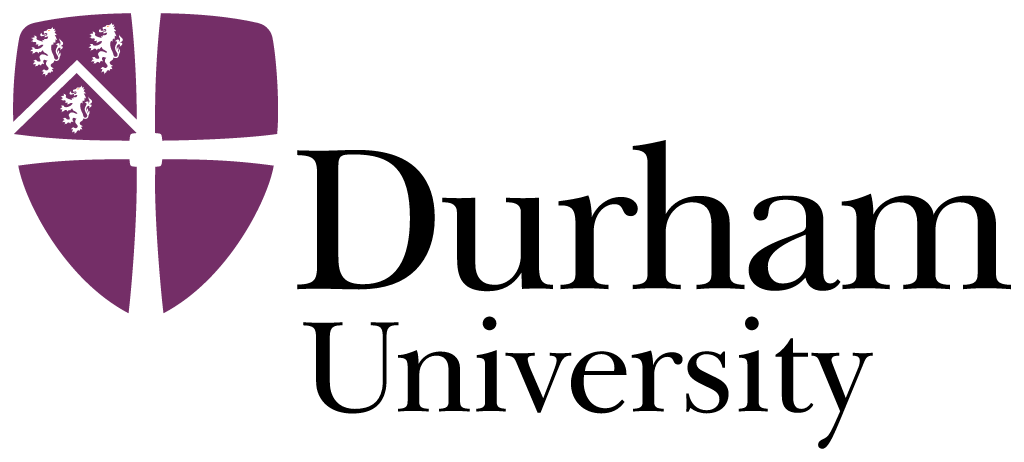
****

**MSc Business Analytics**

**Data Analytics in Action**

**Anonymous Marking Code: Z0149641**

**Question I.**

1. Given that X and Y are related and have no idea which way the causality goes, there is no valid way to find the causality between the variables. The main reason why we can’t say X causes Y or Y causes X is due to the common saying that ‘correlation does not imply causation’. To dive deeper into the meaning of this saying, one must understand the difference between the terms. Correlation is a statistical measurement that indicates the strength of a linear relationship between a pair of variables (Albright and Winston, 2017). In this question, we are told that X and Y are related meaning that some sort of correlation exists between the two variables, but not the why and how the relationship unfolded. Causation, on the other hand, is similar to the idea of cause and effect. Causality between two variables argues that any change to a value of one variable will cause a *change* in the value of the other variable. With the given information, there lacks evidence that if a value of the x-variable is changed that the value in y-variable will also change and vise versa. As far as regression analysis goes, there are “two potential objective of regression analysis: to understand how the world operates and to make predictions” (Albright and Winston, 2017). Even with successfully regressing Y on X and then X on Y, will only construct information on how strong of *predictions* we can make of one variable based on a single explanatory variable. Comparing this to the idea of causality, regression analysis is a technique where one can predict values based on a single explanatory variable, therefore, it does not give any valid explanation whether or not a variable *causes* another variable. In conclusion, neither X causes Y nor Y causes X. This cannot be seen by regression analysis between the two variables because regression predictions are different from true causation.
2. There is a classic example where the direction of causality would be ambiguous due to the idea that ‘correlation does not imply causation’. This example is the correlation between ice cream sales and the rate of murders. While one is a dessert and one is a violent crime, “summer is when people get together. More specifically, casual drinkers and drug users are more likely to go to bars or parties on weekends and evenings, as opposed to a Tuesday morning. These people in the social mix, flooding the city’s streets and neighborhood bars, feed the peak times for murder, experts say” (Peters, 2013). Summer is also known as the peak time of ice cream sales. Thus, one can argue that ice cream sales and murder rates are correlated. This is a prime example of how the direction of causality would be ambiguous:

* Just because an individual buys an ice cream, does not cause a murder.
* Just because there is a murder, does not cause an ice cream purchase.

This mere statistical coincidence shows that just because two variables are correlated does not mean one causes the other and vice versa. For that reason, the ice cream sales and murder rates example demonstrates how the direction of causality would be ambiguous.

