Module: MA0009 Introduction to Probability and Statistics

Television Advertising

Project Title: *Linear Regression for Marketing*

Group: *T9JM_6*

Student Names: *Mustafa Iqbal, Aliyah Zaman, Silvia Botoaca, Deividas Talocka, Baurneegan Kanesalingam*

Group Leader: *Deividas Talocka*

Instructor(s) Name(s): *Jack McKenna*

Date of Submission: *20/12/2024*

Table of Contents

Project Tasks

Write a report for the senior management team addressing the question: Is there a linear relationship between the funds allocated to TV advertising and the number of sales of the company's products?

- Create a scatter plot to explore the relationship between the company's expenditure on TV advertising and their subsequent sales. Justify your choice of the explanatory and response variables.
- Calculate the correlation coefficient and the equation of the regression line. Explain how predictions are made using your model equation (use examples). Interpret the regression line coefficients, the correlation coefficient. Calculate and interpret the coefficient of determination.
- Use residual analysis to assess the validity of your model.
- Build and interpret the 95% and 99% confidence intervals for the TV ads data. Which interval is wider? Explain why.
- The TV Guru CEO claims that the average sum spent on TV advertising by the firm's clients is £14,000. Assess the validity of this claim. Justify using appropriate statistical methods.

Project Implementation

Work Division

The project tasks were divided into five distinct sections with each member taking responsibility for one task:

- **Task 1 (Scatterplot Creation)** – Deividas was responsible for creating scatterplots to visualise the correlation between TV ad expenditure and product sales.

- **Task 2 (Regression Analysis)** – Silvia handled the regression analysis to quantify the relationship between advertising spend and sales.

- **Task 3 (Residual Analysis)** – Baurneegan performed residual analysis to evaluate the accuracy and assumptions of the regression model.

- **Task 4 (Confidence Intervals)** – Mustafa calculated the confidence intervals to provide estimates of the range within which the true TV ads expenditure were likely to fall, with confidence levels of both 95% and 99%.

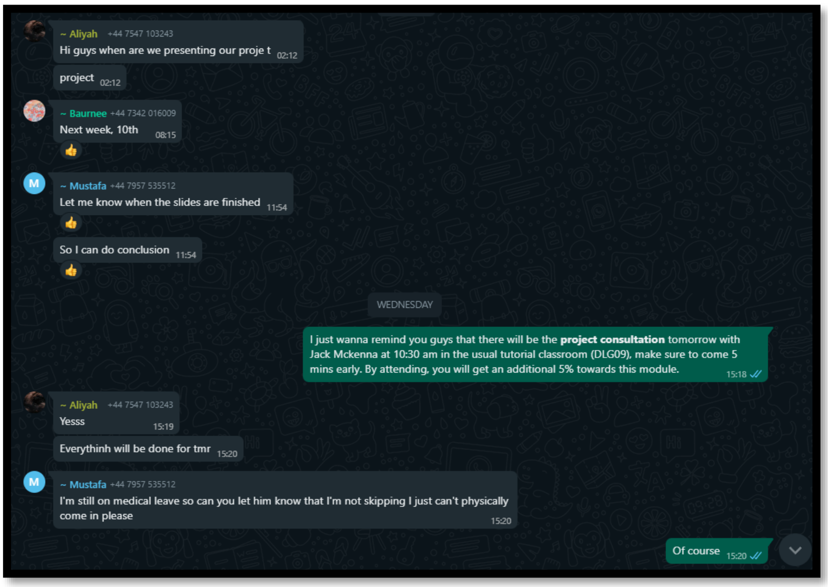
- **Task 5 (Hypothesis Testing)** – Aliyah carried out the hypothesis testing to statistically validate claims regarding the impact of TV advertising on sales.

