Is there a correlation between the air pollution of a country and its number of threatened mammal species?

Ananth Abbaraju

Introduction

This paper is intended to study if there is an association between the air pollution of countries and their number of threatened mammal species. In the past few decades, there has been an alarming spike in the number of endangered species worldwide. In 1996, the International Union for Conservation of Nature listed 1102 animal species as endangered, and since then, this number has increased more than ten-fold. In 2019, 13868 animal species were listed as endangered. Also alarming has been the rise in worldwide pollution due to human activities, including the burning of fossil fuels, the use of synthetic pesticides and fertilizers in agriculture, and the large, ever-increasing number of industries and automobiles. According to the Environmental Protection Agency, fine particulate matter pollution across just the U.S. has increased by a shocking 5.5% from 2016 to 2018. And according to the World Bank Group, in 2017, more than 90% of the world population was exposed to unhealthy air(classified as an annual average $PM_{2.5}$ concentration above 10 $\mu g/m^3$). The increasing seriousness of both these global issues presents an urgent reason to investigate the possible relationship between human-caused environmental issues, such as pollution, and the wide range of connections that they have, such as with the number of threatened mammal species. It is to be hoped that this study will provide strong evidence that will highlight the vulnerability of many aspects of the environment to human activities and the interconnectedness of each of those vulnerabilities by concluding a relationship between an environmental problem known to be caused by humans and another which is not surely proven to be caused by human activities.

To collect data for this study, each country in the world was given a label from 1 to 195, in alphabetical order. Using the TI-84 Plus, 19 unique random numbers were generated, with a seed of 17. For each of the 19 numbers, the mean annual fine particulate matter pollution($\mu g/m^3$ of $PM_{2.5}$), a common measure of air pollution, and the number of threatened mammal species for the respective countries were collected into our data set using reliable sources(see *Works Cited*).

The procedures used in this study will include linear regression analysis, residuals, and inferencing. First, we will create a scatter plot based on the randomly collected data, where each point represents a country's data. Then a linear regression model will be created, and a residual plot will be used to confirm the validity of a linear model. If it is found that a linear model would not be appropriate, other models of regression will be used, and the most appropriate model will be used instead. If a linear model was found to be appropriate, a linear regression significance test will be conducted in hopes of gaining convincing evidence that there is a positive predictive relationship between the mean annual fine particulate matter pollution and the number of threatened mammal species of all countries.

