Structural features

Communication features

Management Studies – Data Analysis Report

An analysis of the market penetration data of a satellite TV company

Summary

This report **presents** a statistically valid model to predict market penetration for a family-oriented TV channel from a number of predictors given in a data set. Firstly, **we** explain the validity of the data set and point out which factors should be excluded in order to develop an accurate model. After having developed a significant, but rather complex, model for predicting market penetration, we suggest that it can be simplified to save data collection costs, whilst still giving accurate predictions. Therefore, the least value adding factors were removed, allowing us to establish a statistically relevant model that is dependent on two factors: computer ownership and education. After validating the model, it is applied to predict market penetration for five new areas and five hypothetical areas, giving further implications for the TV channel's strategy by having identified target audience. Finally, building on the findings from using the model, we make recommendations on which areas should be addressed in order to maximise market penetration, and which areas should not be prioritised.

In business, abstracts can be called summaries or executive summaries, but they perform the same function to summarise the entire contents of the report. See Ch 2, Getting started on your lab report

Present simple tense can be used in abstracts and summaries

Use of personal pronoun 'we' can be acceptable in data analysis reports, when it is a group assignment. However, you need to check with your tutor.

Table of Contents

Data Analysis and other technical reports can include a table of contents. You should check with your tutor.

Introduction

Market penetration in specific areas of the country is an important issue to consider for TV channels. Identifying which factors influence their market penetration will have great

A short, succinct introduction is common in data analysis reports. The introduction identifies the rationale for the analysis, who the report is for, and the aims of the report in the context of its usefulness to the organisation. It follows a standard, essay style structure

implications, since a TV channel can accordingly decide which areas can be targeted to maximise market penetration, for example by advertising campaigns, which could in turn help them achieve a larger profit margin. This report has been produced for the CEO, as well as the Finance, HR and Operations Department of a family-oriented TV channel in order to explain how they can predict their market penetration in specific areas. By analysing a given data set, it provides an analysis of the market penetration data for the TV channel and establishes a statistical model that takes into account a number of different factors[1] determining market penetration. The aim of this report is to provide the TV channel with an appropriate model that can be used to predict their market penetration.

Analysis and Interpretation

Headings are acceptable and desirable in reports, as they help navigation and understanding

a. Assessing the validity of the data set

For the purpose of this analysis, data for 33 trial areas were given. These include information about the numbers of households, the social class (here the classification is affluent, prosperous, comfortable, stretched and needy), the percentage of computer ownership, the average income, the percentage of well-educated people and the household size. Those factors are referred to as predictors for the remainder of this report, as they can potentially predict market penetration. Importantly, for area 32, there seems to be a data collection error since there is no data for the social class. Equally, for area 33, the percentages for the social class add up to 110% instead of 100% showing another collection error. We therefore decided to exclude area 32 and 33 from this analysis in order to ensure an appropriate result.

Mix of past simple (passive) and present tenses. Past is used to describe the experiment that was carried out. Present is used to describe findings.

Use of passive is common in scientific reports, as the action is more important than the actor.

b. Removing predictors to ensure accuracy

No source required for the Figure as this is the students' own work

As a first step, when building a statistical model, it has to be determined if any of the predictors are strongly related to each other, since that would have a significant effect on the statistical model we try to build. For example, one could assume that education and average income are related, since a higher level of education, could mean a higher average income. This is the same for the social class: People who are expected to belong to a high social class, in this case 'affluent', are expected to have a higher income.

Footnotes used to help explain complex terms

[1]This is called 'multiple linear regression' in statistics.

For this analysis, using statistical measures, it was examined if any of the predictors have a worryingly high relationship to each other in order to detect if any predictors have to be excluded for a statistically significant model, i.e. a model that is suitable from a statistical point of view. In contrast to the expectations previously stated, the only two predictors that were found to potentially cause significant deviations due to their relationship, were the 'properous' social class and computer ownership[2]. Their relationship is illustrated in Figure 1 with the line showing that P tends to increase as Computer Ownership increases. Further, the percentages for each social class are related to each other[3]. Using statistical reasoning, it was therefore decided to exclude the social class P from the model in order to simultaneously mitigate both potential problems that may arise from the identified relationships above[4].