Collaboration Tools and Workflow

We relied on WhatsApp for communication so that we can ensure that team members could share updates as well as ask questions and resolve issues promptly. For collaborative work, we shared a PowerPoint presentation to compile and prepare the final visual and verbal presentation as well as a shared Excel file to perform all calculations, including regression and residual analysis and finally a shared Word document to finalise the project report.

Meetings and Online Communication

Our team decided to primarily rely on online communication with WhatsApp instead of physical meetings. This was much more convenient as it allowed us to stay connected and collaborate without the need for in-person meetings so we can save time and be flexible in our work schedules. We frequently shared updates on progress and discuss any issues which came up. This also allowed us to prepare for any important events such as our consultation with Jack McKenna on 05/12/2024.



~ Aliyah +44 7547 103243
Hi guys when are we presenting our proje t 02:12

project 02:12

~ Baurnee +44 7342 016009
Next week, 10th 08:15
👍

~ Mustafa +44 7957 535512
Let me know when the slides are finished 11:54
👍

So I can do conclusion 11:54
👍

WEDNESDAY

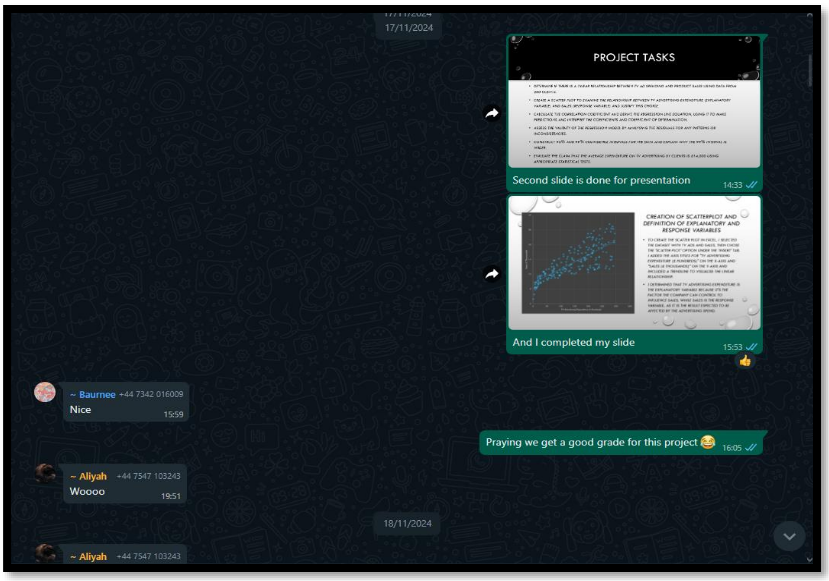
I just wanna remind you guys that there will be the **project consultation** tomorrow with Jack Mckenna at 10:30 am in the usual tutorial classroom (DLG09), make sure to come 5 mins early. By attending, you will get an additional 5% towards this module. 15:18 ✓✓

~ Aliyah +44 7547 103243
Yesss 15:19

Everythinh will be done for tmr 15:20

~ Mustafa +44 7957 535512
I'm still on medical leave so can you let him know that I'm not skipping I just can't physically come in please 15:20

Of course 15:20 ✓✓



17/11/2024

17/11/2024

PROJECT TASKS

- DETERMINE IF THERE IS A LINEAR RELATIONSHIP BETWEEN TV AD SPENDING AND PRODUCT SALES USING DATA FROM 200 CLIENTS.
- CREATE A SCATTER PLOT TO EXAMINE THE RELATIONSHIP BETWEEN TV ADVERTISING EXPENDITURE (EXPLANATORY VARIABLE) AND SALES (RESPONSE VARIABLE) AND JUSTIFY THIS CHOICE.
- CALCULATE THE CORRELATION COEFFICIENT AND DERIVE THE REGRESSION LINE EQUATION, USING IT TO MAKE PREDICTIONS AND INTERPRET THE COEFFICIENTS AND COEFFICIENT OF DETERMINATION.
- ASSESS THE VALIDITY OF THE REGRESSION MODEL BY ANALYSING THE RESIDUALS FOR ANY PATTERNS OR INCONSISTENCIES.
- CONSTRUCT 95% AND 99% CONFIDENCE INTERVALS FOR THE DATA AND EXPLAIN WHY THE 99% INTERVAL IS WIDER.
- EVALUATE THE CLAIM THAT THE AVERAGE EXPENDITURE ON TV ADVERTISING BY CLIENTS IS £14,000 USING APPROPRIATE STATISTICAL TESTS.

