*Exam introduction to modelling*

## Sebastiaan de Jong

## 613234

## Minor data driven decision making

# **Exam introduction to modelling**

Version: 1.0

Date: 22-10-2021

Exam code: MDDFIML1A.5

###### ***Student:***

Name: Sebastiaan de Jong

Studentnumber: 613234

Major: Logistics management

E-mail: ST.dejong@student.han.nl

###### ***Examiner:***

Name: Tijmen Weber

E-mail: Tijmen.Weber@han.nl

###### **Educational institution:**

Institution name: Hogeschool van Arnhem en Nijmegen

Location: Ruitenberglaan 31, 6826 CC Arnhem

# Table of contents

[Chapter 1: preface 3](#_Toc85825797)

[Chapter 2: Question 1 4](#_Toc85825798)

[Chapter 3: Question 2 7](#_Toc85825799)

[Chapter 4: Question 3 9](#_Toc85825800)

[Chapter 5: Question 4 16](#_Toc85825801)

[Chapter 6: Question 5 18](#_Toc85825802)

[Chapter 7: Question 6 21](#_Toc85825803)

[Bibliography 25](#_Toc85825804)

# Chapter 1: preface

This is the exam made by Sebastiaan de Jong. The accompanying r scripts can be found as a github repository at <https://github.com/DeshagoNL/IML-exam> This can be loaded as an R project.

All used packages are combined in libraries.r. Each question has its own R script, named question 1.r, question 2.r, etc.

# Chapter 2: Question 1

## Description of data and cleaning process

The data for this question is collected from a Portuguese school and contains data about the alcohol consumption, sex, age, address, romantic status, status of family relations, health, absences, extra-curricular activities, and grades of a selection of students.

There are no missing values.

For the categorical variables activities and romantic status dummy variables are made, where 1 equals yes and 0 equals no. For the variable address a dummy variable is made where 1 equals a rural address and 0 equals an urban address. This way the correlation between the variables can be checked more easily.

## Description of the regression analysis

For this question, the goal is to see if address, age, romantic status or extra-curricular activities have an effect on alcohol consumption. Therefore, alcohol consumption is used as dependent variable and address\_Dummy, age, romantic\_Dummy and activities\_Dummy are used as independent variables.

Variables are not standardized, as it is only important which variable has an effect and the effect of different variables aren’t compared.

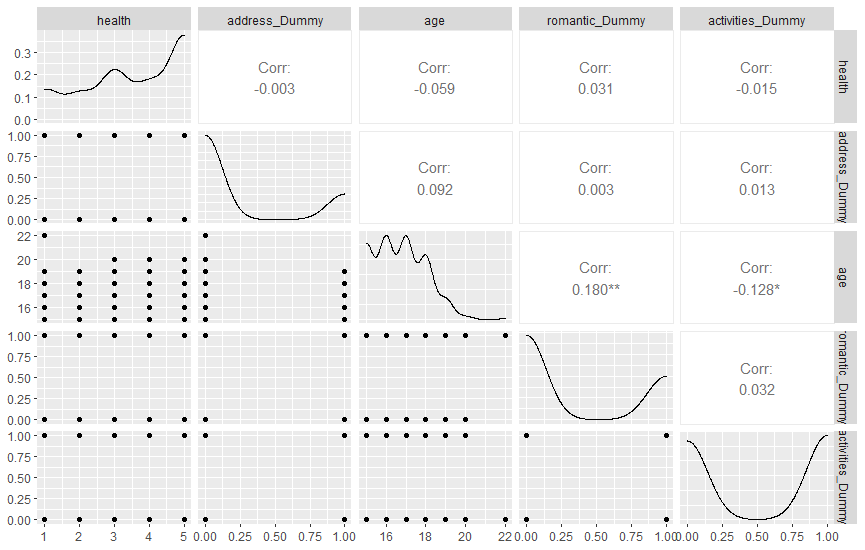
Before running the analysis, the variables were checked for multicollinearity. As all the correlations are below 0.8, it can be concluded that there is no multicollinearity.

Figure 1.1: Correlation between the independent variables

## Regression model

Creating a regression model for these variables gives the following result

**Table 1.1: variables affecting alcohol consumption at Portuguese school**

|  |  |  |  |
| --- | --- | --- | --- |
|  | All students | Male students | Female students |
| (Intercept) | 0.09 | -1.57 | 2.81 \* |
|  | (1.03) | (1.57) | (1.22) |
| health | 0.07 | 0.09 | -0.02 |
|  | (0.05) | (0.09) | (0.06) |
| address | 0.17 | 0.26 | 0.05 |
|  | (0.18) | (0.29) | (0.20) |
| age | 0.12 \* | 0.24 \*\* | -0.04 |
|  | (0.06) | (0.09) | (0.07) |
| romantic status | -0.22 | 0.00 | -0.19 |
|  | (0.16) | (0.27) | (0.18) |
| activities | -0.08 | -0.33 | -0.06 |
|  | (0.15) | (0.25) | (0.17) |
| R^2 | 0.03 | 0.09 | 0.01 |
| Adj. R^2 | 0.01 | 0.05 | -0.02 |
| Num. obs. | 300 | 139 | 161 |

\* = p < 0.05, \*\* = p < 0.01, \*\*\* = p <0.005. Standard errors are between brackets. All variables are standardized.

