Association between a capital and its counties overall pollution levels

18011724, UFMFNA-30-2 - Statistical modelling 19sep_1

Abstract

A countries pollution level can be predicted by the capitals pollution level based off the data provided. There are some draw back to the study's validity discussed in section 4, but if these problems are managed then the findings here show that for an increase in a capitals pollution level the wider country increases by 0.838 of a level. The predicted pollution level for china based off Beijings level of 93 is 86.973, varying from 72.553 to 101.393 within a 95% confidence interval.

Contents

1 Introduction

This report will explore the association between several countries capital pollution level and its corresponding nation. This study is being done because the researcher believes that a recorded pollution level in the capital city would broadly predict the nations pollution level. The statistical hypothesis are as follows:

s_0: A countries pollution level cannot be used to predict the nations wider pollution level.

s_1: A countries pollution level can be used to predict the nations wider pollution level.

The aim of this report is to determine which of these statements is most accurate taking the data into consideration. Furthermore, a prediction for Chinas overall pollution level using the pollution level recorded in Beijing; this will be done at the end of the report.

2 Exploratory Data Analysis

The data set consists of paired data for a countries capital and the wider nations pollution levels. There are no missing values therefor it easier to analyse the data set. The data set is arranged in columns, with the first holding the pollution level for the country and the other holding the capital’s.

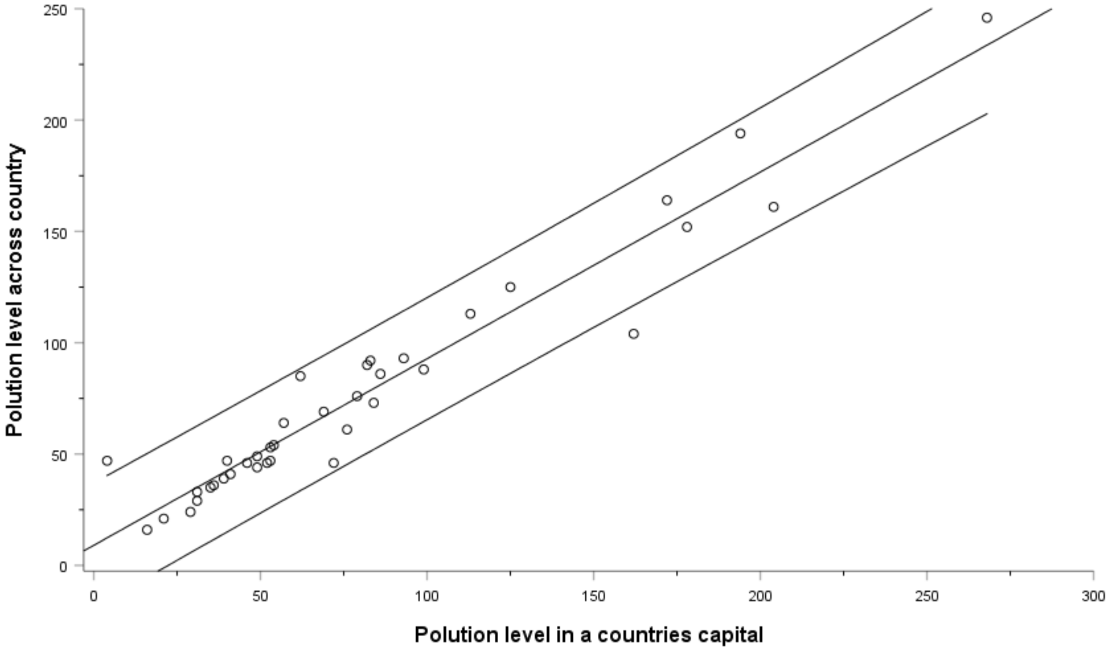

Figure 1. A scatter plot showing the capitals pollution level and its corresponding nation-wide pollution level.

3 Linear Model

Generating a linear regression model gives a way of predicting a countries pollution level from the capitals.

$$country\ pollution\ level = capital\ pollution\ level * 0.838 + 9.039$$

The SPSS output for the linear model includes analysis of the coefficients of the formula above.

Table 1: Regression model output

	Unstandardized Coefficients				95.0% Confidence Interval	
	B	Std. Error	t	Sig	Lower Bound	Upper Bound
(Constant)	9.039	3.671	2.462	.019	1.594	16.484
Capital	.838	.037	22.576	.000	.763	.913

This shows that the calculated coefficients of the linear model are below the 5% significance level, and thus there is strong evidence to accept them. The 95% confidence interval shows that as the pollution level increases, the variability from the line of best fit stays the same. This is known because the constant has a far wider confidence interval proportionally to the gradient of the line. This implies that the variability of the data does not increase with higher values. If the variability in values where swapped, such that the gradient varies a lot but the constant did not, it would show that as the pollution level increased the variation between that and the nationwide pollution level would increase as well. A graph displaying this would have a cone-like shape as the capital’s pollution level increases. This is shown in figure 1, where the nearly parallel lines indicate constant variability regardless of the capital’s pollution level, meaning that if a correlation is found there are unlikely to be outliers where the capital has a high level of pollution. In summary, the deviation from the capitals pollution level is constant, regardless of its pollution level.

