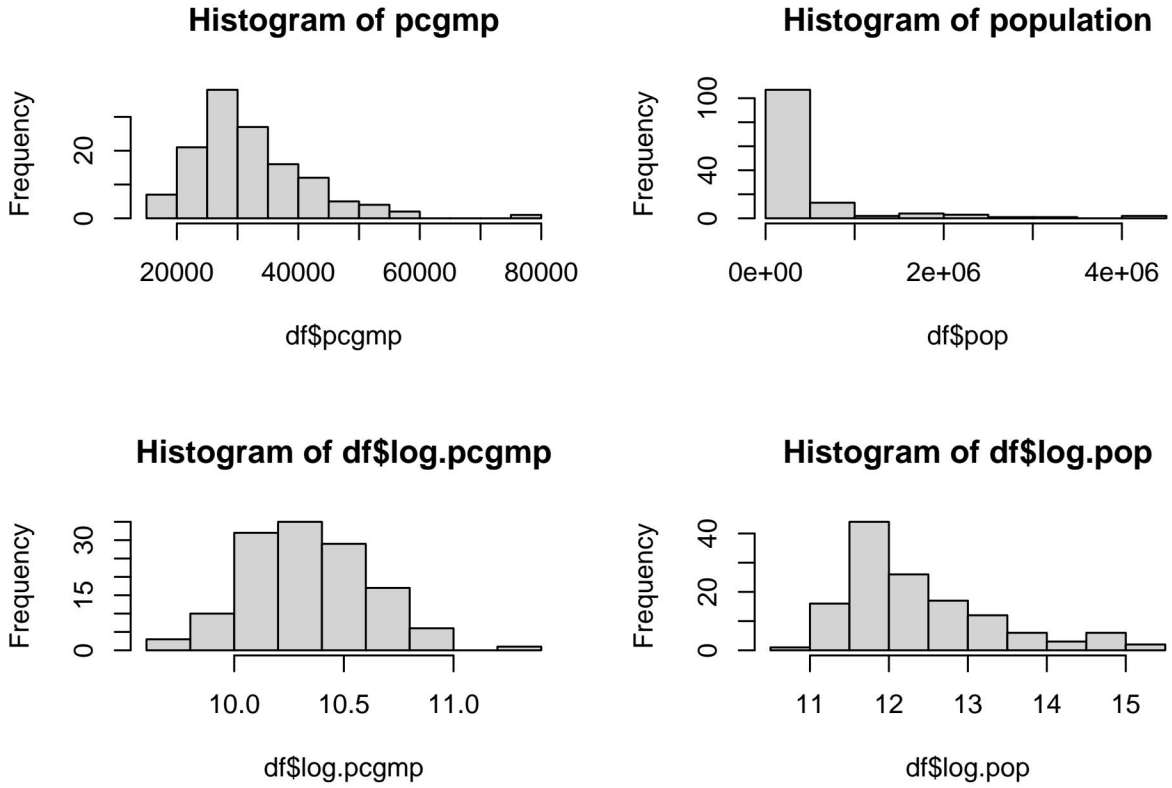

Figure 1: Histograms showing the distributions of variables pcgmp and population before and after log transformation

One thing that caught our attention is that 120 out of 133 data points have population of less than 1000000. This would be important to remember as we develop our models and make predictions/assumptions , since we do not have sucient data on metropolitan statistical area with population more than 1000000.

Then, for multivariate visualization, we created a pairwise plot to visualize the relationships between relevant continuous variables. The scatterplots between pcgmp and population reinforced the idea that log transformation on these two variables would be good.