## Explanation of the results

The wildly differing intercepts suggest that there is a difference in the effect of the variables on alcohol consumption. Where boys consume more alcohol if all the variables increase the same amount, girls actually consume less alcohol if all the variables increase the same amount.

If we look at the mean of alcohol consumption, it shows that the mean of alcohol consumption for boys is 2.7, and for girls is 1.9. Therefore, boys consume more alcohol on average

The only statistically significant value in this table is age. If you only look at all students, it will seem that students drink more as they get older. However, if you look closer it shows that age only has a significant positive effect for boys. For girls age has almost no influence on alcohol consumption.

The data also shows that rural boys consume more alcohol than those with an urban address, whereas for girls address makes almost no difference. However, this result is not statistically significant, thus should not be relied upon. The same goes for activities and romantic status, with boys that have extra-curricular activities consuming less alcohol, whereas for girls it has next to no impact and girls that have a relationship consuming less alcohol, whereas for boys this has no significance. These results however have no statistical significance, so these conclusions should be taken with a grain of salt.

# Chapter 3: Question 2

## Description of data and cleaning process

This data file contains data on books and is gathered from amazon.com there are 270 entries which all list the title, author, listprice, amazonprice, if it’s a hardcover or paperback, number of pages, publisher, publishing year, isbn nr, height, width, thickness, and weight of the books.

There is 1 missing value. As this only pertains to a single row, it is decided to remove this row from the data.

For this question the following variables are used: AmazonPrice, Height, HardOrPaper, NumPages, Weight, Thick

First, a dummy variabls is made for HardOrPaper, where Hard is 1 and Paper is 0.

Then, AmazonPrice, Height, NumPages, Weight and Thick are standardized.

## Description of the regression analysis

The variables are checked for multicollinearity. As shown in figure 2.1, there is a high correlation between thickness and number of pages. As thickness is usually a result of a combination of the number of pages and if it is a hardcover or paperback, it is decided to leave thickness out of the model.

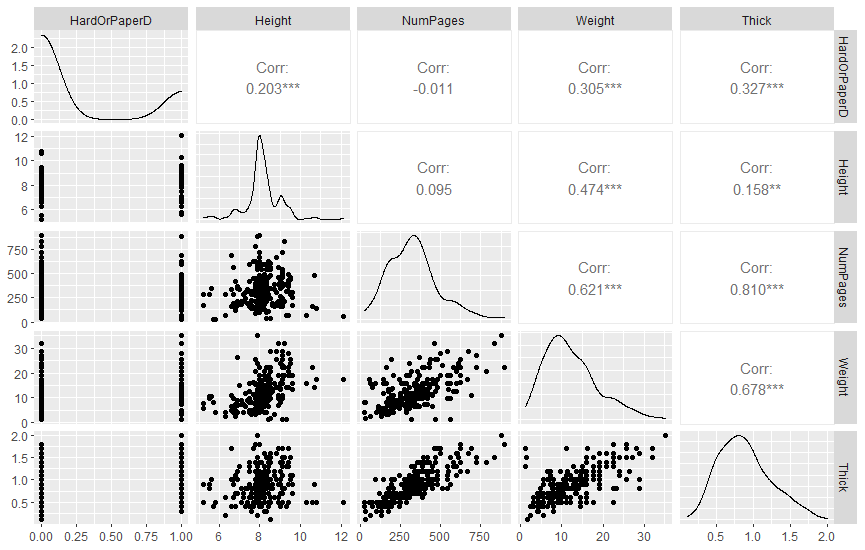


Figure 2.1: Correlation between the independent variables

## Regression model

**Table 2.1: variables affecting book price on Amazon.com**

|  |  |
| --- | --- |
|  | Effect on Amazon Price |
| (Intercept) | 0.08 |
|  | (0.06) |
| Hardcover or Paperback | -0.30 \* |
|  | (0.14) |
| Height | 0.35 \*\*\* |
|  | (0.06) |
| Number of Pages | -0.10 |
|  | (0.08) |
| Weight | 0.23 \* |
|  | (0.09) |
| R^2 | 0.21 |
| Adj. R^2 | 0.20 |
| Num. obs. | 269 |

\* = p < 0.05, \*\* = p < 0.01, \*\*\* = p <0.005. Standard errors are between brackets. All variables are standardized.

## Explanation of the results

According to the model, the height of the book has an extremely statistically significant effect on the amazon price, where books that are one standard deviation bigger are also 0.35 standard deviation more expensive. This makes sense, because books that have more height have more have more words per page and need more material to make, especially when controlled for number of pages as in this model.

The weight also impacts the amazon price in a statistically significant way, with heavier books costing more. This also makes sense for the same reason as height.

Strangely enough, Hardcovers on average cost less than paperbacks once controlled for the other variables. This is very strange, as usually hardcovers of the same book cost more than paperbacks. I am not completely sure why, but I think it might be because you play less for the extra height and weight of the hardcovers than you would when the book is higher or heavier normally.