Second slide is done for presentation 14:33 ✓✓

CREATION OF SCATTERPLOT AND DEFINITION OF EXPLANATORY AND RESPONSE VARIABLES

- TO CREATE THE SCATTER PLOT IN EXCEL, I SELECTED THE DATASET WITH TV ADS AND SALES, THEN CHOSE THE "SCATTER PLOT" OPTION UNDER THE "INSERT" TAB. I ADDED THE AXIS TITLES FOR "TV ADVERTISING EXPENDITURE (£ HUNDREDS)" ON THE X-AXIS AND "SALES (£ THOUSANDS)" ON THE Y-AXIS AND INCLUDED A TRENDLINE TO VISUALISE THE LINEAR RELATIONSHIP.
- I DETERMINED THAT TV ADVERTISING EXPENDITURE IS THE EXPLANATORY VARIABLE BECAUSE IT'S THE FACTOR THE COMPANY CAN CONTROL TO INFLUENCE SALES, WHILE SALES IS THE RESPONSE VARIABLE, AS IT IS THE RESULT EXPECTED TO BE AFFECTED BY THE ADVERTISING SPEND.

And I completed my slide 15:53 ✓✓
👍

~ Baurnee +44 7342 016009
Nice 15:59

Praying we get a good grade for this project 😂 16:05 ✓✓

~ Aliyah +44 7547 103243
Woooo 19:51

18/11/2024

~ Aliyah +44 7547 103243

Methodology and results

Task 1 (Scatterplot Creation) – Deividas Talocka

The goal of this task was to create a scatterplot to explore the relationship between the company's expenditure on TV advertising and their subsequent sales. This analysis aimed to identify trends and provide insights into how advertising spend influences sales.

Steps Taken

I used the provided Excel file containing both datasets that being TV ad expenditure and sales. To create the scatterplot, I had selected all data within the dataset and navigated to the insert tab and chose the scatterplot option. (Academy, 2021)

I had also added a trendline by right-clicking on the scatterplot selecting add trendline and choosing the linear trendline option to the scatterplot. (Bradburn, 2021). In addition, I enabled display equation on chart to show the equation of the trendline and display R-squared value on chart under the trendline settings to show the coefficient of determination value directly on the chart. (Bradburn, 2021).

For visual appeal and clarity, I customised the chart by renaming the axes to show where the sales and TV advertising expenditure are and adjusting the visual appearance with a black background and blue data points. (Academy, 2021). The scatterplot is in the excel file within the tab labelled '*Task 1 (Scatterplot Creation)*'. (Learning, 2017)