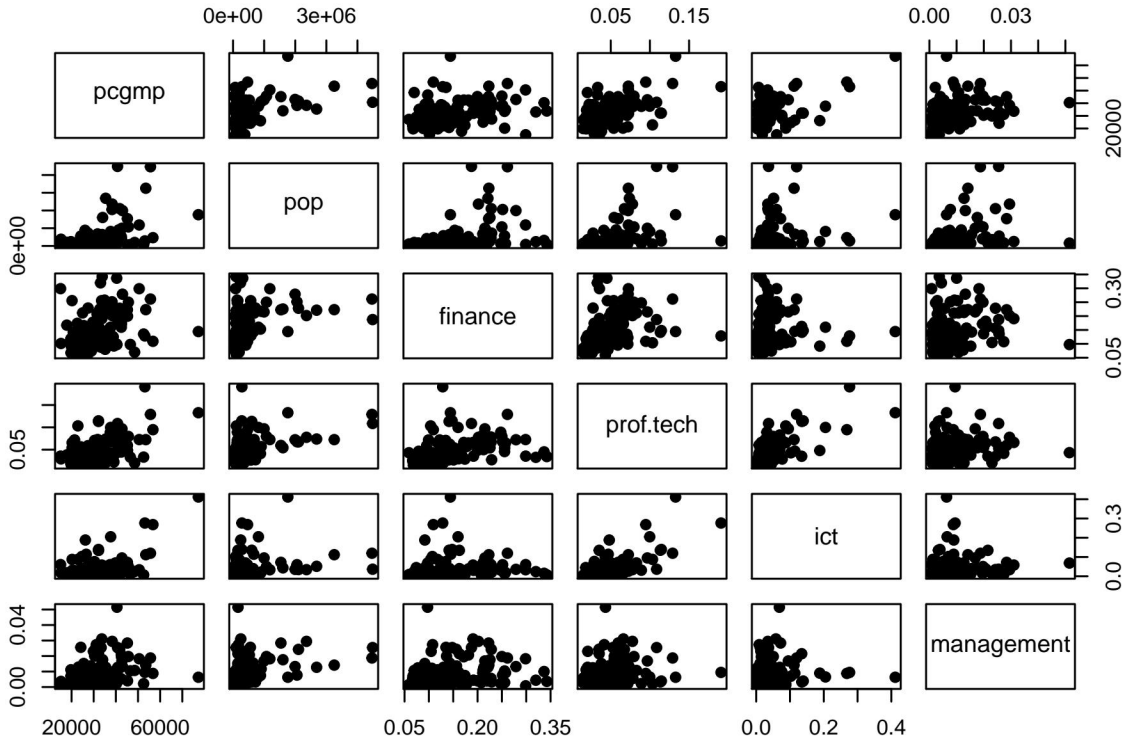

Figure 2: Scatterplot matrix showing relationships between relevant continuous variables

Interestingly, while finance, prof.tech and ict had somewhat clear positive correlationship with pcgmp, management was the only economic variable that had unclear relationship with the pcgmp. This perhaps suggests that management might not be helpful in predicting pcgmp and it is worth noting for future investigations.

Modeling & Diagnostics

For our model 1, we constructed a linear model to test the power-law hypothesis that our client asked us. To be specific, we fitted the power-law-scaling model to relate pcgmp to population size, using the appropriate log transformations to pcgmp and pop. **2**

Linear: log(pcgmp)~$beta_0$+$beta_1$*log(pop)

After establishing model 1, we constructed a flexible kernel smoother to predict pcgmp from the four economic variables for our model 2, in order to test urban hierarchy hypothesis. Again, we had pcgmp log-transformed. The relationships between the four economic variables and log.pcgmp are shown below. **1**

2 Exploratory Data Analysis 15 / 15

✓ **- 0 pts Correct**

Data Description

- 0 pts Correct

- 1 pts Missing meaning of relevant variables given with units

Response Variable

- 0 pts Correct

- 3 pts Missing

- 2 pts Missing distribution of the response variable

Univariate EDA

- 0 pts Correct

- 3 pts Missing

- 1 pts Missing interpretation

- 2 pts Missing graphs

- 2 pts Missing results on variables other than PCGMP

- 2 pts Some results (e.g. average, sd, shape) are not connected to the modeling decisions or results

- 1 pts It's better to show the figure "log y has a distribution resembling a normal distribution"

Multivariate EDA

- 0 pts Correct

- 3 pts Missing

- 1 pts It's better to say something here, such as what transformations they might need or what variables appear to be the strongest predictors

- 2 pts No graphs about the predictors and the response variable.

- 0.5 pts Scatterplots are a little small/unclear

- 1 pts Scatterplots may provide more information than correlation plots since we can observe nonlinear associations from scatterplots, which probably would help us build our models.