# Chapter 4: Question 3

## Description of data and cleaning process

The data is data about subscriptions to economic journals at US libraries for the year 2000. The data contains the following variables: Title, Abbreviation, Publisher, if it is part of the Society, Price, Pages, Characters per page, number of citations, founding year, number of subscriptions and the field it is about.

There are no missing values

Price, Characters per page, number of citations and founding year are all used and compared in this model, so they are standardized.

## Description of the regression analysis

This model will be used to check if the price of the journal, characters per page, number of citations and founding year have an impact on the number of subscriptions the journal has.

First, the independent variables are checked for multicollinearity. Figure 3.1 shows that there is no large overlap between the different independent variables, so there should be no issues with multicollinearity.

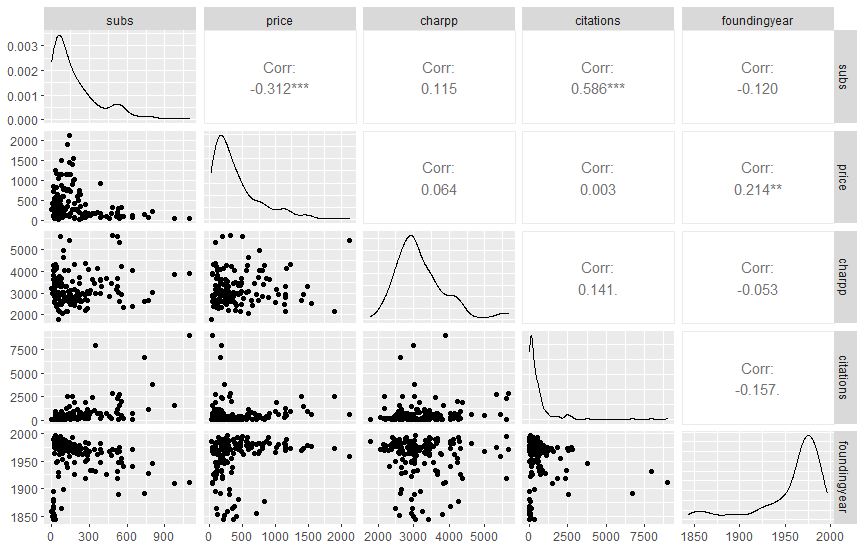


Figure 3.1: Correlation between independent variables

There is a suspicion that there is a non-linear relationship between the variables. To check this, the variables have been graphed with a loess curve.

Figure 3.2 suggests that there is a logarithmic relationship between price and subscriptions. If we analyze the fit, it is shown that the fit without transformation has an r2 of .10, when transformed for a logarithmic relationship it has a fit of .21 and when transformed for a poly relationship, it has a fit of .25. This shows that there is a 3rd degree polynomial relationship between price and number of subscriptions.