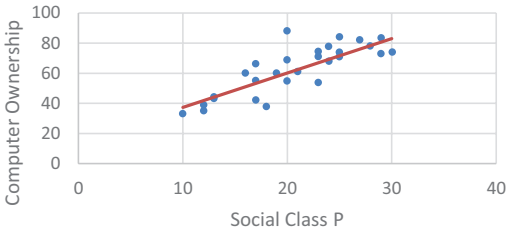

Figure 1: Relationship between Social Class P and Computer Ownership

c. Establishing a model with many predictors

After having ensured that the different predictors are not highly related to each other, the remaining predictors can be used to create a model for forecasting market penetration. Using statistical analysis in Excel[5], a formula was derived that can be used to predict market penetration from the remaining predictors (social class A, C, S and N; Computer Ownership; average income; education; household size). For each area, the market penetration can be predicted using the following formula:

---

[2]Correlation = 0.82 > 0.75 benchmark. This means significantly high correlation, sign for multicollinearity.
Correlation between education and average income = 0.64 < 0.75 benchmark; correlation between affluent and income = 0.58 < 0.75 benchmark. This means relationship is not worryingly high.

[3]Percentages of affluent, prosperous, comfortable, stretched and needy have to add up to 100%.

[4]Excluding P from the model solves both the issue of multicollinearity with computer ownership and the relationship between the social classes. Moreover, adjusted R2 for the model excluding P is 0.73, whilst it is 0.69 when only computer ownership is excluded. This further proves that it is more sensible to exclude P from the model.

[5]Data analysis tool pack

*Market Penetration in % = -15.04 – 0.0003**__no. of households__ *+ 0.39*% of* __affluent households__ *+ 0.61*% of* __comfortable households__ *+ 0.63*% of* __stretchy households__ *+ 0.8*% of* __needy households__ *– 0.03*% of* __computer ownership__ *+ 0.22*__*average income (in £1000)* *– 0.46*% of* __education level__ *+ 8.18*__*household size*

Writer provides a clear and succinct interpretation of results

This highly complex formula can be interpreted as follows: If the percentage of affluent household increases by 1%, other predictors remaining the same, the market penetration will increase by 0.39%. On the other hand, when the percentage of computer ownership increases by 1%, other predictors remaining the same, the market penetration will decrease by 0.03%, etc.

Using this model, Figure 2 shows the predicted and the actual market penetration for each area.

According to Figure 2, the percentages of market penetration were predicted relatively accurately. However, this can also be tested by two different statistical measures. It was found that 73% of the deviations can be explained by this model[6]. Further statistical analysis showed that the model predicts the actual data well[7]. Therefore, based on statistical evaluation, the model represents a good fit of the actual data.

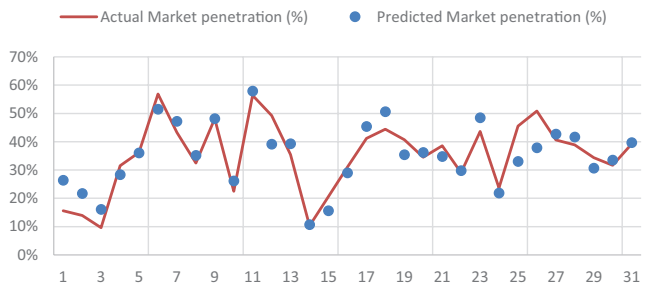

Figure 2: Predicted and Actual Market Penetration in % for each area

Clear, simple, and easy to read figures, with captions (below)

d. Removing the least value adding predictors

Although the model in Section 2.3 predicts market penetration accurately, it is highly complex and over-complicated. To mitigate this complexity, there is an approach commonly used

[6] $R^2_{adjusted} = 0.73$

[7] From ANOVA: F statistic = 9.83 and $F_{0.5,9,21} = 2.37$, 9.83 > 2.37, so regression is significant.

in statistics to test whether specific predictors add significantly to the accuracy of the model or if they can be excluded. The aim of this process[8] is to simplify the model by reducing the number of predictors, but at the same time ensuring that a good fit of the actual data is still obtained. As a result, costs of data collection can be reduced since less data would be needed in order to make predictions.