- 1.5 pts No graphs about the predictors and the response variable other than population

Trends and Feature Identification

- 0 pts Correct

- 3 pts Missing

- 2 pts Wrong analysis on the trend

Relevant Plots

- 1 pts More data visualization needed

- 3 pts No Plots

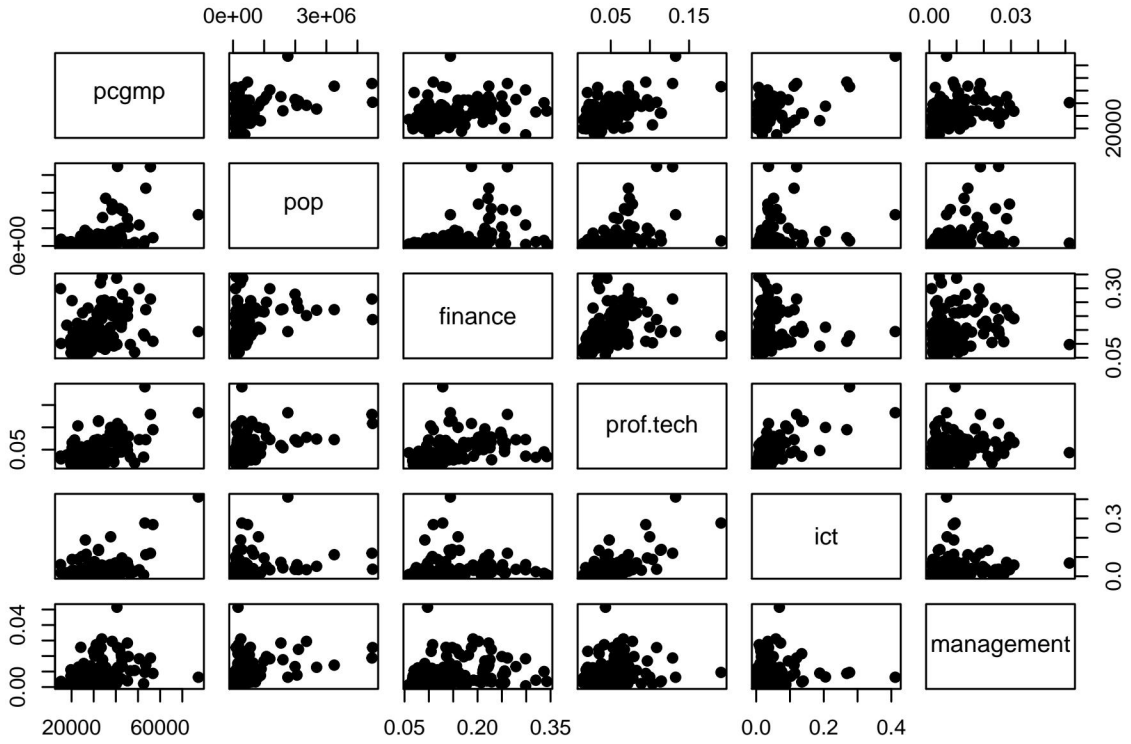

Figure 2: Scatterplot matrix showing relationships between relevant continuous variables

Interestingly, while finance, prof.tech and ict had somewhat clear positive correlationship with pcgmp, management was the only economic variable that had unclear relationship with the pcgmp. This perhaps suggests that management might not be helpful in predicting pcgmp and it is worth noting for future investigations.

Modeling & Diagnostics

For our model 1, we constructed a linear model to test the power-law hypothesis that our client asked us. To be specific, we fitted the power-law-scaling model to relate pcgmp to population size, using the appropriate log transformations to pcgmp and pop. **2**

Linear: log(pcgmp)~$beta_0$+$beta_1$*log(pop)

After establishing model 1, we constructed a flexible kernel smoother to predict pcgmp from the four economic variables for our model 2, in order to test urban hierarchy hypothesis. Again, we had pcgmp log-transformed. The relationships between the four economic variables and log.pcgmp are shown below. **1**

From our model diagnostics below, we see some violations of the model assumptions for both models.