Results

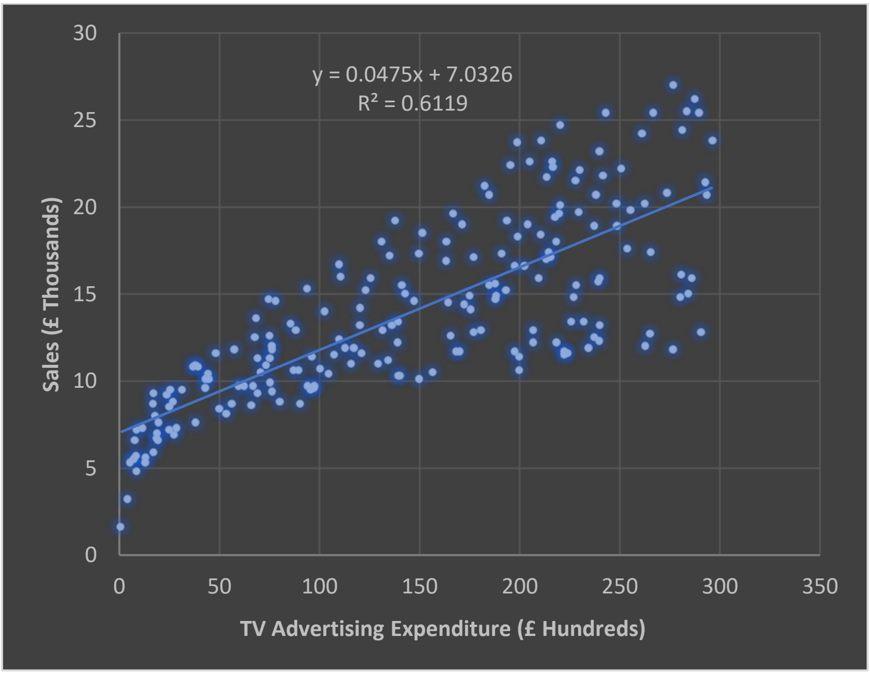


Figure 1. Scatter plot of TV Advertising Expenditure and Sales

The scatterplot (Figure 1) indicates a positive linear relationship between TV advertising expenditure and sales. It can be seen that as the company increase their TV advertising expenditure, their sales tend to increase.

The equation of the trendline is y = 0.0475x + 7.0326y. The trendline indicates that for every £100 spent on TV advertising sales increase by approximately 47.5 items with a baseline of 7,032 items sold even with no advertising.

The coefficient of determination value of 0.6119 suggests that approximately 61.19% of the variation in sales can be explained by the expenditure on TV advertising.

It was determined that the TV advertising expenditure was the explanatory variable as it represents the controllable factor that influences sales. Sales was chosen as the response variable as it is affected by the level of advertising expenditure.

Conclusion

The analysis confirms that TV advertising expenditure is a strong predictor of product sales. By investing more in TV ads, the company generally sees better sales performance as highlighted by the positive slope of the trendline and the coefficient of determination value.

Task 2 (Regression Analysis) – Silvia Botoaca

Steps Taken

Firstly, the value of **r** was calculated using the Excel function CORREL(A2:A201, B2:B201) to give a correlation coefficient of 0.7822 (4 s.f.). (Sharpe, 2019)

The SLOPE(B2:B201, A2:A201) and INTERCEPT(B2:B201, A2:A201) functions were used to calculate the value of the **slope/b** and **y-intercept/a** respectively. (Tutor, Excel Basics - Linear Regression - Finding Slope & Y Intercept, 2018). These values were substituted into the form y = **b**x + **a** to obtain the equation of the regression line:

$$y = 0.0475x + 7.0326$$

The coefficient of determination is obtained by squaring the value of r, which results in r^2 = 0.6119.

Results

Interpreting the regression line coefficients:

> Slope/b: For a 1 hundred pounds (£100) increase in the amount spent on TV ads by the client, product sales are expected to increase by approximately 0.0475 thousands (47.5).
>
> y-intercept/a: Where the client has spent £0 on TV ads, product sales are predicted to be approximately 7.0326 thousands (7,032.6).

Predictions can be made using:

- Interpolation: substituting values from within the range of the observed dataset into the regression line equation.

  $$x = 143.5$$
  $$y = 0.0475(143.25) + 7.0326 = 13.837$$

  If the client spends £14,325 on TV ads, product sales are predicted to be 13,837.

- Extrapolation: substituting values from outside the range of the observed dataset into the regression line equation.

  $$x = 550$$
  $$y = 0.0475(550) + 7.0326 = 33.1576$$

If the client spends £55,000 on TV ads, product sales are predicted to be 33,157.6.