Starting with the model from Section 2.3, the predictor that has the weakest relationship with market penetration is excluded and using statistical measures, it will be decided if it significantly adds to the accuracy of the model. If it does not, this process is repeated for each predictor up to the point, when a predictor is found that does make the model significantly more accurate. Using Excel, it was identified that the number of households has the weakest relationship with market penetration, and it does not add significantly to the accuracy of the model[9]. This is similarly the case for the following other predictors: household size[10], stretched social class[11], comfortable social class[12], affluent social class[13], average income[14] and needy social class[15]. The process concludes that both computer ownership[16] and education level are predictors that are worth to be included in the model.

e. Establishing the reduced model

According to the findings from Section 2.4, the model has been reduced and is just dependent on two predictors. Using Excel, the following simplified formula can be established to predict market penetration:

---

[8]This process is called 'Backward Elimination' in Statistics.

[9]Correlation with market penetration = -0.05; T statistic from ANOVA = 0.43, $F_{0.5,1,22} = 4.32$. T<F, so not worth including in model.

[10]Correlation with market penetration = 0.07; T statistic from ANOVA = 2.53, $F_{0.5,1,22} = 4.30$. T<F, so not worth including in model.

[11]Correlation with market penetration = 0.18; T statistic from ANOVA = 1.54, $F_{0.5,1,23} = 4.28$. T<F, so not worth including in model.

[12]Correlation with market penetration = -0.24; T statistic from ANOVA = 0.02, $F_{0.5,1,24} = 4.26$. T<F, so not worth including in model.

[13]Correlation with market penetration = -0.38; T statistic from ANOVA = 0.10, $F_{0.5,1,25} = 4.24$. T<F, so not worth including in model.

[14]Correlation with market penetration = -0.43; T statistic from ANOVA = 3.22, $F_{0.5,1,26} = 4.23$. T<F, so not worth including in model.

[15]Correlation with market penetration = 0.57; T statistic from ANOVA = 0.85, $F_{0.5,1,27} = 4.21$. T<F, so not worth including in model.

[16]Correlation with market penetration = -0.58; T statistic from ANOVA = 13.59, $F_{0.5,1,28} = 4.20$. T>F. It is worth including this predictor in model.

Market Penetration in % = 76.99 – 0.30% of* ***computer ownership*** *– 0.45*% of* ***education level***

Effectively, this means, that if computer ownership increases by 1%, percentage of education level remaining the same, market penetration will decrease by 0.3%. Similarly, if education level rises by 1%, computer ownership remaining constant, market penetration will fall by 0.45%. Hence, the higher the percentage of computer ownership and education level is in a specific area, the lower will be the market penetration for the TV channel in that area. This relationship is further illustrated in Figure 3: For instance, the blue line shows that as computer ownership increases, education remaining the same, market penetration will decrease. Notably, education has a larger impact on market penetration than computer ownership, which is illustrated by the red line in Figure 3 being steeper than the blue line.

Further statistical evaluation of results to ensure accuracy and reliability

f. Validating the model

Having established a model that uses computer ownership and education to predict market penetration, **it is important to evaluate if the model is correct and accurate.** Again, two statistical measures are used to identify if this is given. It can be found that 71% of the deviations can be explained by this model[17]. Although this value is less that the one that we found for the more complex model, it is still sufficiently high and the model predicts the actual data well[18]. The 2% change in this value is sacrificed in order to obtain a more simplified model that will incur less data collection costs.

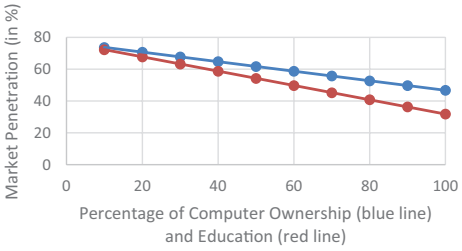

Figure 3: Illustration of the Model

When analysing data using this specific statistical procedure, one has to make assumptions before being able to create a model[19]. It is essential to test these assumptions once the model was found in order to identify if the model is appropriate and can actually be used. The first assumption that is made is that there is a linear relationship between each of the two predictors and market penetration. In order to test this, one has to plot the so-called 'residuals' from each predictor, which is the

[17] $R^2_{adjusted} = 0.71$

[18] From ANOVA: F statistic = 37.57 and $F_{0.5,2,28} = 3.34$, 37.57 > 3.34, so regression is significant

[19] Full analysis is based on the assumption that observations are independently and identically distributed.

difference between the predicted market penetration and the actual market penetration for each area, against the predicted values. If this assumption holds true, no pattern should be identified in either of these plots. As can be seen in Figure 4 and 5, there is no pattern and thus the first assumption is confirmed.

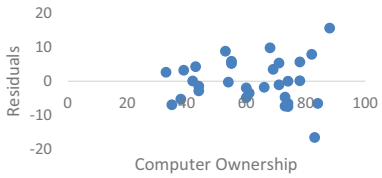

Figure 4: Test for Assumption 1

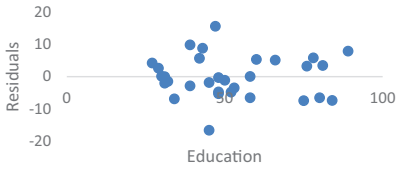

Figure 5: Test for Assumption 1

The second and third assumption[20] can be tested in a similar way: One needs to generate further plots. If no pattern can be seen from the plotted data, it can be said that the assumptions made, in order for the analysis to be accurate, are true. Figure 6 and 7 do not reveal any patterns, which proves that the assumptions hold.

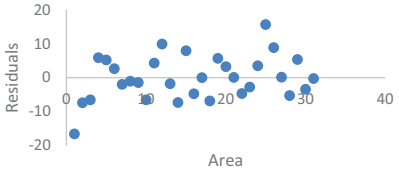

Figure 6: Test for Assumption 2

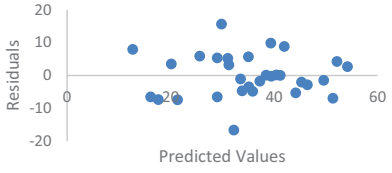

Figure 7: Test for Assumption 3

[20] 2nd Assumption: Independence of Errors, 3rd Assumption: Equal Variances

For the fourth assumption, one would expect to see a straight line if it is true. As can be seen in Figure 8, the data presents a line that is relatively straight, so this assumption is true as well. Hence, it can be said, that this model is validated and can be used in an appropriate setting.

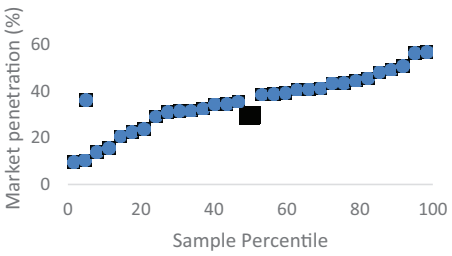

Figure 8: Test for Assumption 4

g. Applying the model

Application of findings to real world setting

Having validated this model, it can now be applied to the areas given in order to predict market penetration. From Figure 9, we can see the predicted market penetration for the new areas 1 to 5. Notably, area 1 and 5 have the highest predicted market penetration. For area 1, the reason for this high prediction is the low level of education (35%), whereas for area 5 it is due to the relatively low level of computer ownership (60%). Applying the same model to hypothetical areas (Figure 10), each containing 100% of households in one of the 5 social classes, there is a clear trend for increasing market penetration, as the level of social class decreases (from affluent to needy). This can be explained because both the percentage of computer ownership and the percentage of education decrease moving from the affluent social class to the needy social class in those hypothetical areas. The social class seems to amplify the effect that level of computer ownership and education have. Although the social class is not a predictor for the simplified model, there seems to be a relationship between the social class and the two predictors, reflecting information about the social class. This is in line with Section 2.2 where a relationship between the prosperous social class and computer ownership was found.

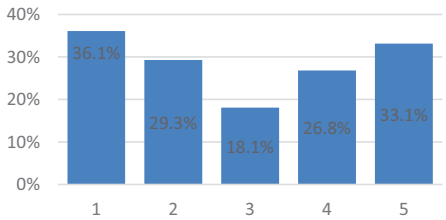

Figure 9: Predicted Market Penetration in % (for new areas 1-5)

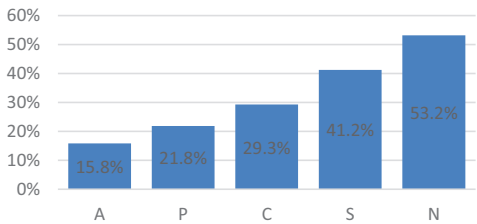

Figure 10: Predicted Market Penetration in % (for hypothetical areas)

h. Interpretation and recommendations

The report includes critical analysis and interpretation of results, and subsequent (relevant and justified) recommendations

The analysis has shown that market penetration for the TV channel is larger in areas with less computer ownership and less education. There are a few possible explanations for this: If computer ownership is higher is one specific area, people

might spend more of their spare time using a computer, rather than watching TV. Further, if there is an internet TV channel that covers similar family-oriented content, people may prefer watching it on their computer rather than the TV. Addtionally, people with higher levels of education may prefer to watch a different TV channel. This refers back to personal preference. It may be assumed, therefore that the programmes shown on the family-oriented TV channel do not appeal to well-educated people due to their content.

Analysing the hypothetical areas shows that further assumptions can be made about market penetration with regards to social class. Indeed, as pointed out in Section 2.2, there is a relationship between the prosperous social class and computer ownership, showing that, to a certain extent, market penetration may be dependent on social class as well. From the hypothetical areas one could therefore recommended the TV channel to address the stretched and needy social class. **However**, one has to be careful with those assumptions as the clear relationship between social class and market penetration only becomes apparent from the hypothetical areas which may not reflect reality, and thus the real data set. Moreover, the analysis of the data did not show that the social class adds significantly to the accuracy of the model.

'However' indicates qualification of first interpretation/ assumption. This is good evaluation of results and implications.

Therefore, it can be said that generally the TV channel **should** address areas with lower education levels and lower computer ownership levels in order to maximise market penetration, and prioritise areas with high levels of education and computer ownership less. Regardless of maximising the market penetration, it is also relevant for the TV channel to address an area with a large number of households. Even if market penetration is high in one specific area, it might only consist of a small number of households. This may mean that less households are addressed than in an area with a large number of households and low market penetration. Hence, the number of households is important when it comes to addressing revenue.

Modal verb (should) used to emphasise recommendations

Limitations

There are a number of limitations that need to be considered and **may** restrict the validity of the results from this analysis.

Cautious language is used to emphasise the possible limits of the study

- **The amount of data given is not extensive. Data for more than 33 areas should have been collected.**

- The fact that 2 out of the 33 areas contained data collection errors **could mean** that there were more errors made for the rest of the data set but which are not as easily detected. This **could have** further implications for the analysis.
- There may be other predictors that influence market penetration, that no data was collected for. For example, it would be interesting to have gender information or percentage of TV license ownership in order to establish a more accurate model.

Use of bullets can be acceptable in both business and scientific reports. Check with your tutor.

Conclusion

This report has aimed to provide a model that can be used to predict market penetration of a family-oriented TV channel by area. Working with a data set that was provided, we have firstly excluded factors that did not add significantly to the accuracy of the forecast. Using statistical analysis, we then derived a model that uses the percentage of computer ownership and education as the two factors that can give useful predictions. To support our findings, we have validated the model using several statistical tools which have proven the validity, relevance and accuracy of our model. Consequently, this model can be confidently used by the TV channel to forecast market penetration. Using this model, we then predicted the market penetration for 10 different areas and recommended the TV channel to address areas with low levels of computer ownership and education to maximise market penetration. The report concludes by pointing out limitations.

Clear, succinct , essay style conclusion summarising the report

Diana and Tom's Comment

This data analysis report provides a thorough statistical analysis of the focus topic. The report is well-structured and easy to follow. In addition, data collection and results analysis is in-depth, with a good range of clear, succinct, and accurate figures used. The writers' identify a number of significant findings which they go on to evaluate and interpret with critical insight and originality. While incorporation of sources to support their ideas and recommendations may have strengthened the report, (though the assignment may not have required them to include these), their recommendations provide a meaningful and highly useful contribution to the topic area.

No reference list. Some reports do not require this. Check with your tutor.