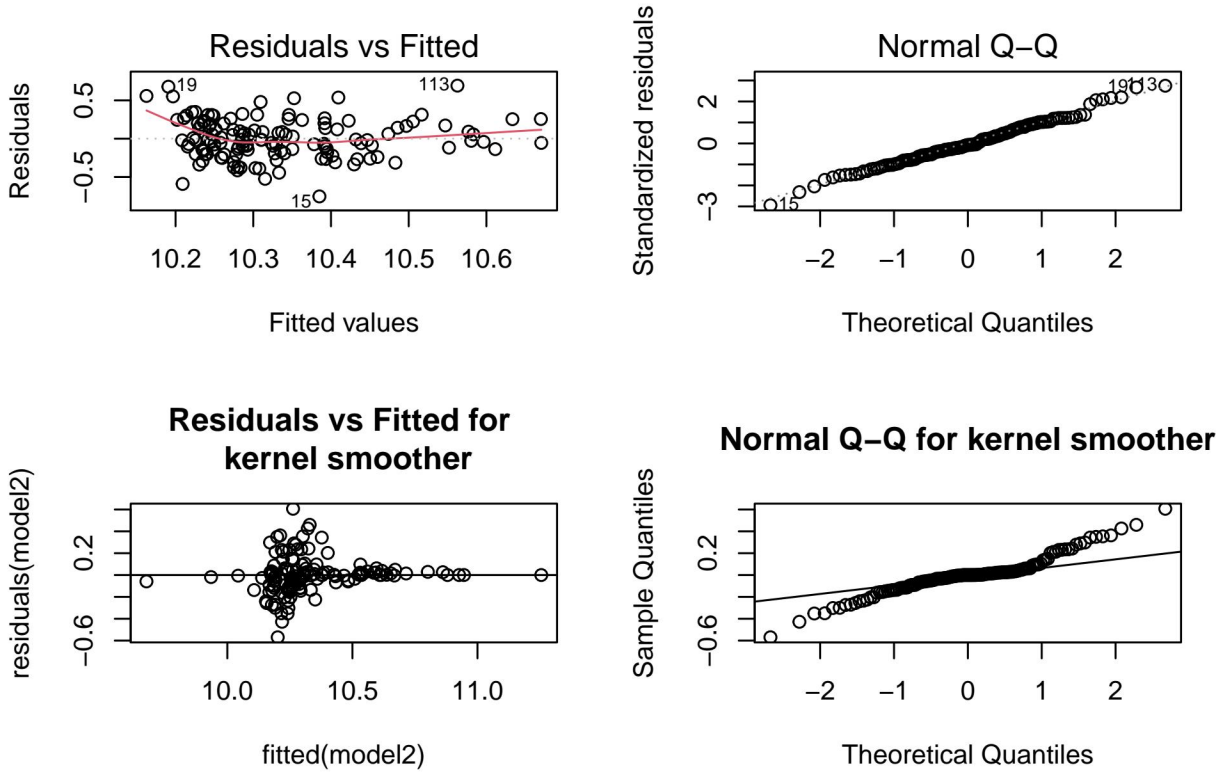

Figure 3: Diagnostic plots analyzing fitted values versus respective residuals and QQ-plots for models 1 and 2

For both models' residuals plots, the residuals are not equally spread out, so we know both violate the homoscedasticity assumption. Heteroscedasticity is a problem because ordinary least squares (OLS) regression assumes that all residuals are drawn from a population that has a constant variance. This might affect the confidence intervals of the two models' predictions to be less reliable. Moreover, see violation of normality in normal quantile-quantile(QQ) plots.While residuals deviate from the dotted line by slight amount at the left and the right tails for model 1, the residuals deviate by great amount for model 2. The fact that the residuals are not completely normal does slightly undermine the validity of each model's ability to perform inference. Although non-normality is ok when sample size is large, it might not be applicable to us, given that our dataset is size of 133.

We then performed a 10-fold cross-validation to estimate the prediction errors of both models in order to determine which model fits the data given that we were not able to

come to a clear conclusion just from the diagnosticis and goodness-of-fit. Below in table 1, we see the results of our cross validation. Our kernel smoother had lower mean squared error(MSE) by 74382819 and lower standard error by 14900998.**1**

Table 1: Estimated prediction error for each model and standard error based on 10-fold cross validation

Results of CV	Linear Model	Kernel Smoother
Estimated MSE	74422791	39971.97
Standard Error	14902620	1621.966

Given the results of our cross-validation along with the fact that both models had very similar model diagnostics and nearly identical residuals, we decided that the kernel smoother would be better predicting pcgmp.**(1)**

Our next interest would be to evaluate whether population size matters after the economic variables are accounted for. However, before proceeding to make a model 3, we need to consider one problem with our power-law-scaling models. If our goal is to predict pcgmp, using a log transformation lets us use linear regression to estimate the power law, but it is biased. Hence, it is worth investigating the quantifiable value of the bias present. We decided to quantify the bias in our model's estimates of pcgmp for cities the size of Pittsburgh, which has population size of 2361000 and is in between 50% and 75% percentiles of population range.