Interpreting the correlation coefficient & coefficient of determination

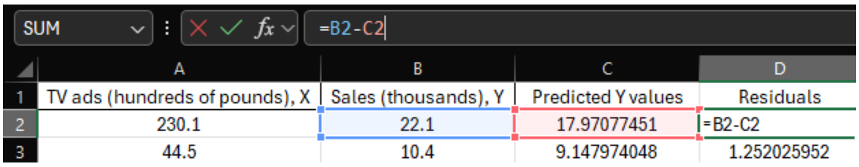
The correlation coefficient r = 0.7822 illustrates that there is a **strong positive** correlation between the explanatory variable (TV ads) and response variable (Sales).

The coefficient of determination shows that 61.19% of variation in Sales is due to the change in the amount spent on TV ads, and 38.81% is accredited to other factors.

Task 3 (Residual Analysis) – Baurneegan Kanesalingam

Steps Taken

Using the equation of the line of regression: y = 0.0475x + 7.0326, we found our *predicted Y* values. Using these and our actual *Y* values- we found the residuals between these using the simple formula of : actual y value - predicted y value. *In excel it was written as =B2-C2.*



	A	B	C	D
1	TV ads (hundreds of pounds), X	Sales (thousands), Y	Predicted Y values	Residuals
2	230.1	22.1	17.97077451	=B2-C2
3	44.5	10.4	9.147974048	1.252025952

Figure 2. Formula for calculating residuals

This gave me the list of residuals that I used to create my residuals plot graph, against an x axis of TV ads (hundreds of pounds). Using excel, I highlighted the explanatory variable (TV Ads Data) and the residuals and pressed insert to create my residuals plot graph which can be seen below as Figure 3. (Freeman, 2020)

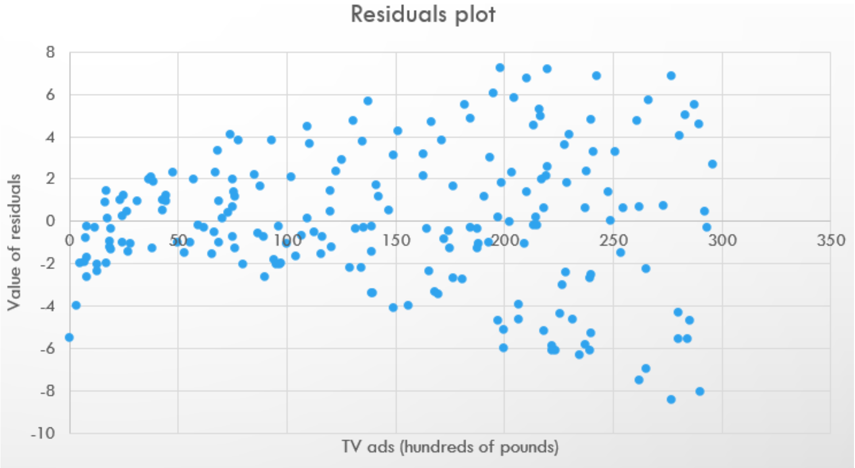


Figure 3. Residual Plot Graph of TV ads data

Results

We can see that as our explanatory variable increases, so does the variability in our residuals. This is suggestive that our model does not take a variable into account, which results in there being no constant variance (the variability of our residuals increase and spread out as x increases).

This kind of trend, seeing the residual variance increase as our explanatory variable increase, is called heteroscedasticity.

Conclusion

Our results showing heteroscedasticity suggest that our model may be missing a variable and is therefore incomplete. The missing variable(s) could include the time frames that the adverts are put up, as if it is an advert for a summer product that runs through the whole year, our results and therefore model validity would be compromised.

Task 4 (Confidence Intervals) – Mustafa Iqbal

The objective of this task was to calculate and interpret the 95% and 99% confidence intervals for the sample data on TV advertising expenditure. The analysis aimed to determine the range within which the true population mean is likely to fall. Additionally, the task involved comparing the widths of the two intervals and explaining the reasons for differences.