4 Study Design

4.1 Internal validity

The problem with the data collection is how it's collected or calculated, it could range from taking 1000's of samples that are equally apart using a specialised piece of equipment, to ranking it out of 100 in air quality by repetitively breathing though the nose as the researcher walked around the city jotting down the score. As for the samples taken in the wider country, they could be from an area next to a factory in the middle of nowhere, or from the middle of a jungle. In the city, samples could be taken from next to the industrial district or at the top of a skyscraper, far away from the dense pollution particulates below. If these factors are not controlled, then the pollution level could be artificially increased or lowered to fit a possible external agenda. It's not known if there are repeated measures across the city and country. This would give little confidence when using the data, unless more information was received about how it was collected, as well as the general study design.

4.2 External validity

The data given is collected only from Africa, which raises many problems for external validity. For example, the population density in Africa is 44 per Km^2 [1], which is far from that of Chinas a population density of 153 per Km^2 [2]. If we presume that greater population density could correlate to more pollution, this would be an important factor to highlight when applying findings from Africa's data to China's population. This therefore implies that the findings would have to be extrapolated if applied elsewhere, making it unclear as to how generalisable the findings are to other countries. In order to amend this problem, a wider sample of countries should be used to in order to introduce variation in data, which would allow the findings to become applicable to more countries.

Additionally, the different possible sources of pollution emit different levels and types of pollution, an example being the burning of coal is more harmful than noise pollution which is what could be being studied. Moreover, mining minerals doesn't occur in a capital city, but manufacturing goods does, which would affect the data because if the county did little manufacturing but lots of mining the country pollution levels may be affected more than the city centres. So, if you took the findings here and applied them to China, which manufactures a huge number of items, the pollution levels may differ.

5 Testing the Statistical Hypothesis

The statistical hypothesis are as follows:

s_0: A countries pollution level cannot be used to predict the nations wider pollution level.

s_1: A countries pollution level can be used to predict the nations wider pollution level.

$P_{co,n}$ = the nth countries pollution level

$P_{ca,n}$ = the nth capitals pollution level

H_0: B = 0

H_1: B ≠ 0

If B = 0 then there is no correlation between capital and country pollution levels, because it would mean that regardless of any change in the capital pollution level the predicted country-wide level would not change. Table 1 shows that B > 0. This shows that the capital pollution level correlates to the wider country.

Table 2: Statistical test for correlation

	Sum of Squares	df	Mean Square	F	Sig.
Regression	90476.224	1	90476.224	509.665	.000
Residual	6390.749	36	177.521		
Total	96866.974	37			

To determine the confidence in the correlation a statistical test is run by SPSS to asses this. Looking at the F statistic and the significance level there is extremely strong evidence to reject the null hypothesis and show that there is a correlation between a capital and the wider countries pollution levels. To show this visually, Figure 2 is the probability distribution for the test done above, the F value of 509.665 is so far away from the mean that its impossible to graphically represent its value.

There is very strong evidence to reject H_0 at the 0.1% significance level showing that there is a strong correlation between the capital pollution level and the wider country.

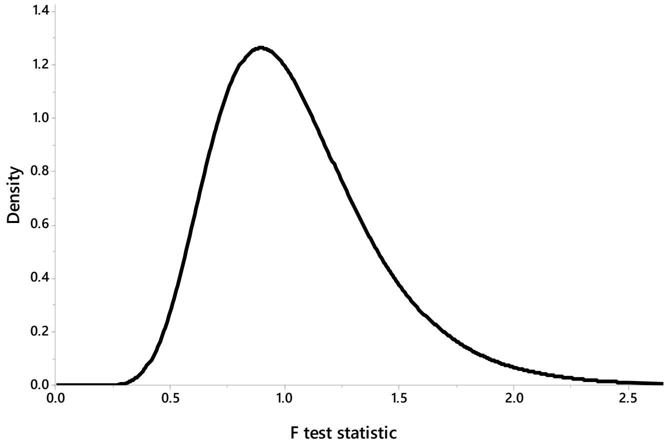

Figure 2: F distribution plot, df1=36, df2=37

6 China Prediction

So using the linear model from the Africa data set, and applying that to the capital pollution level of Beijing:

$$Chinas\ polution\ level = Beijings\ pollution\ level * 0.838 + 9.039$$
$$Chinas\ polution\ level = 93 * 0.838 + 9.039$$
$$Chinas\ polution\ level = 86.973$$

If a 95% confidence interval is applied then the true value could vary from 72.553 to 101.393; this is quite a large range so going out to find the true value may be more suitable, unless that range is of acceptable width.

7 Conclusion

A countries pollution level can be predicted by the capitals pollution level based off the data provided. There are some draw backs to the study's validity discussed previously in section 4, but if these problems are managed then the findings here show that, for an increase in a capitals pollution level, the wider country increases by 0.838 of a level. It isn't known if this is correlation or causality because in a small country the capital pollution level may directly affect its surrounding country but in Russia, for example, the pollution level in Moscow has almost no effect on the pollution level on the whole country.

8 References

[1] - worldometers https://www.worldometers.info/world-population/africa-population/ [accessed 12/12/19]

[2] - worldometers https://www.worldometers.info/world-population/china-population/ [accessed 12/12/19]