Our estimates of the population mean are less and less variable with increasing sample size, and they converge towards the true population value. For small samples, the typical sample mean tends to underestimate the true population value. However, regradlessness of the skewness of the sampling distribution with small n, the average of the 1000 simulations/experiments is very close to the population value, for all sample sizes

From the diagnostic earlier on in Fig 3, we know the residuals are not completely normal; we do not know the distribution of the residuals. For bootstrap, we nonparametrically resampled rows, because the only assumption this approach makes is that the residuals are independent.Thus, we performed 1000 simulations, in which we bootstrapped 133 points to train a new estimator on and predict pcgmp for a city the size of Pittsburgh. We gathered 1000 prediction values, got the average and subtracted the Pittsburgh's pcgmp from the average to quantify the bias.

Pittsburgh has pcgmp of 38350,whereas the average prediction of Pittsburgh's pcgmp from estimators modelled on bootstrapped data was 39971.97. The difference between

the mean of the mean estimates and the population value is called bias. Thus, 1621.966 is the bias.

Now that we have quantified the bias present in our model, we constructed our 3rd model, a a linear regression of the *residuals* of the previous model 2 and population, with appropriate log transformation. To be specific, we applied log transformations to pcgmp and population. This is essentially equivalent to fitting the power-law model after accounting for the economic variables. For this model, it is worth noting that the p-value of log(pop) in model 3 was 0.119; as it was much greater than 0.05, the covariate was not considered as statistically significant. **(2)**

Results

Overall, between our linear model and kernel smoother, we observed better prediction performance from our non-parametric method that uses other economic variables to predict instead of using population. As shown in table 1, while the linear model(model 1) had the estimated MSE of 74422791 after 10-fold cross validation, the kernel smoother had the estimated MSE of 39971.97 after 10-fold cross validation. Our kernel smoother had lower MSE by 74382819 and lower standard error by 14900998. Considering these stark contrasts,it is likely that economic variables have more useful data information for prediction than population alone. In addition, through results of our bootstrap analysis, we determined the bias of model to be 1101.723. Therefore, we do not feel confident with model 1's suitability for predicting pcgmp and feel more confident regarding the model 2's suitability for predicting pcgmp. **(1)**

Then, to evaluate whether population size matters after the economic variables are accounted for, we fitted a linear model to the residuals of the economic model, using log N as a covariate. Using the coefficient from the summary of this model 3, we report -0.02387579 as an estimate of the scaling exponent. Then, considering how the residual plots earlier on seem to show that residuals depend on X, we used the bootstrap method of resampling whole cases to get good estimates of the uncertainty. As a result, we derived [-0.07652439 -0.01781071] to be the 95% confidence interval for the scaling exponent.After accounting for the economic variables, we see that log(pcgmp) and log(population) are negatively correlated. In addition, the interval does not contain zero, so this finding is statistically significant. This shows higher population size negatively affects pcgmp after the economic variables are accounted for. **(2)**

3 Modeling & Diagnostics **28 / 30**

Variable Selection

✓ **+ 1 pts** **Goal (i).** The power-law-scaling model has $$\log Y$$ as the outcome and have $$\log N$$ as the predictor.

✓ **+ 1 pts** **Goal (iii).** The nonparametric model has $$\log Y$$ as the outcome and the other economic variables (`finance`, `prof.tech`, `ict`, and `management`) as predictors.

✓ **+ 1 pts** **Goal (iv).** The linear model for the residuals uses $$\log N$$ as its covariate.

Model Construction & Explanation

✓ **+ 3 pts** **Goal (i).** The power-law-scaling model is described clearly. (Either a statement of the form, "We fit a linear model of the log of...," or a formula will suffice.)