Data Collection

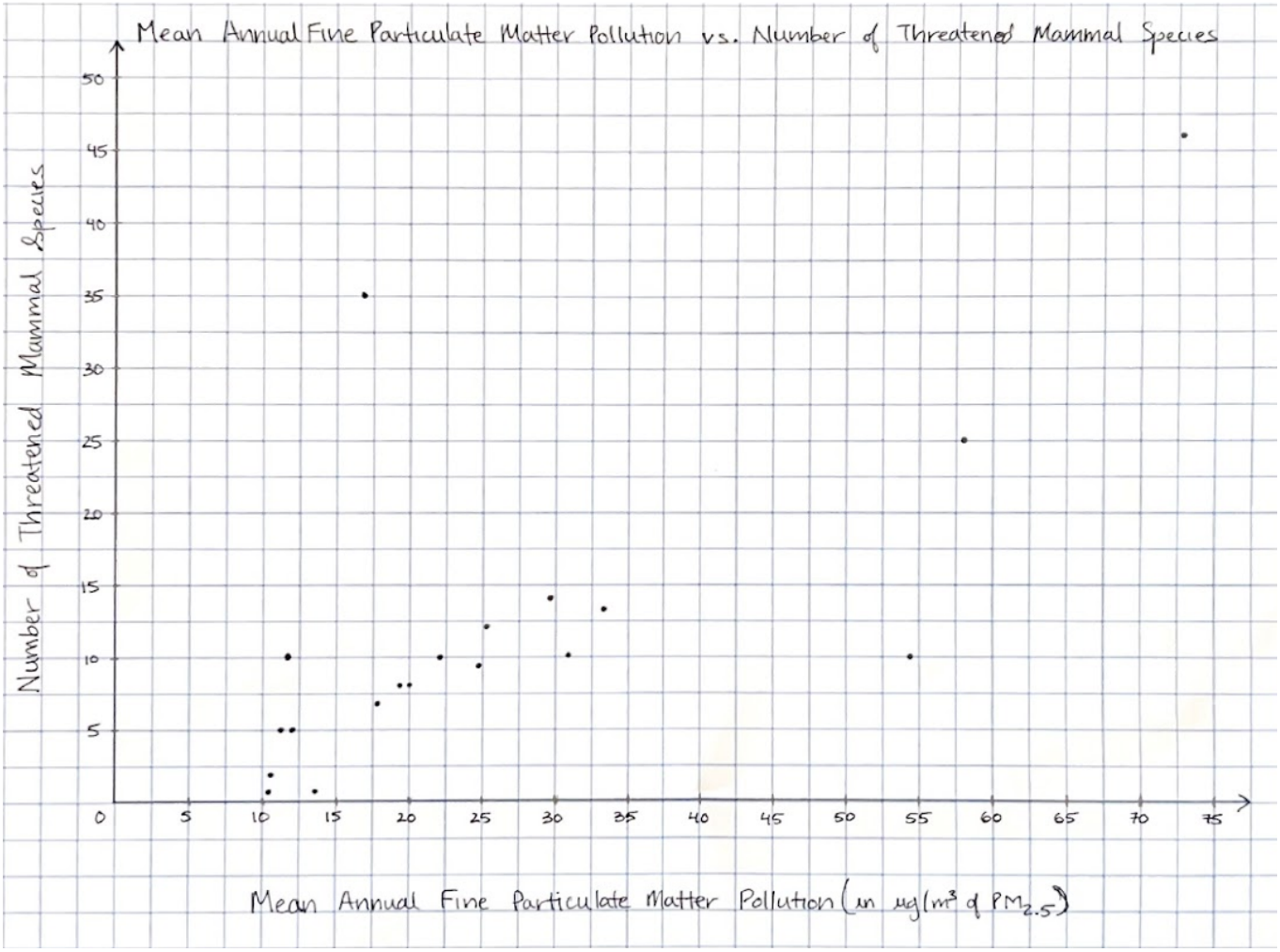
The table below contains the data collected for each of the 19 countries in the random sample. This data will be used to determine whether there is an association between a country's mean annual $PM_{2.5}$ concentrations and its number of threatened mammal species.

	Randomly Generated Number(seed → 17)	Country	Mean Annual Fine Particulate Matter Pollution(in µg/m^3 of $PM_{2.5}$)	Number of Threatened Mammal Species
1	132	Pakistan	58.28	25
2	108	Marshall Islands	10.24	1
3	163	South Korea	25.04	12
4	191	Venezuela	17.01	35
5	87	Jordan	33.01	13
6	71	Guinea-Bissau	29.77	14
7	112	Micronesia	11.28	5
8	32	Cameroon	72.79	46
9	98	Libya	54.25	10
10	94	Latvia	13.43	1
11	26	Bulgaria	19.15	8
12	195	Zimbabwe	22.25	10
13	168	Suriname	24.78	9
14	11	Azerbaijan	19.93	8
15	95	Lebanon	30.62	10
16	65	Germany	12.03	5
17	4	Andorra	10.31	2
18	137	Paraguay	11.91	10
19	125	Nicaragua	17.61	7

First, we need to create a scatter plot to present this data in a layout that makes it easy to visualize a possible relationship between the pollution and the number of threatened mammal species. We must then calculate and plot the least-squares regression line(LSRL) to further support our understanding of the strength, direction, and form of the relationship based on the scatter plot. The differences between our LSRL's predicted values and the observed values will be used to create a residual plot, which will allow us to determine whether a linear model is truly appropriate for this data.

Scatter Plot

The following scatter plot contains the data from the random sample. Each point represents a country from the sample. The x-axis represents our explanatory variable, the mean annual fine particulate matter pollution. The y-axis represents the response variable, the number of threatened mammal species. The reason for creating this scatter plot is to allow us to gain a general understanding of the relationship between the explanatory and response variables by visualizing the data.