Steps Taken

To begin, I utilised only the data for the TV ads provided from the Excel file. Since the population standard deviation (σ) is unknown, I used the t-critical formula to calculate the confidence intervals because the t-distribution is appropriate when working with sample data and unknown population parameters. (University, Confidence Intervals Lecture Notes, n.d.)

The first step was calculating the sample mean ($\bar{x}$) and the sample standard deviation (s). I used the functions =AVERAGE(A2:A201) to calculate the sample mean (Tutor, How To Calculate The Average In Excel, 2020) and =STDEV.S(A2:A201) function to calculate the sample standard deviation of the TV ads expenditure. (Jardin, 2020). This resulted in sample mean ($\bar{x}$) being 147.0425 and sample standard deviation (s) being 85.85424 (Both in Hundreds of pounds).

I worked out the t-critical values by looking them up on the t-distribution lookup sheet using the degree of freedom (df) of 199, which was calculated from the sample size of 200 (df = n - 1). The t-critical value for a confidence level of 95% came out to be 1.972 and 2.626 for 99% confidence. Next, I needed to calculate the margin of error (ME) for both the 95% and 99% confidence intervals. I used the mathematical formula for ME so I could create a excel formula to do same operations =P8*(M9/SQRT(M10)). (Guide, 2022). The mathematical formula is shown below.

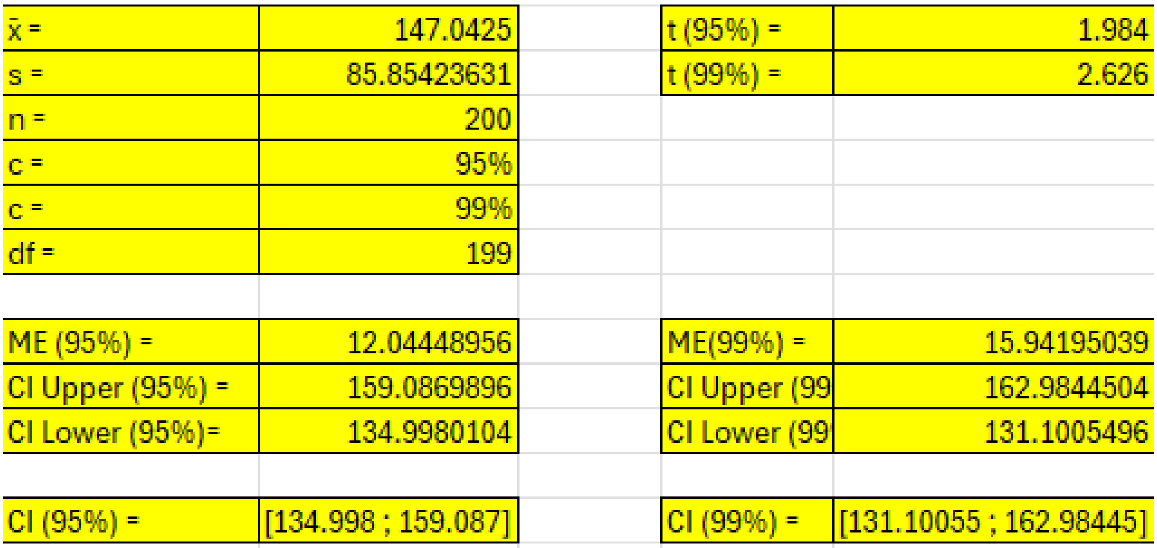
$$ME = t_{\frac{a}{2}} \cdot \frac{s}{\sqrt{n}}$$

Finally, I calculated the confidence intervals by adding and subtracting the margin of error from the sample mean (147.0425) by using a simple excel formula =M8+M15 and =M8-M15 (Plus and minus sample mean by ME of 95%) and =M8+P15 and =M8-P15 (Plus and minus sample mean by ME of 99%) so I can get both upper and lower bounds for both intervals.

$$\bar{x} \pm ME$$

The calculated intervals came out to be [134.998; 159.087] with 95% confidence and [131.10055; 162.98445] with 99% confidence. You can find the Excel work on the sheet tab called Task 4 (Confidence Intervals).