+ 2 pts **Goal (iii).** A nonparametric model (e.g. a kernel smoother with normal kernel) was fitted and described clearly. (Bandwidths may be chosen by less expensive procedures than cross-validation, but the procedure used should be clearly described.)

✓ **+ 2 pts** **Goal (iv).** A linear model was fitted to the residuals of the nonparametric model and it is described clearly. (In particular, it is clarified that the residuals are regressed on the log of the population.)

- 0.5 pts Minor issue with explanation

Cross Validation

✓ **+ 1 pts** Cross validation was performed on model 1.

✓ **+ 2 pts** Described (or at least reported) the CV errors **and** those measurements are on the original scale, not the log-transformed scale. (Error estimates should be on the order of $$8\times 10^7$$.)

✓ **+ 2 pts** Described (or at least reported) the estimated standard error. (The standard error should be on the order of $$1\times 10^7$$.)

- 0.5 pts CV errors are not summarized in this section, or they are summarized as a list.

- 1 pts Calculated CV errors are far off of their expectation. (See roughly the order of the error and its standard deviation in the above rubric items.)

- 2 pts Other minor error

Model diagnostics

✓ **+ 1.5 pts** Plotted appropriate diagnostic plot(s)

✓ **+ 1.5 pts** Plot has good/okay format (ie one of: properly labeled axis, no code, easily viewable, etc)

+ 0.5 pts comment on/describe the plots [with minor error (ie explicitly said that ____ assumption is true, ie need to explain/describe more)]

✓ **+ 1 pts** comment on/describe the plots

+ 0.5 pts Is there any violation of assumptions? if not, state it and summarize why. If so, suggest improvements if needed/addressed this [with minor error, ie suggested improvement doesn't really help or needs more explanation]

✓ **+ 1 pts** Is there any violation of assumptions? if not, state it and summarize why. if so, suggest improvements if needed/addressed this

+ 0 pts Missing

+ 0 pts Figures should have captions. Or plots have bad axis names (Removed this rubric item because Intro already penalized for missing caption)

Model Comparison

✓ **+ 1.5 pts** **Presented the standard errors for the CV errors.**

✓ **+ 2.5 pts** **Compared the difference between the CV errors against their standard errors.**

✓ **+ 1 pts** **Explained why you chose this 'best' model.**

+ 0 pts Missing

Bootstrap

✓ **+ 1 pts** **Proper argument using residual diagnostics. (Note that linear relationship is not necessary for parametric or nonparametric bootstrap by resampling residuals.)**

✓ **+ 1 pts** **Clearly addressed the assumptions for (at least the chosen) bootstrap methods**

✓ **+ 2 pts** **Sufficient residual diagnostic plots for the argument.**

✓ **+ 1 pts** **Choice of a valid bootstrap method.**

+ 0 pts No credible effort

- 1 pts Pages not selected

- 1 pts Formatting problem (e.g. one long page, font too small)

the mean of the mean estimates and the population value is called bias. Thus, 1621.966 is the bias.

Now that we have quantified the bias present in our model, we constructed our 3rd model, a a linear regression of the *residuals* of the previous model 2 and population, with appropriate log transformation. To be specific, we applied log transformations to pcgmp and population. This is essentially equivalent to fitting the power-law model after accounting for the economic variables. For this model, it is worth noting that the p-value of log(pop) in model 3 was 0.119; as it was much greater than 0.05, the covariate was not considered as statistically significant. **(2)**

Results

Overall, between our linear model and kernel smoother, we observed better prediction performance from our non-parametric method that uses other economic variables to predict instead of using population. As shown in table 1, while the linear model(model 1) had the estimated MSE of 74422791 after 10-fold cross validation, the kernel smoother had the estimated MSE of 39971.97 after 10-fold cross validation. Our kernel smoother had lower MSE by 74382819 and lower standard error by 14900998. Considering these stark contrasts,it is likely that economic variables have more useful data information for prediction than population alone. In addition, through results of our bootstrap analysis, we determined the bias of model to be 1101.723. Therefore, we do not feel confident with model 1's suitability for predicting pcgmp and feel more confident regarding the model 2's suitability for predicting pcgmp. **(1)**