**Question II.**

Given the number of cars per 1000 people is known for virtually every country in the world, there are many countries, however, where per capita income is unknown. To go about estimating per capita income for these countries, we first have to understand what per capita income is. Per capita income is “a measure of the amount of money earned per person in a nation or geographic region” (Kenton, 2019). Per capita income can also be used to determine an area’s standard of living or quality of life. This is the calculation:

The next step of estimating per capita income for countries where it is unknown is to understand what information is already given; which is, the number of cars per 1000 people for the country. This can be represented with the formula:

With both of these formulas in mind, we are also told that *“for many*” countries, per capita income is **unknown**; therefore, there are some countries where per capita income is **known**. Using the values of cars per 1000 people as the *x*-variable, explanatory and independent variable, and the values of per capita income for the known counties as the *y-*variable, response and dependent variable, one can create a scatterplot. After building the scatterplot, regression analysis can be used to see if there is a linear relationship between the two variables, cars per 100 people and per capita income. If there is a linear relationship, then a least-square line can be fitted to the scatterplot based on minimizing the distance from the points to the line. This least-squares estimation can help predict and estimate per capita income for countries where it is unknown based on the known values of cars per 1000 people with the country.

**Question III.**

Yes, as the procurement manager of the company, it is possible to calculate the probability that the delivery is from an unhealthy supplier given that it is rejected. The first step of this calculation is to examine what information is given to us.

* The probability of an “unhealthy supplier” that produces a certain amount of defective hi-tech component is computed as 0.5%.
* If the supplier is labeled unhealthy, the probability of rejecting the delivery is 85%.
* If the supplier is labeled healthy, the probability of rejecting the delivery is 10%.

With this information in mind, I have created a table with the probabilities of true rejects, false rejects, true accepts, and false accepts of deliveries based on healthy and unhealthy labels.

|  |  |  |
| --- | --- | --- |
|  | Unhealthy Label (0.5%) | Healthy Label (99.5%) |
| Rejecting | *True Reject*  0.005 x 0.85 = 0.00425 | *False Reject*  .995 x 0.1 = 0.0995 |
| Accepting | *False Accept*  0.05 x 0.15 = 0.0075 | *True Accept*  .995 x 0.9 = 0.8955 |

In addition to the table, I have constructed a probability tree with the likelihoods of the rejects and accepts given unhealthy and healthy labels. Something to pay attention to is that Event A is the delivery has an unhealthy label and Event B is that the delivery is rejected.

Reject

P(A and B) = (0.005) (0.85) = 0.00425

*P(B|A) =* 85%

Unhealthy

*P(A)* = 0.5%

P(A and ) = (0.05) (0.15) = 0.0075

Accept

*P()* = 15%

P( and B) = (0.995) (0.1) = 0.0995

Reject

*P() =* 10%

Healthy

*P() =* 99.5%

Accept

*P() =* 90%

P( and ) = (0.995) (0.9) = 0.8955

Unfortunately, to calculate the probability that the delivery is from an unhealthy supplier given that the delivery is rejected, *P(A|B)*, cannot be performed using the table or probability tree. This is because *P(A|B)* has to take into account the true rejections, *P(),* as well as the false rejections, *P().* Thus, to calculate *P(A|B)* has to be done using Bayes’ Theorem:

This what each part of the theorem means in correlation to the problem:

* *P(B|A)*, the probability that the delivery is rejected given that it is has an unhealthy label.
* *P(A)*, the probability that the delivery has an unhealthy label.
* *P(B)*, the probability that the delivery is rejected. This is a *mutually exclusive event* meaning that it has to take into account the delivery being rejected from with a unhealthy label and with a healthy label. Thus the denominator turns into: *P(B|A) P(A) + P(B|) P()*.

Thus, I have substituted Bayes’ Theorem with the variables in question:

Before we put together the final calculation, here are the individual calculations for each probability:

|  |  |  |
| --- | --- | --- |
|  |  | Calculations: |
| *P(Reject | Unhealthy)* | *P(B|A)* | 0.85 |
| *P(Unhealthy)* | *P(A)* | 0.005 |
| *P(Reject | Healthy)* | *P(B|)* | 0.10 |
| *P(Healthy)* | *P()* | 0.995 |
| *P(Reject | Unhealthy) P(Unhealthy)* | *P(B|A) P(A)* | 0.85 0.005 = 0.00425 |
| *P(Reject | Healthy) P(Healthy)* | *P(B|) P()* | 0.10 0.995 = 0.0995 |
| *P(Reject | Unhealthy) P(Unhealthy) + P(Reject | Healthy) P(Healthy)* | *P(B|A) P(A) + P(B|) P()* | 0.00425 + 0.0995 = 0.10375 |

Using the calculations in the table above and substituting them into Bayes’ Theorem, the final probability calculation of *P(Unhealthy |* Reject) or *P(A|B)* is the following:

In conclusion, the probability that the delivery is from an unhealthy supplier given that the delivery has been rejected is approximately **4.1%**.