Results



x̄ =	147.0425	t (95%) =	1.984
s =	85.85423631	t (99%) =	2.626
n =	200		
c =	95%		
c =	99%		
df =	199		
ME (95%) =	12.04448956	ME(99%) =	15.94195039
CI Upper (95%) =	159.0869896	CI Upper (99	162.9844504
CI Lower (95%)=	134.9980104	CI Lower (99	131.1005496
CI (95%) =	[134.998 ; 159.087]	CI (99%) =	[131.10055 ; 162.98445]

Figure 4. Calculated Confidence Interval Variables in Excel

The 95% confidence interval suggests that we are 95% confident that the true mean of TV advertising expenditure and sales falls between 134.998 and 159.087 (Hundreds of Pounds) and the 99% confidence interval suggests that we are 99% confident that the true mean falls between 131.10055 and 162.98445 (Hundreds of Pounds).

The 99% confidence interval is wider than the 95% interval as expected. The margin of error increases as the confidence level increases because a higher confidence level implies that we need to account for more variability in the population thus resulting in a broader range.

Conclusion

The analysis shows that the 95% confidence interval is [134.998; 159.087] and the 99% confidence interval is [131.10055; 162.98445]. This means that we are 95% and 99% confident that the true population mean falls within these ranges. As expected, the 99% confidence interval is wider than the 95% interval due to the higher level of confidence which leads to a larger margin of error.

Task 5 (Hypothesis Testing) – Aliyah Zaman

Steps Taken

To begin, the first thing I did was define my hypotheses. My null hypothesis was that the average sum spent on TV advertising by the firm's clients is £14,000 whereas my alternate hypothesis was that the average sum spent on TV advertising by the firm's clients is not £14,000. I gathered all my data and organised it in such a way that it was clearly legible and easy to compute. Sample size (N): 200. Sample Mean (x̄): £14,704 (147.04 in hundreds of pounds). Sample Standard Deviation (S): £8,585 (85.85 in hundreds of pounds). Hypothesized Population Mean (μ_0): £14,000 (140 in hundreds of pounds)

Since I didn't know the population standard deviation, I had to use a t-test instead of a z-test. (University, Lecture Notes Chapter 9, n.d.). I had later calculated the test statistic using the formula aligned with the T-test and chose the variables from my data to work this out.

The test statistic (t) informed me of how many standard errors the sample mean is away from the hypothesised population mean. In this case, the test statistic is approximately 1.16. This value was used in the next steps to determine whether to reject the null hypothesis.

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{147.04 - 140}{\frac{85.85}{\sqrt{200}}} = 1.16$$

For a t-test, the degrees of freedom are calculated as DF = N - 1 where N is the sample size. So, the degrees of freedom (DF) would be DF = 200 - 1 = 199. With 199 degrees of freedom, I used this number to look up the critical t-value from the t-distribution table. The t-distribution table provides the values that correspond to the significance level we chose (0.05) and our calculated degrees of freedom. Using the degrees of freedom (DF = 199) and a chosen significance level (0.05 for a two-tailed test), I refer to the t-distribution table. For DF = 199 and significance level of 0.05, the critical t-value is approximately 1.96.

Next, I compared the value of the test statistic to the critical value. If the test statistic is greater than the critical value, I reject the null hypothesis. And to make the decision, I had to analyse the comparison. Finally, I had to draw a conclusion and either decide whether I was going to reject or fail to reject the null hypothesis. As our test statistic is lower than our critical value (1.16 < 1.96), we fail to reject the null hypothesis.