Then, to evaluate whether population size matters after the economic variables are accounted for, we fitted a linear model to the residuals of the economic model, using log N as a covariate. Using the coefficient from the summary of this model 3, we report -0.02387579 as an estimate of the scaling exponent. Then, considering how the residual plots earlier on seem to show that residuals depend on X, we used the bootstrap method of resampling whole cases to get good estimates of the uncertainty. As a result, we derived [-0.07652439 -0.01781071] to be the 95% confidence interval for the scaling exponent.After accounting for the economic variables, we see that log(pcgmp) and log(population) are negatively correlated. In addition, the interval does not contain zero, so this finding is statistically significant. This shows higher population size negatively affects pcgmp after the economic variables are accounted for. **(2)**

4 Results 20 / 20

CV estimate of Model 1 prediction error and the bias

✓ **- 0 pts** **Correct**

- 2 pts Incorrect/Missing estimated prediction error

- 2 pts Incorrect/Missing estimated bias

- 2 pts Incorrect/Missing comment on the suitability of Model 1

CV estimate of Model 2 prediction error

✓ **- 0 pts** **Correct**

- 2 pts Incorrect/Missing estimated Model 2 prediction error

- 2 pts Incorrect/Missing contrast of Model 1 and 2

Estimate of a scaling exponent and its CI

✓ **- 0 pts** **Correct**

- 2 pts Incorrect/Missing estimate of a scaling exponent

- 2 pts Incorrect/Missing CI for a scaling exponent

- 2 pts Poor/Missing justification for the bootstrap choice (which bootstrap to use and why).

Interpreting a CI for scaling exponent

✓ **- 0 pts** **Correct.**

1. A negative exponent implies that the relationship between `pcgmp` and `pop` appears to be the reverse of what was described in the problem.

2. The CI, however, gets very close to 0 or may overlap 0.

3. Economic variables account for the majority of `pcgmp`, and `pop` is a weak predictor.

- 2 pts Poor/Incorrect interpretation of the relationship between `pcgmp` and `pop`.

- 2 pts Missing/Incorrect statistical (in)significance of estimate for a scaling exponent.

Conclusions

Through the findings from models 2 to 3, we can conclude the urban hierarchy hypothesis is true, meaning that economically productive companies in large cities are responsible for their large GMPs **(1)** .On the other hand, findings from models 1-3 show that power law scaling hypothesis is false, meaning larger cities' greater pcgmp do not result from large cities' ability to support more specialized businesses. **(2)**. Thus, the client can claim that the company moving to the city will improve the economy, and request the city to give them huge tax breaks to convince them to move there.

However, the client should be cautious when relying upon these findings to make accurate predictions for cities not present in the data due to several limitations. Firstly, the dataset's size is 133; in the grand scheme of things, the data's size is very small. From our EDA in Fig 1, we discovered how 120 out of 133 data points have population of less than 1000000; with this in mind, we should be cautious of using our models to predict the pcgmp of metropolitan statistical area with population more than 1000000. We should collect more data on cities not included in the dataset, hopefully those that have population of more than 1000000. Secondly, only four sectors(financial, professional and technical, information and communication and technology, corporate management) have been chosen for economic variables. For research purpose, it would be worth considering other sectors such as biotech and art. Lastly, there might be other confounding variables that we have not accounted for, which play pivotal roles in driving cities' economic growth. We should examine different research papers done on cities' economic growth and attempt to account for the influential variables covered in those papers.

5 Conclusions 14 / 15

- **0 pts** Correct

Overall Requirements: Writing & Style

✓ **- 0 pts Awesome**

- **1 pts** Too much stats jargon.
- **3 pts** Is hard to understand for non-statisticians

Conclusions: Main Findings

- **0 pts** Perfect.
- **2 pts** This is not strong support for the power law scaling model.

✓ **- 1 pts You should not conclude this proves Urban hierarchy hypothesis.**

- **3 pts** Partially missing
- **5 pts** Missing

Conclusions: Discussion & Limitations

✓ **- 0 pts Excelent.**

- **2 pts** You should mention that this was not a causal analysis.
- **2 pts** not enough discussion on limitations
- **5 pts** Missing

- **1 pts** Do not write out a list!
- **1 pts** No labels for answers
- **15 pts** Missing