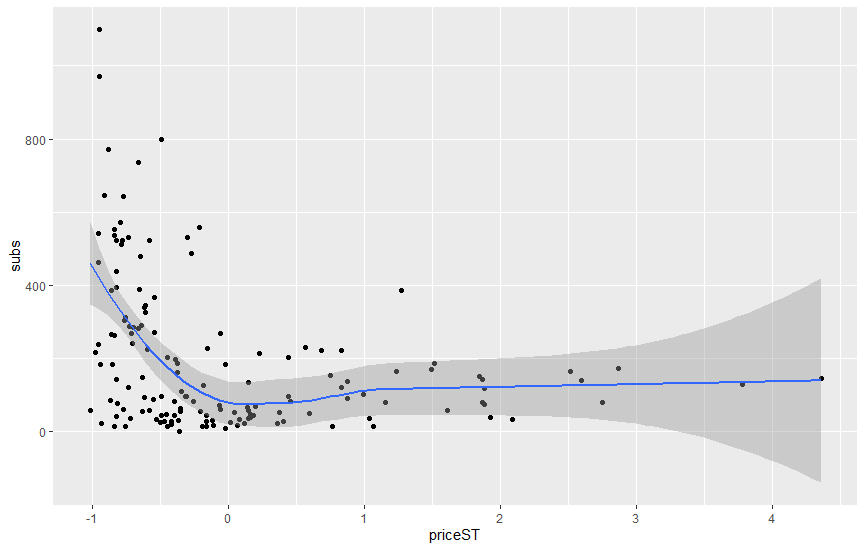


Figure 3.2: Relationship between price and subscriptions

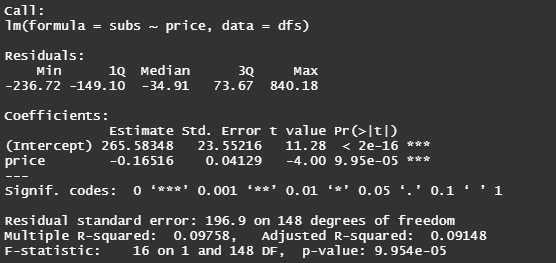


Figure 3.2.1: Fit without transformation

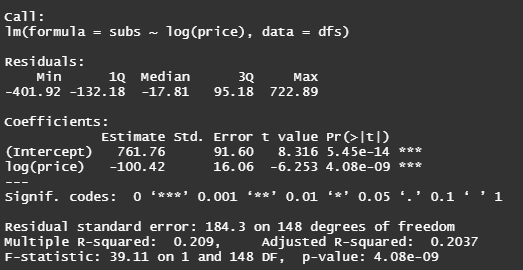


Figure 3.2.2: Fit when transformed for logarithmic relationship

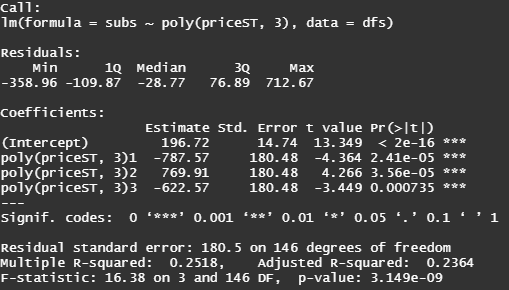


Figure 3.2.3: Fit when transformed for 3rd degree poly relationship

Figure 3.3 shows the relationship between characters per page and number of subscriptions. The current r2 is 0.01. If we transform it using a 3rd degree polynomial, we can get the r2 to 0.017, and with a 5th degree polynomial to 0.029. However, both these transformations have a negative adjusted r2, meaning they improve the model less than would be expected by adding more independent variables (Investopedia, N.D.). Therefore, no transformation will be applied to the model.

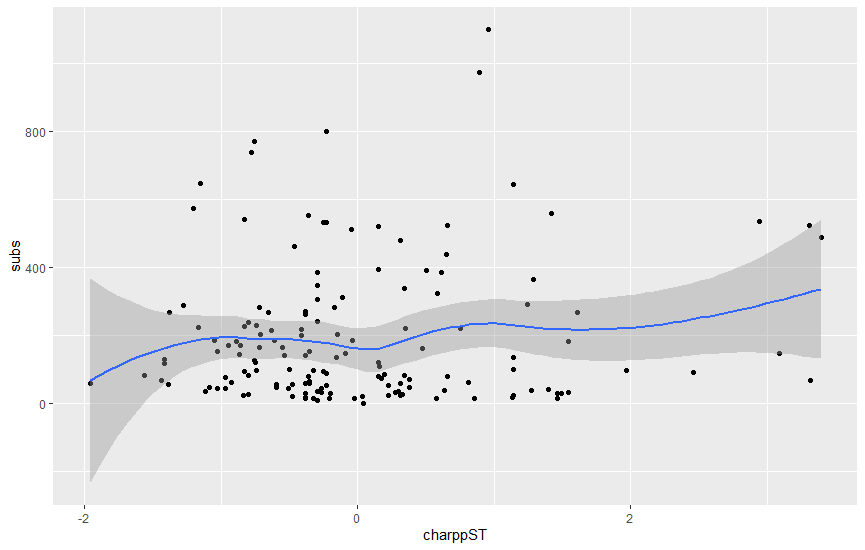


Figure 3.3: relationship between characters per page and subscriptions

Figure 3.4 shows the relationship between the amount of subscriptions and the amount of citations. It seems like there might be either a logarithmic relationship or a 3rd degree polynomial relationship. If we check both options it is shown that when the model is transformed using a logarithmic function, the r2 lowers from .3428 to .3401, showing that the fit using a logarithmic function is a little bit worse. However, if a 3rd degree polynomial transformation is used, the r2 jumps to .4127, and the adjusted R-squared also increases. Therefore, the relationship will be transformed using a 3rd degree polynomial.

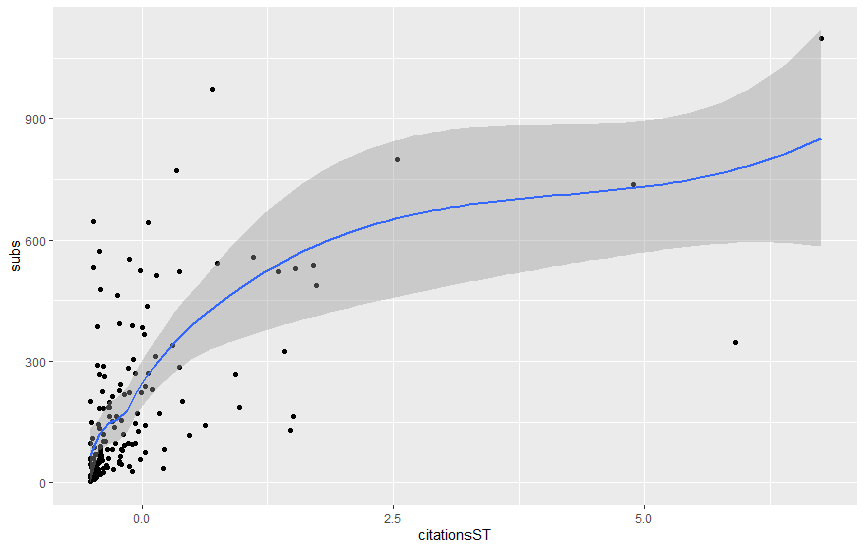


Figure 3.4: relationship between amount of citations and subscriptions

Finally, the relationship between founding year and subscriptions looks a lot like a 2nd degree polynomial. Indeed, when transformed the r2 increases from .01 to .33, as shown in figures 3.5.1 and 3.5.2