Conclusion

To conclude**,** we found that our test statistic of 1.16 did not exceed the critical value of 1.96, this therefore led to me failing to reject the null hypothesis indicating that the observed sample mean (£14,704) is not significantly different from the hypothesized population mean (£14,000). Furthermore, the hypothesis test supports the CEO’s claim and provides a sound statistical foundation for the firm to trust in their current understanding of client advertising expenditures.

Conclusions

The analysis revealed several significant insights into the relationship between TV advertising expenditure and product sales. Firstly, there is a moderate positive linear relationship between the TV advertising expenditure and sales which was heavily indicated by the correlation coefficient of 0.7822.

This highlights that increased investment in TV advertising tends to result in sales. Secondly, the coefficient of determination (61.19%) indicates that a substantial proportion of the variation in sales can be explained by TV advertising expenditure although other factors account for the remaining 38.81%. Additionally, the confidence interval analysis showed that the average TV advertising expenditure falls between £13,110 and £16,298 with 99% confidence validating the claim that advertising investments have a predictable range.

Key Learnings

We gained hands-on experience in applying statistical techniques such as linear regression, residual analysis, and confidence interval estimation. These techniques helped deepen our understanding of how data-driven decisions can be made in real-world scenarios.

The process improved our efficiency in Excel, especially in using built-in functions for regression analysis, scatterplot creation, and calculating confidence levels.

Challenges

The group encountered challenges due to varying levels of Excel proficiency. Some team members performed calculations manually instead of utilising automated functions due to lack of Excel knowledge which caused some delays.

However, this became an opportunity to learn and collaborate as a team. In terms of project management, our reliance on online communication through WhatsApp allowed flexibility and highlighted the importance of clear and consistent updates to prevent misunderstandings.

References

Academy, E. T. (2021, March 15). *How to Make a Scatter Plot in Excel*. Retrieved from YouTube: https://www.youtube.com/watch?v=MfEAEmdFOBo

Bradburn, S. (2021, March 15). *Adding The Trendline, Equation And R2 In Excel*. Retrieved from YouTube: https://www.youtube.com/watch?v=JoLpTefCIzk

Freeman, A. (2020, October 25). *Using Excel - Creating a Residual Plot*. Retrieved from YouTube: https://www.youtube.com/watch?v=EHnkc6evkiQ

Guide, H.-T. (2022, February 23). *How to Calculate Square Root in Microsoft Excel :Tutorial*. Retrieved from YouTube: https://www.youtube.com/watch?v=DCgSynk_hSQ

Jardin, D. E. (2020, July 23). *Calculate Standard Deviation in Excel (30 Seconds)*. Retrieved from YouTube: https://www.youtube.com/watch?v=KcCA061tpyk

Learning, E. (2017, April 13). *Adding Sheets in Excel*. Retrieved from YouTube: https://www.youtube.com/watch?v=EPJ2DdsLaoI

Sharpe, M. (2019, October 31). *Calculating Correlation Coefficient Excel*. Retrieved from YouTube: https://www.youtube.com/watch?v=Nc0cTp4UdBk

Tutor, T. O. (2018, May 29). *Excel Basics - Linear Regression - Finding Slope & Y Intercept*. Retrieved from YouTube: https://www.youtube.com/watch?v=KwQsV77bYDY

Tutor, T. O. (2020, June 18). *How To Calculate The Average In Excel*. Retrieved from YouTube: https://www.youtube.com/watch?v=1xD_pdAnU0c

University, C. S. (n.d.). *Confidence Intervals Lecture Notes*. Retrieved from Chapter 8.2.2: https://moodle4.city.ac.uk/pluginfile.php/1194761/mod_resource/content/1/Chapter%208%20proba-68-75.pdf

University, C. S. (n.d.). *Lecture Notes Chapter 9*. Retrieved from Hypothesis Testing 9.2.2: https://moodle4.city.ac.uk/pluginfile.php/1194769/mod_resource/content/1/Chapter%209%20proba-76-81.pdf