At a glance of this scatter plot, there seems to be a moderate, positive, linear relationship between a country's mean annual amount of fine particulate matter pollution and its number of threatened mammal species. However, this is simply an assumption based on a quick glance of the scatter plot; to confirm this, the LSRL along with the other regression statistics need to be calculated.

Regression Line from Summary Statistics

The LSRL is a "line of best fit" that minimizes the sum of the residuals squared. It will be calculated using summary statistics, which include means and standard deviations. An LSRL is needed for two main reasons. The first reason is that it visually represents the general trend of the data, meaning the trend of dispersion or tightness of the data points can be assessed in relation to the "line of best fit". The other reason is that the equation of the LSRL can be used to calculate the expected value($\hat{y}$) of the response variable for a given value of the explanatory variable(x). The differences between the observed and expected values of the response variables can be used to create a residual plot, which will then help in determining the appropriateness of this linear model.

The LSRL will be calculated using the summary statistics of x and y, where x is the mean annual fine particulate matter pollution($PM_{2.5}$), and y is the number of threatened mammal species. The summary statistics that are needed for this calculation are the mean and standard deviation of the x values($\bar{x}$, s_x), the mean and standard deviation of the y values($\bar{y}$, s_y), and the correlation coefficient(r). First, the slope(b) of the LSRL needs to be calculated using the formula, $b = r(s_y/s_x)$, which will indicate by how much the number of threatened mammal species changes for each additional $\mu g/m^3$ of mean annual $PM_{2.5}$ on average. Then, using this slope, the y-intercept(a) will be calculated with the formula, $a = \bar{y} - b\bar{x}$, which tells us the average number of threatened mammal species when there is no fine particulate matter pollution present. Then, using the slope and y-intercept, the LSRL will be expressed as $\hat{y} = a + bx$.

x = mean annual fine particulate matter pollution ($\mu g/m^3$ $PM_{2.5}$)

y = number of threatened mammal species

$\bar{x} = 25.9836421$

$s_x = 17.683066698$

$\bar{y} = 12.15789474$

$s_y = 11.50006356$

$r = 0.6979861243$

slope

$\hookrightarrow b = \frac{r(s_y)}{s_x} = \frac{.69799(11.5001)}{17.6831} = 0.4539305769$