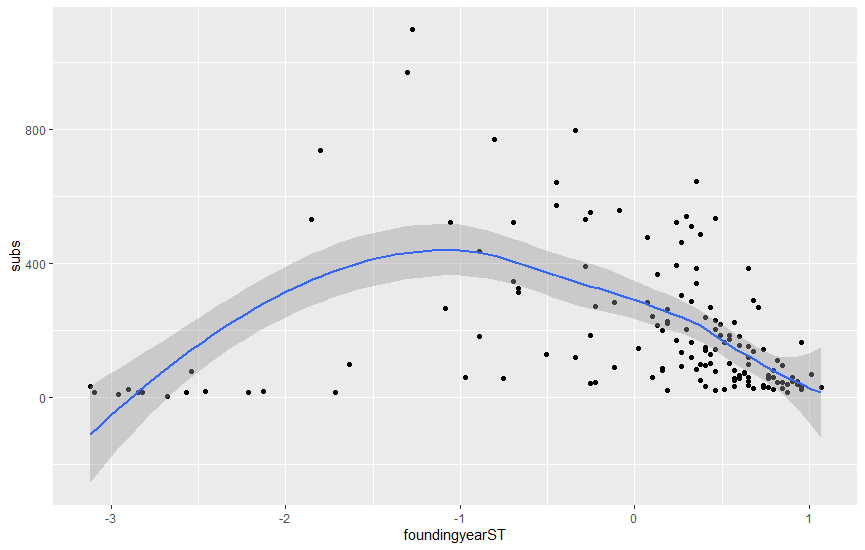


Figure 3.5: relationship between founding year and subscriptions

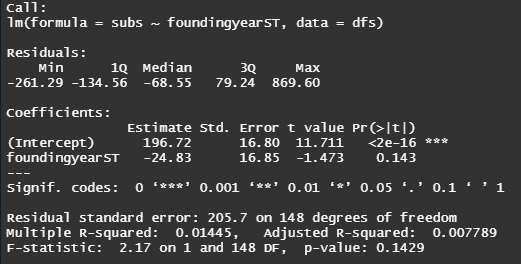


Figure 3.5.1: relationship between founding year and subscriptions with no transformations

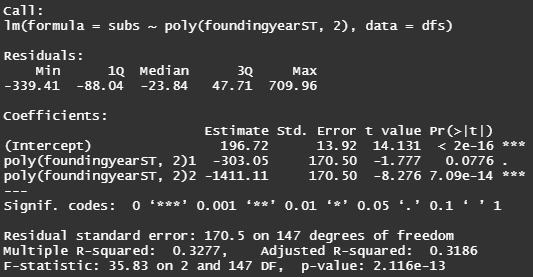


Figure 3.5.2: relationship between founding year and subscriptions with 2nd degree polynomial transformation

## Regression model

**Table 3.1: variables affecting the amount of subscriptions on economic journals, with and without correcting for non-linear relationships**

|  |  |  |
| --- | --- | --- |
|  | Without corrections | With corrections |
| (Intercept) | 196.72 \*\*\* | 196.72 \*\*\* |
|  | (12.72) | (10.41) |
| Price of journal | -67.66 \*\*\* |  |
|  | (13.11) |  |
| Character per page | 11.69 | 1.60 |
|  | (12.93) | (11.13) |
| Amount of citations | 120.97 \*\*\* |  |
|  | (13.04) |  |
| Founding year | 9.21 |  |
|  | (13.25) |  |
| Price, 1st polynomial |  | -886.42 \*\*\* |
|  |  | (142.20) |
| Price, 2nd polynomial |  | 277.72 \* |
|  |  | (136.27) |
| Price, 3rd polynomial |  | -184.74 |
|  |  | (138.81) |
| Citations, 1st polynomial |  | 1212.06 \*\*\* |
|  |  | (142.45) |
| Citations, 2nd polynomial |  | -558.87 \*\*\* |
|  |  | (142.17) |
| Citations, 3rd polynomial |  | 401.78 \*\* |
|  |  | (138.48) |
| Founding year, 1st polynomial |  | -3.67 |
|  |  | (134.56) |
| Founding year, 2nd polynomial |  | -565.55 \*\*\* |
|  |  | (151.58) |
| R^2 | 0.45 | 0.64 |
| Adj. R^2 | 0.43 | 0.62 |
| Num. obs. | 150 | 150 |

\* = p < 0.05, \*\* = p < 0.01, \*\*\* = p <0.005. Standard errors are between brackets. All variables are standardized.