**Question IV.**

A screenshot of a cell phone

Description automatically generated

1. By observing the data points in the scatterplot above, one can observe that a high amount of motors are returned if inspection costs are too low or too high. Unfortunately, the manger cannot make strong conclusions by just observing the data points with the scatterplot because the graph does not quantify the relationship between the two variables, the Inspection expenditures and motors returned. This is where a simple linear regression can help quantify the relationship where there is a single explanatory variable. The steps that were taken to construct this linear regression are the following:

* Drawing a scatterplot is a good way to begin the linear regression analysis. The scatterplot is a graphical plot of two variables, *x* and *y*. The *x*-variable, also known as the independent or explanatory variable, will be the inspection expenditures. The *y*-variable, also known as the dependent or response variable, will be the number of motors returned.
* After selecting both of the columns, within the Excel spreadsheet, that contains both the *x* and *y* variables, I inserted the scatterplot.
* Next, I was able to add a chart element under Chart Design, which was a *linear treadline*. This is where the least-squares line is constructed: a straight line is fitted in the scatterplot that “minimizes the sum of squared residuals” (Albright and Winston, 2017). A residual is a difference between the actual and fitted values of the dependent variable, the number of motors returned (Albright and Winston, 2017). Therefore, the fundamental equation for regression is:
* The linear treadline, also known as the least-squares line, is calculated by an equation containing the slope and the y-intercept.
  + The equation for the slope is: =
  + The equation for the y-intercept is:
* With the help of Excel, the formulated simple linear regression equation is:
  + **.**

A screenshot of a cell phone

Description automatically generated

The simple linear regression equation contains two important features where the manager can draw conclusions from are the y-intercept and slope. The y-intercept allows the manager to interpret how many motors will be returned when the inspections costs are $0. The y-intercept = 66.476 so therefore, when inspections costs are $0, the manager can predict approximately 66.5 motors to be returned. Importantly, this demonstrates that no matter the inspection costs, there will still be motor returns. The slope allows the manager to interpret the rate at which the motors are being returned at per inspection dollar spent. The slope = -0.00002 so therefore, every dollar spent on an inspection there will be a 0.00002 decrease in the amount of motors returned. Overall, the slope is negative which concludes that as inspection expenditure increases, the number of motors returned will decrease.

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

1. The R-Square value, of this linear regression model is **0.01798.** R-Square is also known as the coefficient of determination, which is a statistical measurement of how close the datapoints are to the fitted to the least squares line of the regression. The R-Square value is calculated using the formula:

.

Specifically, in simple linear regressions, the R-Square value is the square of the correlation between the dependent variable, number of motors returned, and explanatory variable, inspection expenditures. The R-Square value is always between 0 and 1; the closer to 1, the better the fit. In this problem, the R-Square value is 0.018 which is closer to 0 than it is to 1; therefore, the linear regression of this dataset is poorly fitted. This can also be seen in the residual lot as well as it shows the linear regression “systematically over and under-predicts the data at different points along the curve” (Rieuf, 2017). The Standard Error of Estimate, Se, of the regression model is **1.4495** motors returned. The Se, essentially standard deviation of the residuals, indicates the level of accuracy of predictions made from the regression equation. The Se can be measured with the formula:

.

While 1.45 seems like a smaller number, a standard error of 1.45 models returned is quite large with a dataset’s range between 64 and 68 motors returned. This can conclude that the predictions based on the linear regression model of electric motors returned based on inspection costs are not accurate.

**Question V.**

1. The most representative average commute time for individuals living in London metropolitan area is approximately **42.65 minutes**. Assuming that this sample is a portion of the whole population and randomly selected, then in theory, a sample mean can be used as a measure of central tendency. Unfortunately, the distribution of the data is skewed which means the measure of central tendency for this problem is the *median*. The median is calculated by sorting the data from smallest to largest and selecting the middle observation:

30.5 32.9 33.3 33.6 34.9 35.1 36 36.6 37.5 37.5

37.6 38 38.8 39.6 40.8 41.6 41.7 41.9 42.3 42.4

42.9 43.2 43.8 43.9 44.1 45.4 46.5 46.6 47.3 47.6

47.7 48.1 49.1 49.5 49.9 52.5 52.8 56.9 58.6 62.5

With a sample of 40 commuters, the middle observation is defined as the average of the two middle observations due to an even sample size. In this case, the average between 42.4 and 42.9. Therefore, the final calculation is:

This can also be calculated by selecting the data and using Excel’s MEDIAN function.

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

1. Looking at the histogram above, it appears that the distribution of average commute time is not approximately symmetric. In addition to the histogram, the descriptive statistics area of the Data Analysis Toolpak demonstrates the calculated skewness in the formulated table above. In general, this is what skewness means when it come to the distribution:

* If skewness is negative, the distribution is skewed to the left.
* If skewness is positive, the distribution is skewed to the right.
* If the skewness is approximately zero, the distribution is symmetric/normal.

Because the skewness is approximately **0.57**, then the distribution of average commute time is skewed to the right. For a distribution to be skewed to the right, the mean should be greater than the median. The mean and median calculations for this dataset can be seen in the table below:

|  |  |
| --- | --- |
| Median | 42.6 |
| Mean | 43.3 |

As we can see, the mean is greater than the median, therefore, the distribution of average commute time is skewed to the right. Another way to see that this dataset is not a symmetric distribution is introducing the technique of using minimum, maximum, percentiles and quartiles.

A screenshot of a social media post

Description automatically generated A screenshot of a cell phone screen with text

Description automatically generated

Quartiles divide the data into four groups, with approximately a quarter of all observations within each of the quartiles. A boxplot is constructed according to these four quartiles from this dataset and here are the steps to constructing such a boxplot (Support.office.com, 2013):

* Calculate quartile values (seen in the table above):
  + minimum value, first quartile, median value, third quartile, maximum value.
* Calculate quartile differences:
  + First quartile – minimum value
  + Median value – first quartile
  + Third quartile – median value
  + Maximum value – third quartile
* Using these differences, I used Excel’s stacked column design.
* Next, the topmost box and bottom box are turned into *whiskers*. The middle two quartile boxes are turned into the same color and thus creating the final product of the constructed boxplot.

The interquartile range (IQR) is represented as the light blue box and contains the middle 50% of the data. The line in the middle of the IQR represents the median, the middle observation of the sorted dataset. The green dot in the boxplot represents the mean of the dataset. In the boxplot, the green dot is above the middle line, meaning the value of the green dot is greater than the middle line. Thus, the mean is greater than the median. This concludes that the distribution average commute times is positively skewed as the mean is greater than the median. Therefore, with the help of Toolpak’s descriptive statistic skewness feature and boxplot’s interquartile range, the average commute time is not approximately symmetric.

**Bibliography:**

Albright, S. and Winston, W. (2017). *Business Analytics: Data Analysis & Decision Making*. 6th ed.

Boston, MA: Cengage Learning, pp. 419, 430-431.

Kenton, W. (2019). *How Per Capita Income is Calculated and Used by Companies*. [online]

Investopedia. Available at: https://www.investopedia.com/terms/i/income-per-capita.asp [Accessed 14 Jan. 2020].

Peters, J. (2013). *When Ice Cream Sales Rise, So Do Homicides. Coincidence, or Will Your Next Cone*

*Murder You?*. [online] Slate Magazine. Available at: https://slate.com/news-and-politics/2013/07/warm-weather-homicide-rates-when-ice-cream-sales-rise-homicides-rise-coincidence.html [Accessed 14 Jan. 2020].

Rieuf, E. (2017). *How To Interpret R-squared and Goodness-of-Fit in Regression Analysis*. [online]

Datasciencecentral.com. Available at: https://www.datasciencecentral.com/profiles/blogs/regression-analysis-how-do-i-interpret-r-squared-and-assess-the [Accessed 14 Jan. 2020].

Singh, S. (2018). *Why correlation does not imply causation?*. [online] Medium; towards data science.

Available at: https://towardsdatascience.com/why-correlation-does-not-imply-causation-5b99790df07e [Accessed 14 Jan. 2020].

Statistics.laerd.com. (2018). *FAQs on Measures of Central Tendency - Mean, Mode and Median | Laerd*

*Statistics*. [online] Available at: https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median-faqs.php [Accessed 15 Jan. 2020].

Support.office.com. (2013). *Create a box plot*. [online] Available at: https://support.office.com/en-

gb/article/create-a-box-plot-10204530-8cdf-40fe-a711-2eb9785e510f [Accessed 15 Jan. 2020].