y-intercept

$\hookrightarrow a = \bar{y} - b\bar{x} = 12.1579 - 0.45393(25.9836)$

$= 0.3631059728$

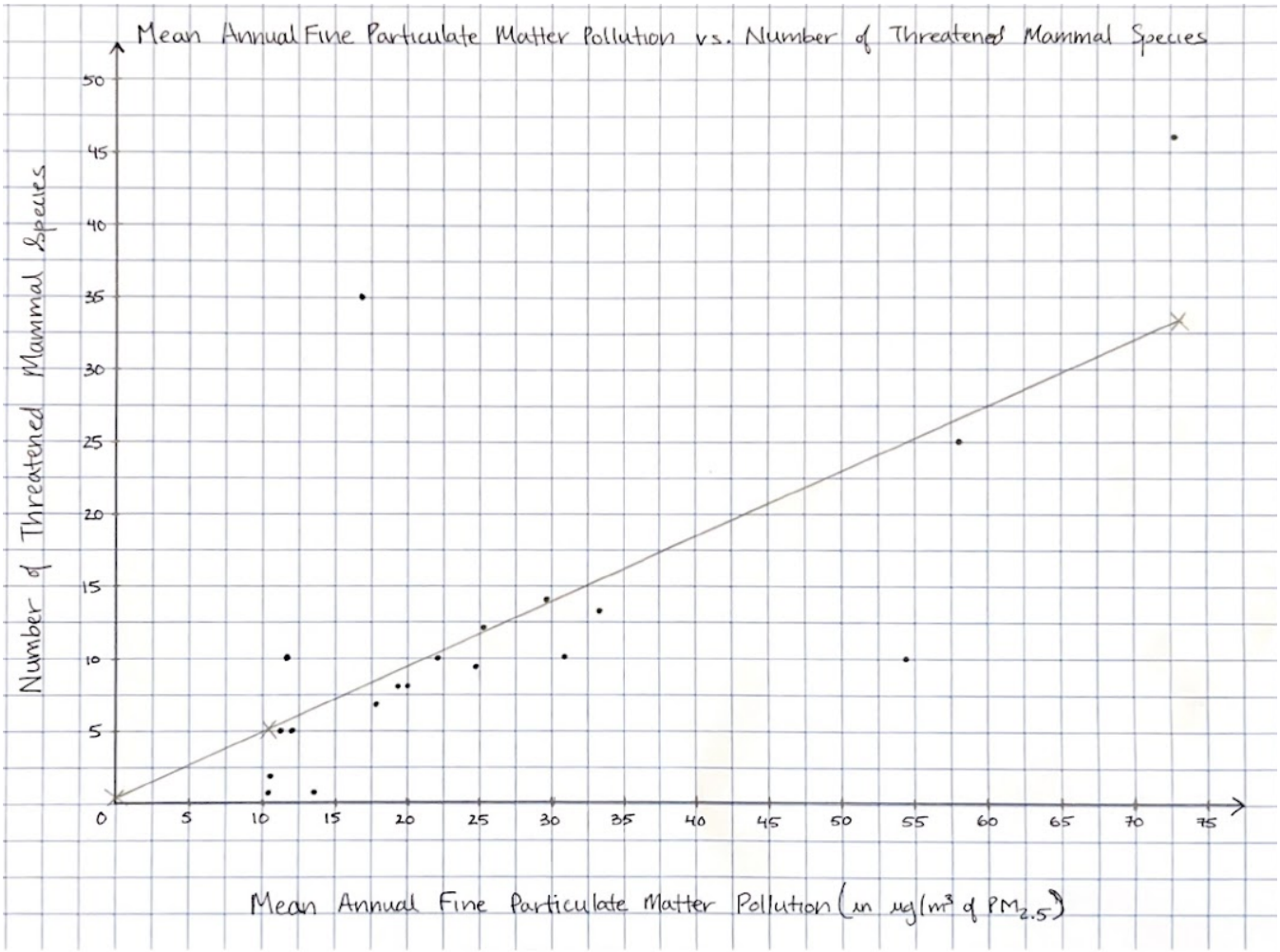
least squares regression line

$\hookrightarrow \hat{y} = 0.3631 + 0.45393x$

Now that we have calculated the LSRL, $\hat{y} = 0.3631 + 0.45393x$, we have all the statistics necessary to interpret this relationship. However, it would be informative to interpret the relationship in the context of the regression model. In other words, before moving on to the interpretations, we should make a scatter plot that contains the LSRL, which we can use later as a reference to better understand the meaning of the statistics that we have obtained.

In general, the LSRL can be most easily graphed by finding the predicted($\hat{y}$) value of the response value for the minimum and maximum values of the explanatory variable. In this case, the minimum amount of mean annual fine particulate matter pollution was 10.24 $\mu g/m^3$, in the Marshall Islands. The maximum amount was 72.79 $\mu g/m^3$, in Cameroon. Once these values are plugged into the LSRL's equation, we can simply draw a straight line to connect the resulting coordinates(x, $\hat{y}$).

The predicted number of threatened mammal species for the Marshall Islands, which had the minimum amount of mean annual fine particulate matter pollution, was 5.011. The predicted number of threatened mammal species for Cameroon was 33.405. The resulting coordinates are (10.24, 5.011) and (72.79, 33.405). To be cautious, we can also plot the y-intercept before drawing the LSRL, since it would increase the precision of the LSRL drawing. The scatter plot with the LSRL is shown below.



With the LSRL now graphed on the scatter plot, we are fully prepared to interpret the relationship between a country's mean annual fine particulate matter pollution and its number of threatened mammal species. We found the slope to be approximately 0.45393, which means that for each additional $\mu g/m^3$ of mean annual $PM_{2.5}$, a country's number of threatened mammal species increases by 0.45393 on average. The *y*-intercept was calculated to be approximately 0.3631. This means that when a country has a mean annual fine particulate matter pollution of 0 $\mu g/m^3$ $PM_{2.5}$(no pollution), it would have 0.3631 threatened mammal species on average. The value which tells us the most about this relationship is the correlation coefficient(*r*), which was calculated to be 0.69799. Based on this value, we can confirm our previous statement that there is a moderate, positive, linear relationship between a country's mean annual fine particulate matter pollution and its number of threatened mammal species. The scatter plot with the LSRL can visually support this interpretation: positive since the response variable generally increases as the explanatory variable increases, moderate since there is moderate variation about the linear trend, and linear since the number of threatened mammal species increases at a fairly constant rate in relation to the mean annual fine particulate matter pollution, as described above. Furthermore, we can multiply the correlation coefficient by itself to find the coefficient of determination(r^2), which equates to 0.48718. This means that about 48.72% of the variability in a country's number of threatened mammal species is explained by its mean annual amount of fine particulate matter pollution, once again indicating that our linear model is a decent fit for the data.

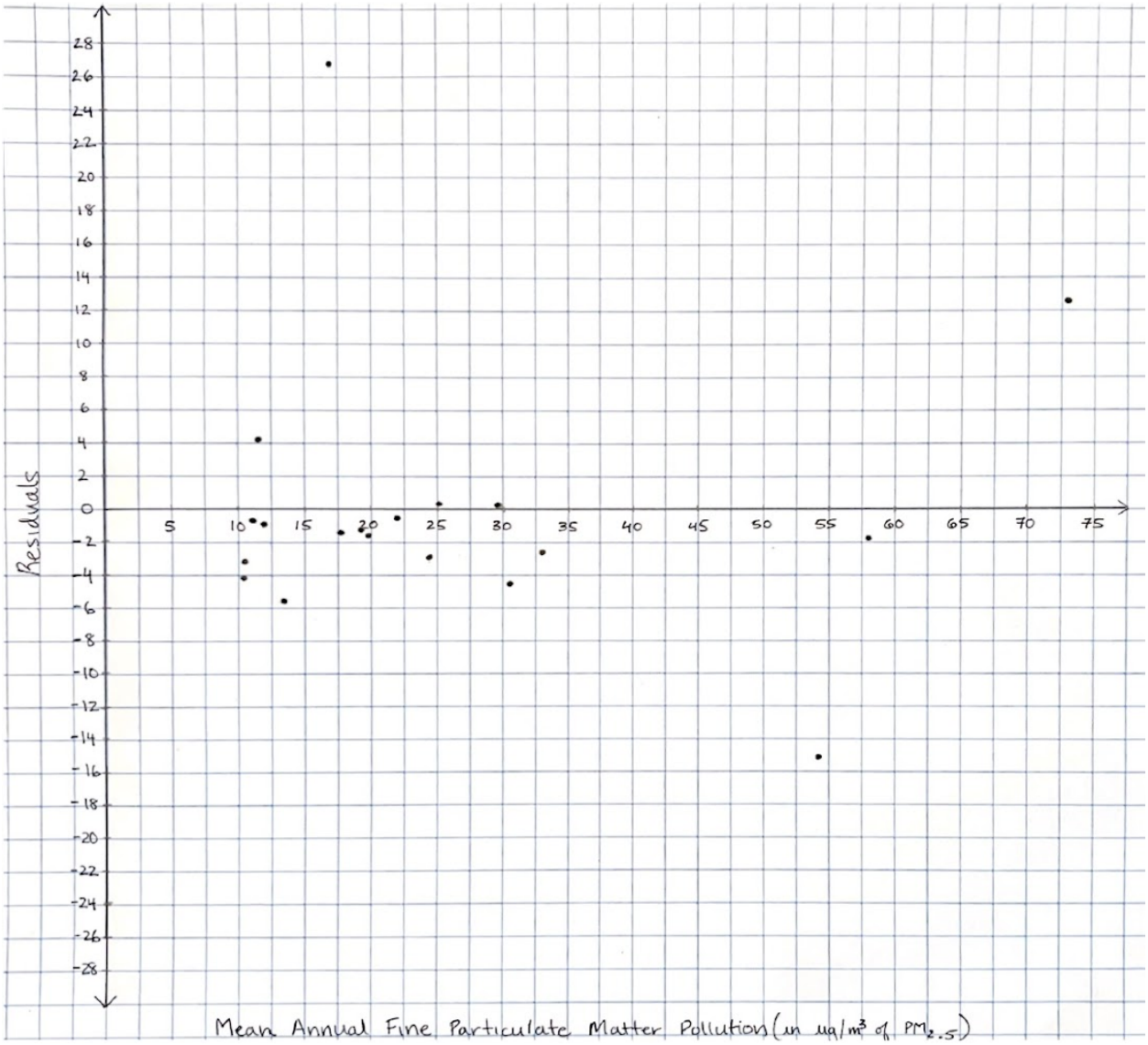
Residuals

A residual is the error between our model's predictions of the response variable and the observed values of that response variable; it is a measure of a point's vertical distance from the regression line. The residuals of each observation is calculated by inserting that value's corresponding explanatory value(x) into the equation of our LSRL that we calculated earlier. The resulting value is our predicted value based on the regression model, and will be subtracted from the observed value to calculate the residual. This process will be repeated for each observation. A positive residual indicates an underestimation of the response variable by our model, and a negative residual indicates an overestimation. Using all 19 residual values, a residual plot will be created to verify that a linear model is appropriate for this data. The residuals have been appended to the initial data table, shown below.

	Country	Mean Annual Fine Particulate Matter Pollution(in $PM_{2.5}$ $\mu g/m^3$)	Number of Threatened Mammal Species	Residual
1	Pakistan	58.28	25	-1.818
2	Marshall Islands	10.24	1	-4.011
3	South Korea	25.04	12	0.27049
4	Venezuela	17.01	35	26.916
5	Jordan	33.01	13	-2.347
6	Guinea-Bissau	29.77	14	0.1234
7	Micronesia	11.28	5	-0.4834
8	Cameroon	72.79	46	12.595
9	Libya	54.25	10	-14.989
10	Latvia	13.43	1	-5.459
11	Bulgaria	19.15	8	-1.056
12	Zimbabwe	22.25	10	-0.463
13	Suriname	24.78	9	-2.611
14	Azerbaijan	19.93	8	-1.41
15	Lebanon	30.62	10	-4.262
16	Germany	12.03	5	-0.8239
17	Andorra	10.31	2	-3.043
18	Paraguay	11.91	10	4.231
19	Nicaragua	17.61	7	-1.357

Residual Plot

Residual plots display the residual value of each response variable on the vertical axis, with the horizontal axis representing the value of the explanatory variable. The plot can be used to determine if a linear model is appropriate by checking if there is a pattern to the residuals. If a pattern is present, then a linear model will not be a good fit because the regression line had a systematic error in the prediction of the response variable for a substantial portion of the data set. Following the same logic, the absence of a pattern in the residual plot suggests that the use of a linear model is appropriate for our data, since there isn't a systematic error for our linear model's prediction. The residual plot is shown below.



The residual plot shows the differences between our predictions of a country's threatened mammal species and its true number of threatened mammal species. The residual plot has no obvious pattern, so a linear model is appropriate; we can continue our study with the model we calculated earlier. However, we should keep in mind that there seem to be three possible outliers that are clearly visible on this residual plot at (17.01,26.916), (54.25,-14.989), and (72.79,12.595).

Linear Regression T-Test and T-Interval

Earlier, when we calculated the LSRL with the summary statistics, we conducted our linear regression analysis to quantify the relationship between a country's mean annual fine particulate matter pollution and its number of threatened mammal species. Whenever this process is performed, the next step is to test whether the relationship between these variables is statistically significant. We conduct this significance test by running a linear regression t-test, in which we test the null hypothesis that there is no predictive relationship between a country's pollution and its number of threatened mammal species.

Significance tests yield a p-value associated with the test statistic(t, z, x^2). When this p-value is lower than a given threshold(represented by α), the results are considered to be statistically significant. This allows us to reject the null hypothesis. However, there are two possible errors that arise from this conclusion: Type I & II errors. Type I errors are when we mistakenly reject a true null hypothesis, and Type II errors are when we fail to reject a false null hypothesis.

In our situation, making a Type I error is preferable to making a Type II error. A Type II error in the context of our study would be when there *is* a predictive relationship between a country's mean annual fine particulate matter pollution and its number of threatened mammal species, but we say such a relationship doesn't exist, and ignore an existing connection between two global issues of extreme importance to our planet's health. This is undoubtedly worse than falsely concluding that there is a predictive relationship even when there is not one present(Type I), because there is no harm in finding a predictive relationship between our variables. In fact, if we make a Type I error, although it is still an error, its results would be better in terms of the well-being of the planet, since we would at least be aware that there is an association between the different aspects of environmental damage that humans' actions have. And as a result, hopefully we would be more considerate toward the environment by realizing that each harmful human activity toward the environment damages it in many more ways than one, even though this increased awareness was the result of an error.

Since we have established that making a Type II would be worse than a Type I, we should take steps to minimize the probability of a Type II error. There are two main ways to go about this. We could increase alpha(α), or we could increase sample size. Increasing the sample size, however, would cause some issues with our significance test as we would violate one of the conditions for conducting the test. Increasing alpha is the best way to minimize the probability of a Type II error in this situation. Alpha itself is equal to the probability of a Type I error, so by increasing alpha, we increase the probability of a Type I error. In many situations, this would be a less desirable way to minimize the probability of a type II error, but as we have made clear, a Type I is a "good" error to make in our situation. Furthermore, as alpha increases, the power of our test increases as well. Power is the probability that our significance test will correctly reject a false null hypothesis. Thus, increasing the alpha has three main benefits. It minimizes the probability of a Type II error, increases the probability of a Type I error, and increases the power of the significance test. A commonly used significance level(α) is 0.05, but since we want to minimize the probability of a Type II error and increase power, we will be running this test with $\alpha = 0.10$.

It should be noted that our alternative hypothesis will be declared as $H_a > 0$, since we are looking for a positive linear relationship between a country's mean annual amount of fine particulate matter pollution and its number of threatened mammal species. However, before we can begin the significance test, we must check the three main conditions for this test: the randomness, 10%, and linearity conditions. We conducted a simple random sample(SRS) of 19 countries, each of whose mean annual fine particulate matter pollution and number of threatened mammal species were collected. This random selection satisfies the randomness condition. The 10% condition is that the sample must be less than 10% of the population, and is used to check that the observations in our sample are close to independent. We randomly selected 19 countries(n = 19), and since there are 195 countries(> 190), the 10% condition is satisfied as well. Next, to check the linearity condition, we can refer to the previously constructed residual plot if needed(pg. 8). As we had observed, there was no obvious pattern to the residuals on the residual plot, suggesting that a linear model is appropriate for the data. This means that the form of the data is linear, satisfying this condition. However, on the residual plot, we had noticed three points that seemed to be outliers, so we should be careful when generalizing our results to the entire population of interest. The t-test is shown below.

Linear Regression t-test

H_0: $\beta = 0$ (no predictive relationship)

H_a: $\beta > 0$ (number of threatened mammal species can be predicted from mean annual fine particulate matter pollution and is a positive relationship)

β = true slope between mean annual fine particulate matter pollution and number of threatened mammal species of all countries

$df = n - 2 = 19 - 2 = 17$

$$t = \frac{b - \beta}{SE_b} = \frac{b - \beta}{\left(\frac{s}{s_x\sqrt{n-1}}\right)} = \frac{.4539305769 - 0}{\left(\frac{8.474079801}{17.68306698\sqrt{19-1}}\right)} = 4.018747762$$

p-value = tcdf(lower: 4.018747762, upper: 1E99, df = 17)

$= 4.4515 \times 10^{-4} \approx .0004 \approx 0$

Since $p(\approx .0004) < \alpha(.10)$, our results are significant. We can reject H_0, and we can say that the true slope is significantly greater than 0, which means that a country's number of threatened mammal species can be predicted from its mean annual fine particulate matter pollution and is a positive relationship

Our linear regression t-test calculations yielded a p-value of approximately 0.0004. This tells us that, if the true slope between the mean annual fine particulate matter pollution and the number of threatened species of all countries is equal to zero(no predictive relationship), then about 4 times out of 10,000, we would find a sample of 19 countries whose slope is greater than 0.

We can also construct a confidence interval to make a prediction about the true slope of the population. In this case, that would be a linear regression t-interval. This procedure has the same conditions as a linear regression t-test, all of which were checked previously. Ideally, we would want a confidence interval with a high confidence level and a low margin of error, but there is usually a trade-off between the confidence level and the margin of error. For our purposes, we will aim for an interval with decently high confidence despite the wider interval, because it would be more insightful to have a high confidence level than to have a tight interval with an inadequately low confidence level. Hence, we will construct a 90% confidence interval for the true slope between the mean annual fine particulate matter pollution and the number of threatened species of all countries. The interval procedure is shown below.

Linear Regression t-interval

$n = 19$

$df = n - 2 = 17$

to construct interval : $b \pm t^*(SE_b)$

$b = 0.4539305769$

$t^* = invT(area: .05, df = 17) = -1.739606667$

$SE_b = \frac{s}{s_x\sqrt{n-1}} = \frac{8.474079801}{17.68306698\sqrt{19-1}}$

$0.4539305769 \pm -1.739606667\left(\frac{8.474079801}{17.68306698\sqrt{19-1}}\right)$

$= (0.25744, 0.65042)$

We are 90% confident that the true increase in a country's estimated number of threatened mammal species for each additional unit of fine particulate matter pollution is between 0.25744 and 0.65042

Conclusion

The original scatter plot we created seemed to have a moderate, positive, linear relationship between the mean annual fine particulate matter pollution and the number of threatened mammal species of a country. To confirm this, a linear regression model was created, and the regression analysis allowed us to conclude that there is a moderate, positive, linear relationship between these variables. A residual plot was then created to support this conclusion by determining whether a linear model is appropriate for the data; it was found to be appropriate.

Then we tested the null hypothesis against the alternative that there is a significant, positive, predictive relationship between a country's mean annual fine particulate matter pollution and its number of threatened mammal species by conducting a linear regression t-test. We also constructed a linear regression t-interval with 90% confidence to estimate the true slope between the mean annual fine particulate matter pollution and the number of threatened mammal species of all countries.

We were able to make the following conclusions based on this study.

- For each additional $\mu g/m^3$ of mean annual $PM_{2.5}$, a country's number of threatened mammal species increases by 0.45393 on average.
- When a country has a mean annual fine particulate matter pollution of 0 $\mu g/m^3$ $PM_{2.5}$, it would have 0.3631 threatened mammal species on average.
- There is a moderate, positive, linear relationship between a country's mean annual fine particulate matter pollution and its number of threatened mammal species.
- About 48.72% of the variability in a country's number of threatened mammal species is explained by its amount of fine particulate matter pollution.
- The significance test yielded a test statistic(t) of 4.0187, with a p-value of .0004. This significant result allowed us to reject the null hypothesis and conclude that the true slope is greater than 0, meaning there is a positive, predictive relationship between the mean annual fine particulate matter pollution and the number of threatened mammal species for all countries.
- The 90% confidence interval allows us to conclude with 90% confidence that for each additional $\mu g/m^3$ of mean annual $PM_{2.5}$, the true increase in a country's number of threatened mammal species is between 0.25744 and 0.65042 on average.
- The confidence level of 90% means that, in repeated samples of 19 countries, 90% of the resulting confidence intervals will contain the true slope between the mean annual fine particulate matter pollution and the number of threatened mammal species of all countries.

Works Cited

"PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)." *World Bank Indicators*, World Bank, data.worldbank.org/indicator/EN.ATM.PM25.MC.M3? Accessed 17 Apr. 2022. Table.

"Mammal species, threatened." *World Bank Indicators*, World Bank, data.worldbank.org/indicator/EN.MAM.THRD.NO? Accessed 17 Apr. 2022. Table.

"Alphabetical list of countries." *Worldometer*, www.worldometers.info/geography/alphabetical-list-of-countries/. Accessed 17 Apr. 2022. Table.