## Explanation of the results

The model with corrections is a much better fit. Both models show which variables affect the amount of subscriptions with a statistical significance, namely the price of the journal and the amount of citations. The corrected model however also shows that founding year does have a statistical significance, while the first model misses this. It is also better in showing what kind of relationship there is between the variables. It shows much more clearly how important price is for instance, as there is much more significance of price differences in the low price range than the first model shows.

I would much rather have the 2nd model, although it might be confusing for people that are less versed in statistics and the meaning of the different results.

# Chapter 5: Question 4

## Description of data and cleaning process

The data used is data from a store about how much money customers spend, some identifying data and multiple variables that might influence this. These variables are: the average amount of time they spent with a stylist, the average amount of time they spent in the app, the average time they spent on the website and how long they have been a member.

There are no missing values

Because this is a prediction model and not a comparison, none of the variables are standardized.

## Description of the regression analysis

First, we check if there is multicollinearity. Figure 4.1 shows that there are no high correlations between the variables

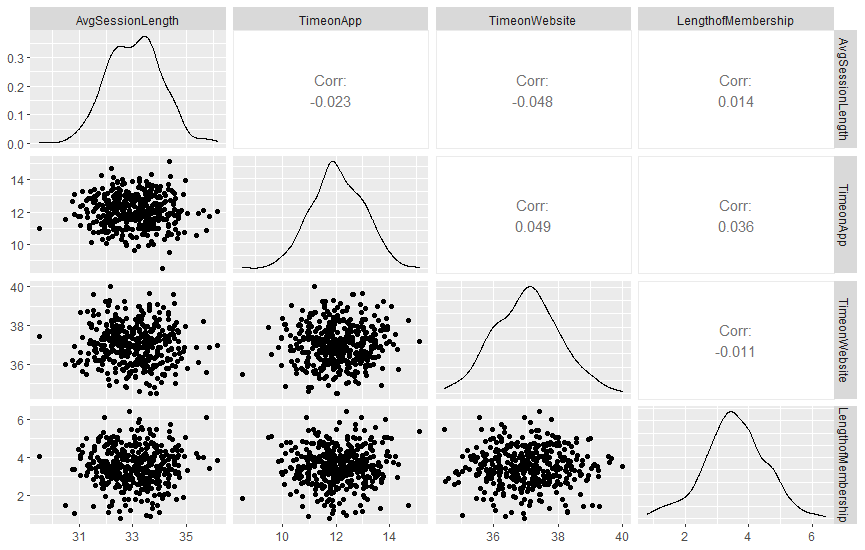


Figure 4.1: correlation between the independent variables

The actual analysis is fairly straightforward. No transformations or manipulations of the data is needed. As time spent on the website has absolutely no effect on the accuracy of the model, it is removed to simplify the model (parsimony).

## Regression model

**Table 4.1: Impact of different variables on the amount of money spent in the store**

|  |  |
| --- | --- |
|  | Impact |
| (Intercept) | -1041.12 \*\*\* |
|  | (17.89) |
| Time spent with stylist | 25.77 \*\*\* |
|  | (0.50) |
| TimeonApp | 39.06 \*\*\* |
|  | (0.51) |
| LengthofMembership | 61.72 \*\*\* |
|  | (0.51) |
| R^2 | 0.98 |
| Adj. R^2 | 0.98 |
| Num. obs. | 400 |

\* = p < 0.05, \*\* = p < 0.01, \*\*\* = p <0.005. Standard errors are between brackets

## Explanation of the results

As the R2 of this model is very good, it is a good tool to predict the amount spent in the store. We know it is accurate, because it explains 98% of all influences on the amount spent.

The amount of money someone will spend on the website is as follows:

If we fill in this formula for the first person in the dataset, we get the following result:

This is very close to the actual amount spent of 330.5944, with a difference of 5.49. Therefore, we can conclude that this model is fairly accurate.

# Chapter 6: Question 5

## Description of data and cleaning process

The data is a file containing a number of variables in different areas, townships and boroughs in the state of Pennsylvania. These variables are number of crimes per 1000 people, county the city is in, average house prices, distance from the center of Philadelphia center and change in population.

There is 1 missing value. This is removed, as it is only impacts a single row.

There are 3 rows of outliers. Removing these outliers leaves us with 81 observations. We are testing 3 variables, so our minimum amount of observations is around 50+3\*8=71. While the amount of observations is not ideal and we would prefer to have more than 100 observations, we can state that the sample size is big enough that we can reliably make a model with these observations.

## Description of the regression analysis

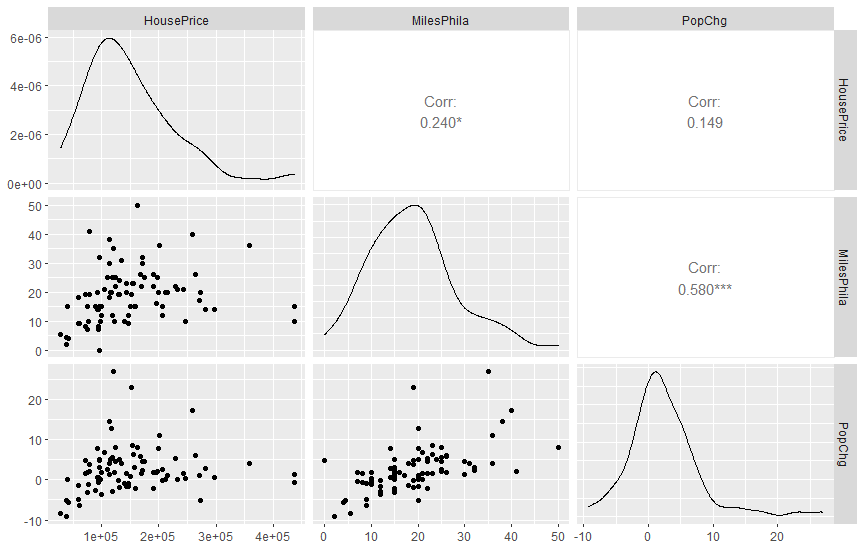
First, we check if there is multicollinearity. Figure 5.1 shows that there is no high correlation between the variables, so there is no risk of multicollinearity.

Figure 5.1: correlation between independent variables

If we look at the fit of the model with and without outliers, it is shown that the fit is far better without outliers, as the r2 is more than doubled. Therefore, we will use the model without outliers.

As population change has no significant effect on the crime rate, it is possible to leave it out of the model in the interest of parsimony. T

## Regression model

**Table 5.1: Variables affecting the crime rate of Pennsylvanian townships**

|  |  |  |  |
| --- | --- | --- | --- |
|  | With outliers | Without outliers | Standardized |
| (Intercept) | 73.50 \*\*\* | 46.05 \*\*\* | 28.56 \*\*\* |
|  | (11.66) | (4.73) | (1.56) |
| HousePrice | -0.00 \* | -0.00 \*\*\* | -8.12 \*\*\* |
|  | (0.00) | (0.00) | (1.63) |
| MilesPhila | -1.42 \* | -0.11 | -0.96 |
|  | (0.55) | (0.24) | (2.20) |
| PopChg | 1.34 | -0.19 | -1.02 |
|  | (0.88) | (0.40) | (2.15) |
| R^2 | 0.15 | 0.29 | 0.29 |
| Adj. R^2 | 0.12 | 0.26 | 0.26 |
| Num. obs. | 84 | 81 | 81 |

\* = p < 0.05, \*\* = p < 0.01, \*\*\* = p <0.005. Standard errors are between brackets. Variables in the Standardized column are Standardized

## Explanation of the results

As said earlier, the r2 when outliers are removed is a lot higher, so we will use that model. It is very interesting that the relationship between house prices and crime rate is extremely significant, but so small that it rounds to 0. This is because house prices are an extremely high number compared to the crime rate. The standardized column clearly shows the large impact of house prices on the crime rate. When house prices are 1 standard deviation larger than average, the crime rate drops by a staggering 8 incidents per 1000 people. This makes a lot of sense, as crime is often a result of poverty and wealthier neighborhoods tend to have a bigger presence of law enforcement.

The model also shows that the further away from Philadelphia, the lower the crime rate. Apparently, Philadelphia is a hub of crime in the area. A higher population growth also gives a lower crime rate. Both these variables however are not statistically significant, so these relationships shouldn’t be relied on.

As the r2 of the model is still only 0.26, a lot of the reasons for a high crime rate are missing from this model.

To calculate the crime rate of a municipality, the formula would be:

\* House prices + -0.11 \* Miles from Philadelphia + -0.19 \* Population change

# Chapter 7: Question 6

## Description of data and cleaning process

The data is a csv file containing data bout the earnings of alumni from different schools in the United States. It contains their school and location, if it is a private or public school, the price of admission, the SAT score, the earnings of the graduates and how much of their tuition was paid through a scholarship.

There are multiple missing values. These values are in the variables public, earn and need\_fraction. If we remove these values the dataset goes from 600 to 474 observations.

2 models will be made, 1 with every row that has a missing value removed and 1 where these values are replaced. The impact of both these methods will be assessed.

To replace the values, dummies have been made for the categorical variable public. For the continuous variables SAT and need\_fraction a dummy has been made so that it can be checked if the missing values are non-random. The missing values have been replaced by the mean of that variable.

## Description of the regression analysis

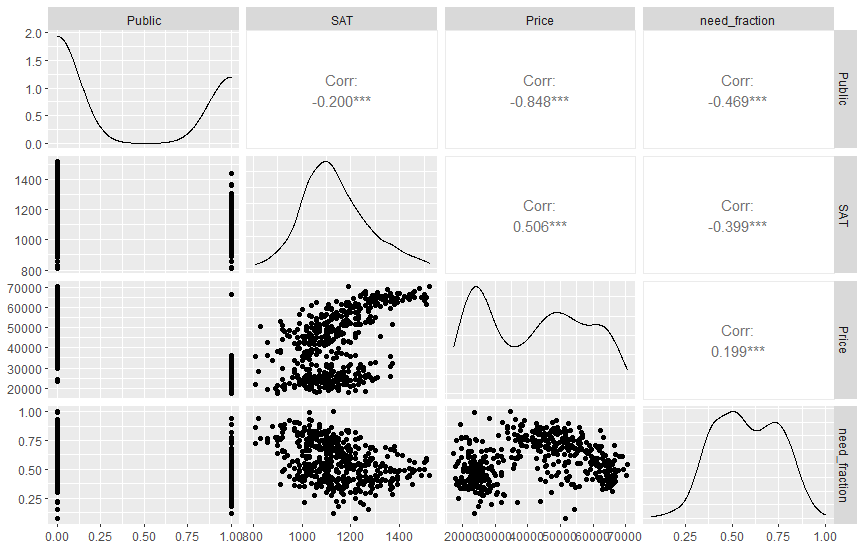


Figure 6.1: correlation between independent variables

The high correlation between price and public means that there might be multicollinearity. However, if we run the model both with and without public, no problems are found. Therefore, we can conclude that there is no issue of multicollinearity.

All the missing values are checked to see if there is a significant relationship between the missing values and the other variables. Figure 6.2 shows that there is no significant relationship between Public\_NA and the other variables, as there is no variable with a p<0.05 or a high estimate. The same goes for the other NA variables. Therefore, we can conclude that these missing variables are random.

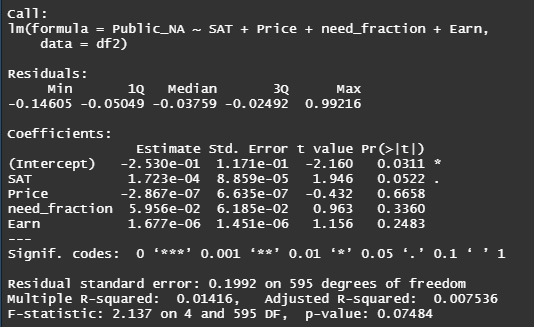


Figure 6.2: relationship between Public\_NA and the other variables

We can make two models, one where the missing variables are removed and one where they are replaced

## Regression model

**Table 6.1: variables affecting earnings of graduates**

|  |  |  |
| --- | --- | --- |
|  | Missing values removed | Missing values replaced |
| (Intercept) | 19490.04 \*\*\* | 26189.65 \*\*\* |
|  | (3483.00) | (3210.49) |
| Public school | 7861.01 \*\*\* |  |
|  | (1190.52) |  |
| SAT | 12.70 \*\*\* | 10.69 \*\*\* |
|  | (2.68) | (2.40) |
| Price | 0.29 \*\*\* | 0.26 \*\*\* |
|  | (0.04) | (0.03) |
| % tuition from scholarships | -5717.50 \*\* | -10093.71 \*\*\* |
|  | (1918.58) | (1745.17) |
| Public school (without missing values) |  | 6109.48 \*\*\* |
|  |  | (995.22) |
| R^2 | 0.39 | 0.37 |
| Adj. R^2 | 0.39 | 0.37 |
| Num. obs. | 474 | 600 |

\* = p < 0.05, \*\* = p < 0.01, \*\*\* = p <0.005. Standard errors are between brackets.

## Explanation of the results

There is some difference between the model with the missing values removed and the model with the missing values replaced. We still get the same explanation from both models, however. There is a significant positive relationship between earnings and public school. This means that those that studied at a public school on average earned more than those who studied at a private school. This makes some sense as public schools in the United States have far stricter requirements for teachers and curriculum. Therefore, on average, public schools are considered better than private schools in the US. If this data came from the UK, this result would probably be reversed as in the UK private schools are usually better regarded, so if this is a variable used, you need to be careful that your data is not contaminated by collecting data from different countries.

There is also a significant positive relation between price and earnings. Those who went to a more expensive school also earned more. The same goes for SAT score, those with a higher SAT score earned more. This makes a lot of sense, as those with a higher SAT score tend to go to better schools and the better schools can charge more tuition.

There is a significant negative relationship between % tuition from scholarships and earnings, meaning that those who relied more on scholarships generally earn less. This might be because those people that get more money from scholarships are generally either from poorer areas or have some sort of disability. Those from poorer areas generally don’t have the network that those from rich areas do, making getting a well-paying job harder. Disabilities also generally make it harder to get well-paying jobs in the United States.

So the conclusions from both models are the same, although the effect of each variable differs in the models. For instance, %tuition from scholarships has far more impact in the model with replacements than it does in the model with missing values removed. It is hard to say which model is better, as both manipulate the data in different ways. The model with replacements just replaces the missing values with the mean, which might have a big influence on the accuracy on the data and makes those values meaningless. Removing the missing values also presents its own problems however, as these missing values might be non-random, making it so removing these rows makes the dataset more biased.

Ultimately, I would choose those remove the rows with missing values in this case. We couldn’t find a significant correlation between the missing values and the other variables, reducing the chance that the missing values were non-random. I think that unless you replace the missing values with carefully imputated values (like with MICE), it just contaminates your dataset too much.

# Bibliography

Investopedia, N.D. *R-Squared vs. Adjusted R-Squared: What’s the difference?* Investopedia. Geraadpleegd op 21 october 2021, van <https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp>