Cihan Açilov

j.achilov@student.han.nl

Abstract

The report presents findings of the data analysis that have been carried on six different cases (datasets).

Introduction to Modelling

Data analysis and statistical modeling on six different cases

# Title Page

|  |  |
| --- | --- |
| Name School | Hogeschool van Arnhem & Nijmegen |
| Study year | 2021-2022 |
| Degree program and target group | Minor |
| Code of Study Unit | MDDF IML1A.5 |
| Name of Study Unit | Introduction to Modelling |
|  |  |
| Name Student | Cihan Acilov |
| Student Number | 556000 |
| Contact Student | [j.achilov@student.han.nl](mailto:j.achilov@student.han.nl) |
|  |  |
| Name Teacher | Tijmen Weber |
| Contact Teacher | [tijmen.weber@han.nl](mailto:tijmen.weber@han.nl) |

Contents

[Title Page 1](#_Toc85837326)

[Introduction 5](#_Toc85837327)

[1. Assignment 1. Student Alcohol Consumption 6](#_Toc85837328)

[1.1. Data Description 6](#_Toc85837329)

[1.2. Data preparation 6](#_Toc85837330)

[1.3. Data Analysis 7](#_Toc85837331)

[1.4. Regression results & Conclusion 11](#_Toc85837332)

[2. Assignment 2. Amazon Book Sales 13](#_Toc85837333)

[2.1. Data Description 13](#_Toc85837334)

[2.2. Data preparation 13](#_Toc85837335)

[2.3. Data Analysis 14](#_Toc85837336)

[2.4. Regression results & Conclusion 17](#_Toc85837337)

[3. Assignment 3. Economic Journal Subscription 19](#_Toc85837338)

[3.1. Data Description 19](#_Toc85837339)

[3.2. Data preparation 19](#_Toc85837340)

[3.3. Data Analysis 20](#_Toc85837341)

[3.4. Regression results & Conclusion 24](#_Toc85837342)

[4. Assignment 4. E-commerce Customers 26](#_Toc85837343)

[4.1. Data Description 26](#_Toc85837344)

[4.2. Data preparation 26](#_Toc85837345)

[4.3. Data Analysis 26](#_Toc85837346)

[4.4. Regression results & Conclusion 26](#_Toc85837347)

[5. Assignment 5. Crime Rate in Philadelphia 27](#_Toc85837348)

[5.1. Data Description 27](#_Toc85837349)

[5.2. Data preparation 27](#_Toc85837350)

[5.3. Data Analysis 27](#_Toc85837351)

[5.4. Regression results & Conclusion 30](#_Toc85837352)

[6. Assignment 6. Graduate Earnings 31](#_Toc85837353)

[6.1. Data Description 31](#_Toc85837354)

[6.2. Data preparation 31](#_Toc85837355)

[6.3. Data Analysis 33](#_Toc85837356)

[6.4. Regression results & Conclusion 36](#_Toc85837357)

[Title Page 1](#_Toc85837273)

[Introduction 4](#_Toc85837274)

[1. Assignment 1. Student Alcohol Consumption 5](#_Toc85837275)

[1.1. Data Description 5](#_Toc85837276)

[1.2. Data preparation 5](#_Toc85837277)

[1.3. Data Analysis 6](#_Toc85837278)

[Variable Sex 6](#_Toc85837279)

[Variable Address 7](#_Toc85837280)

[Variable Romantic 7](#_Toc85837281)

[Variable Grade 8](#_Toc85837282)

[Variable Activities 9](#_Toc85837283)

[Multicollinearity 9](#_Toc85837284)

[1.4. Regression results & Conclusion 10](#_Toc85837285)

[2. Assignment 2. Amazon Book Sales 12](#_Toc85837286)

[2.1. Data Description 12](#_Toc85837287)

[2.2. Data preparation 12](#_Toc85837288)

[2.3. Data Analysis 13](#_Toc85837289)

[Multicollinnearity 13](#_Toc85837290)

[Standardization 13](#_Toc85837291)

[Transformation 14](#_Toc85837292)

[2.4. Regression results & Conclusion 16](#_Toc85837293)

[3. Assignment 3. Economic Journal Subscription 18](#_Toc85837294)

[3.1. Data Description 18](#_Toc85837295)

[3.2. Data preparation 18](#_Toc85837296)

[3.3. Data Analysis 19](#_Toc85837297)

[Initial Regression 19](#_Toc85837298)

[Identifying nonlinear relationships 20](#_Toc85837299)

[3.4. Regression results & Conclusion 23](#_Toc85837300)

[Regression results after transformation 23](#_Toc85837301)

[Regression results after removing outliers 24](#_Toc85837302)

[4. Assignment 4. E-commerce Customers 25](#_Toc85837303)

[4.1. Data Description 25](#_Toc85837304)

[4.2. Data preparation 25](#_Toc85837305)

[4.3. Data Analysis 25](#_Toc85837306)

[4.4. Regression results & Conclusion 25](#_Toc85837307)

[5. Assignment 5. Crime Rate in Philadelphia 26](#_Toc85837308)

[5.1. Data Description 26](#_Toc85837309)

[5.2. Data preparation 26](#_Toc85837310)

[5.3. Data Analysis 26](#_Toc85837311)

[5.4. Regression results & Conclusion 29](#_Toc85837312)

[6. Assignment 6. Graduate Earnings 30](#_Toc85837313)

[6.1. Data Description 30](#_Toc85837314)

[6.2. Data preparation 30](#_Toc85837315)

[Missing values 31](#_Toc85837316)

[6.3. Data Analysis 32](#_Toc85837317)

[6.4. Regression results & Conclusion 35](#_Toc85837318)

# Introduction

At HAN University of Applied Science the student from different backgrounds and schools are able to follow the minor program Data Driven Decision Making. During the first half of the program the students are gently introduced into to topics such us statistics, data analysis and data visualization.

The study unit Introduction to Modelling focus solely on statistics, explanatory and predictive modeling. This rapport is created as the final product/assignment for that study unit. The student who are following this lesson get six different datasets, each having an individual assignment that have to be handed in at the end of the first half of the minor program.

The datasets that are used are as follow:

1. Student Alcohol Consumption (Assignment 1)
2. Amazon Book Sells (Assignment 2)
3. Economics Journal Subscription (Assignment 3)
4. E-commerce Customers (Assignment 4)
5. Crime Rate of Philadelphia (Assignment 5)
6. Graduate earnings (Assignment 6)

There are six assignments that the students have to carry out that correlate with the datasets above. Each dataset is different in design, this leads to six different data analysis, cleaning and modelling methods that each dataset requires.

The report is structured in way that each assignment has its own standard sections as shown below:

1. Data description: here is the data described in terms of assignment goal, data source, the number of variables and observations and statistical summaries. This section covers the data cleaning and data preparation as well.
2. Data analyses: this section focuses on the individual variables in relation to the target variable, meaning that the variables are controlled for multicollinearity, different transformation methods are applied.
3. Results and Conclusion: This section presents the results of the regression analysis and the findings are described based on the results. Finally and overall conclusion is given per dataset.

# Assignment 1. Student Alcohol Consumption

## Data Description

In this section the dataset of *Student\_Alcohol\_Consumption.csv* is analyzed. The data contains the alcohol consumption behavior of students at a Portuguese school which is trying to figure out whether the alcohol consumption of students are related to private and school characteristics.

The original dataset contains **300 observations and 10 variables** (no missing values) as shown below.

* **Alcohol**: weekly alcohol consumption measured in scale between 1 (Very Low) to 5 (Very High).
* **Sex**: the gender of the students (F=female & M=male)
* **Activities**: extra-curricular activities (Yes or No)
* **Address**: in which area the students are living (U=urban & R=rural)
* **Grade**: grade scores in scale between 1 (Lowest) and 20 (Highest)
* **Absences**: number of classes missed this year
* **Romantic**: whether a student is involved in a romantic relationship (Yes or No)
* **Age**: the age of the students
* **Famrel**: quality of family relationship in scale between 1(Very Bad) and 5 (Very Good)
* **Health**: health status in scale between 1 (Bad Health) and 5 (Good Health)

Even though there are 10 variables in the dataset, only the first 7 variables are used for the analyses and modeling purposes (Alcohol, Sex, Activities, Address, Grade, Absences and Romantic).

## Data preparation

This section describes the steps that are taken to clean the data.

Firstly, a new data frame is created with the 7 variables that will be used in the analyses. Tables below represent the summary of the data frame; in Table 1 the continuous variables are summarized and in Table 2 the categorical variables:

**Table 1. Summary of Continuous Variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Continuous variables | **Min** | **Median** | **Mean** | **Max** |
| **Alcohol** | 1.00 | 2.00 | 2.27 | 5.00 |
| **Absences** | 0.00 | 4.00 | 5.79 | 75.00 |
| **Grade** | 0.00 | 11.00 | 10.56 | 20.00 |

There are 300 observation and no missing values. The values of continuous variables are proportionally close to each other. Thus a standardization might not needed.

**Table 2. Summary of Categorical Variables**

|  |  |  |
| --- | --- | --- |
| Categorical variables | **Level 1** | **Level 2** |
| **Sex** | Female: 157 | Male: 143 |
| **Address** | Rural: 62 | Urban: 238 |
| **Activities** | Yes: 150 | No: 150 |
| **Romantic** | Yes: 101 | No: 199 |

There are 300 observations and no missing values. All four categorical values have 2 levels.

To be able to see the exact influence of categorical variables, new dummy columns are created from Sex, Address, Activities and Romantic. Also, the names of the dummy columns are modified for better readability.

* From the sex columns two dummy columns are created: male and female
* From the address columns two dummy columns are created: rural and urban
* From the activities column two dummy columns are created: activity\_yes and activity\_no
* From the Romantic column two dummy columns are created: ramntic\_yes and romantic\_no

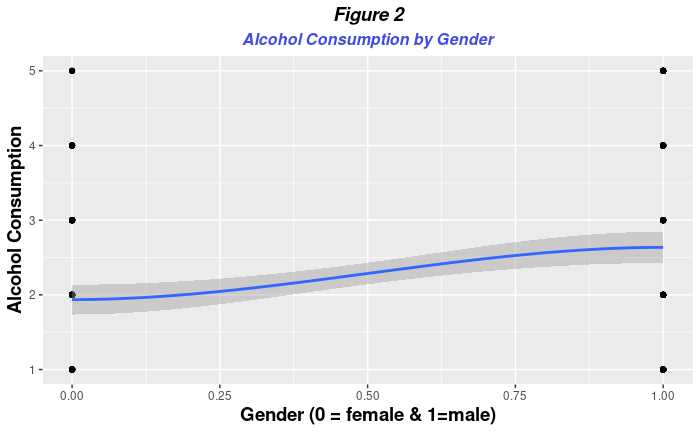
## Data Analysis

This section presents the findings of the analysis. Firstly, each independent variable is analyzed and compared to the dependent variable. Secondly, the variables are checked for multicollinearity and outliers.

### Variable Sex

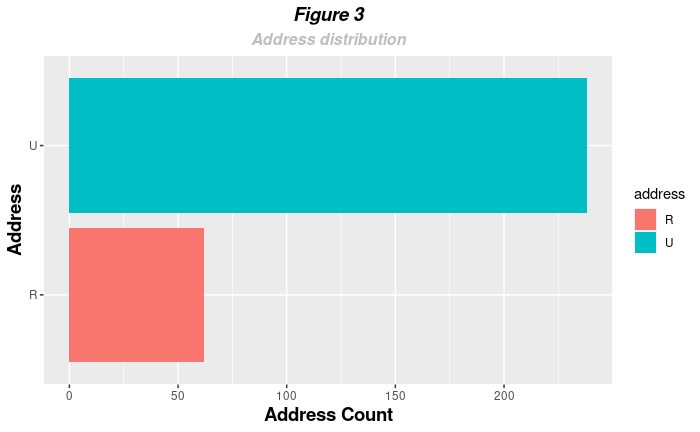
The dataset contains 157 female students and 143 male students as shown in Figure 1:

Chart, bar chart

Description automatically generated

As shown in Figure 2, the alcohol consumption by males seems to be higher than females (males 2.7 and females 1.9).

### Variable Address



The address variables contain two categories: urban and rural. As shown in Figure 3, 238 students are living in urban areas, and 62 students are living in rural areas.

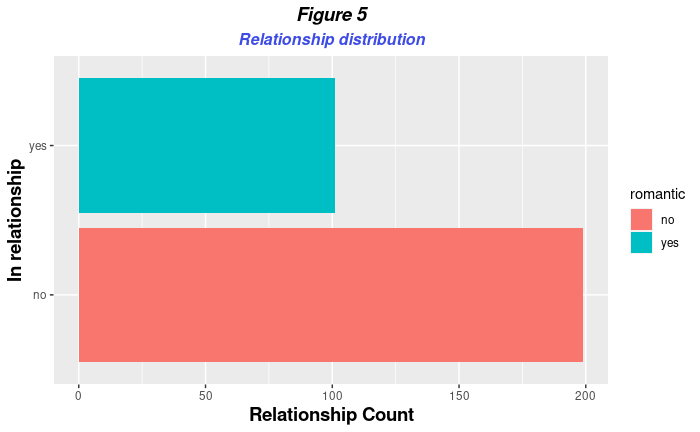
Chart, square

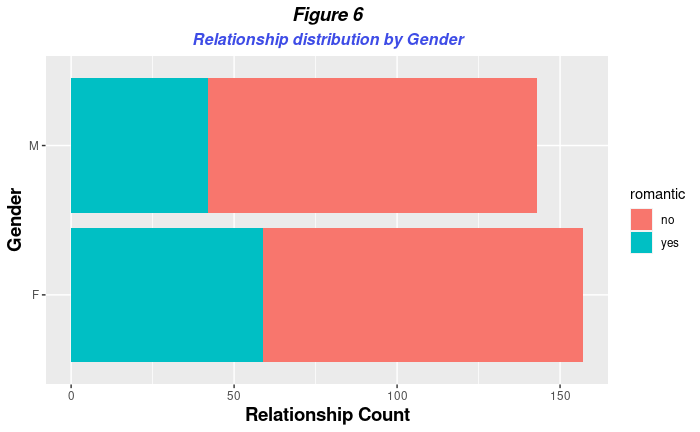
Description automatically generated

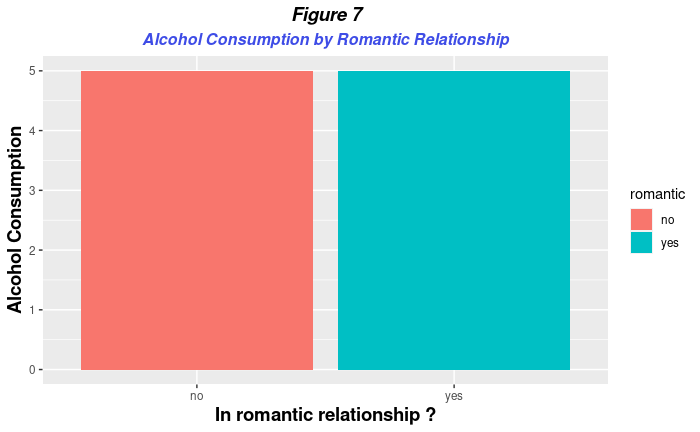
Figure 4 shows that there are no differences in alcohol consumption between urban and rural areas.

### Variable Romantic

This binary variable shows whether a student is engaged in aromantic relationships. Figure 5 shows that a 2/3 of the students are not engaged in a romantic relationship while 1/3 of the students are in a romantic relationship, as shown below in Figure 5:



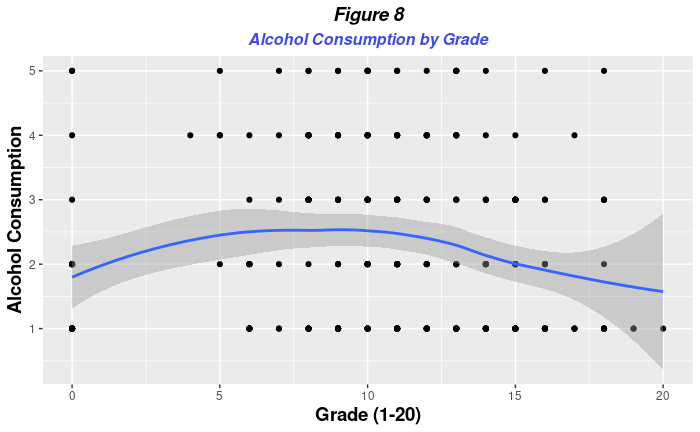
Figure 6 illustrates the distribution of romantic relationships by gender of the students. The number of Female students, who are in a romantic relationship, are slightly higher than the number of male students:



As shown in Figure 7, there are no differences in alcohol consumption between students who are in a relationship and those who are not.

### Variable Grade

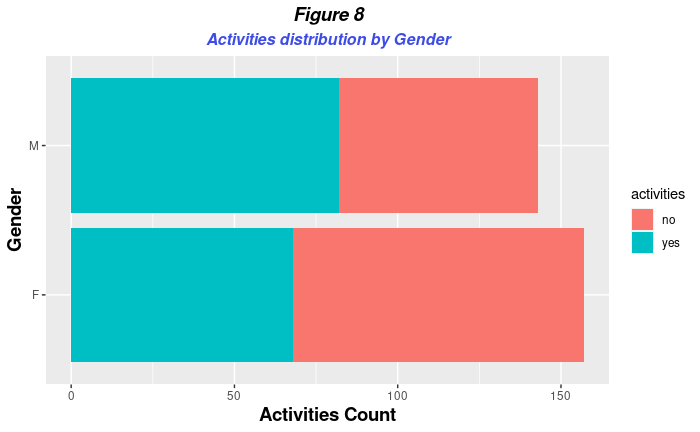
As shown in Figure 8, the students with a grade between 0 and 10 seem to consume more alcohol than students with a grade above 10:



### Variable Activities

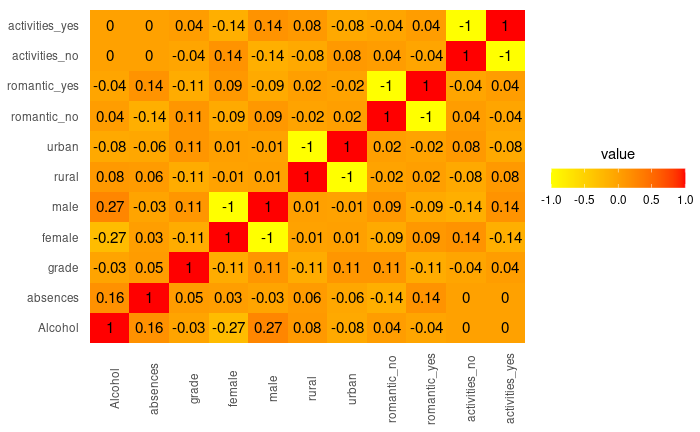
Half of the students (150) are involved in extracurricular activities, and the other half not.

In Figure 8 the gender of the students is compared to whether they are engaged in any extracurricular activity. Male students are slightly more engaged in extracurricular activities than female students:



### Multicollinearity

In this section the variables are checked for multicollinearity.

As shown in Figure 9, in the correlation heatmap there is no multicollinearity between the variables, excluding the dummy variables: which will always have a negative correlation due to its binary nature. For example, variable rural will always be negatively correlated with the variable urban.

**Figure 9 Correlation Heatmap**

## Regression results & Conclusion

In this section regression analyses are carried out on the dataset. The amount of alcohol consumption is used as the depended variable and independent variables are set in three different models as shown below:

1. **Model1:** Alcohol consumption by all genders
2. **Model2:** Alcohol consumption by males
3. **Model3:** Alcohol consumption by females

**Table 3. Regression Results only by gender**

|  |  |  |  |
| --- | --- | --- | --- |
| *Alcohol Consumption behavior of students (between 15 and 21 years old)* | | | |
|  | **DV: Alcohol** | | |
|  | All Genders | Male | Female |
|  | | | |
| sexM | 0.700\*\*\* (0.147) |  |  |
| male |  | 0.700\*\*\* (0.147) |  |
| female |  |  | -0.700\*\*\* (0.147) |
| Constant | 1.936\*\*\* (0.101) | 1.936\*\*\* (0.101) | 2.636\*\*\* (0.106) |
|  | | | |
| Observations | 300 | 300 | 300 |
| R2 | 0.071 | 0.071 | 0.071 |
| Adjusted R2 | 0.068 | 0.068 | 0.068 |
| Residual Std. Error (df = 298) | 1.270 | 1.270 | 1.270 |
| F Statistic (df = 1; 298) | 22.748\*\*\* | 22.748\*\*\* | 22.748\*\*\* |
|  | | | |
| Significance levels | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | |

The results of the regressions in Table 3, shows that the gender of students has a significant relationship with the amount of alcohol consumed. Male students are consuming 0.700 alcohol per week than female students. R square is 0.071, which means the model applies only for the 7 % of the dataset which is an extremely low.

Here below in Table 4 the gender variable is controlled by other independent variables:

**Table 4. Regression Results Gender controlled by other variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Alcohol Consumption behavior of students (between 15 and 21 years old) | | | |
|  | DV: Alcohol | | |
|  | All Genders | Male | Female |
|  | | | |
| sexM | 0.737\*\*\* (0.148) |  |  |
| male |  | 0.737\*\*\* (0.148) |  |
| female |  |  | -0.737\*\*\* (0.148) |
| absences | 0.227\*\*\* (0.073) | 0.227\*\*\* (0.073) | 0.227\*\*\* (0.073) |
| grade | -0.093 (0.074) | -0.093 (0.074) | -0.093 (0.074) |
| activities\_yes | -0.114 (0.147) | -0.114 (0.147) | -0.114 (0.147) |
| romantic\_yes | -0.146 (0.156) | -0.146 (0.156) | -0.146 (0.156) |
| urban | -0.222 (0.181) | -0.222 (0.181) | -0.222 (0.181) |
| Constant | 2.201\*\*\* (0.199) | 2.201\*\*\* (0.199) | 2.938\*\*\* (0.205) |
|  | | | |
| Observations | 300 | 300 | 300 |
| R2 | 0.112 | 0.112 | 0.112 |
| Adjusted R2 | 0.094 | 0.094 | 0.094 |
| Residual Std. Error (df = 293) | 1.252 | 1.252 | 1.252 |
| F Statistic (df = 6; 293) | 6.168\*\*\* | 6.168\*\*\* | 6.168\*\*\* |
|  | | | |
| Significance levels | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | |

Both gender and absence have a significant relationship with alcohol consumption. Alcohol consumption is higher with male students. Compared to the first regression results, in Table 1, the coefficient is slightly higher, at 0.737.

Absence has a positive relationship, if absence increases by 1, Alcohol consumption increases by 0.227.

The rest of the independent variables have no significant relationship with alcohol consumption. Also, there are no differences between models All Gender, Female and Male.

Because there are multiple independent variables involved into this model adjusted R square is more accurate, which is 9%, slightly higher than the first regression model with only gender as independent variable which had a R square of 7%.

# Assignment 2. Amazon Book Sales

## Data Description

Dataset that is analyzed in this section contains data regarding books crawled from amazon.com. The data is provided by HAN as the final assignment of the lesson Introduction to Modeling by HAN..

The dataset contains the variables below:

1. Title: the title of the book
2. Author: the author of the book
3. ListPrice: the list price
4. AmazonPrice: The amazon price
5. HardorPaper: Hardback or Paperback
6. NumPages: Number of pages
7. Publisher: name of the publisher
8. PubYear: Publication year
9. ISBN.10: ISBN ID of the book
10. Height: height
11. Thick: thickness of the book
12. Weight: the weight of the book

The goal of this assignment is to check whether there is multicollinearity between the variables and fix it appropriately. The next goal is to determine which variable has the strongest influence on the dependent variable AmazonPrice.

## Data preparation

This section describes the steps that are taken to clean and prepare the data for the analysis.

A new data frame is created with the variables that will be used in the analysis (AmazonPrice, PubYear, HardorPaper, NumPages, Height and Thick). Here below the summary of the newly created data frame:

**Table 5. Variables Summary**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | **Min** | **Median** | **Mean** | **Max** | **Observations** | **Missing values** |
| **AmazonPrice** | 0.770 | 10.200 | 12.578 | 118.210 | 270 | 0 |
| **NumPages** | 24.0 | 320.0 | 336.2 | 896.0 | 270 | 0 |
| **PubYear** | 1936 | 2004 | 2002 | 2011 | 270 | 0 |
| **Height** | 5.200 | 8.000 | 8.125 | 12.100 | 270 | 0 |
| **Thick** | 0.1000 | 0.8000 | 0.8989 | 2.0000 | 270 | 0 |

As shown above there are 270 observations with no missing values. The values of the variables differ proportionally, which means that they must be standardized.

**Table 6. Variables Summary**

|  |  |  |
| --- | --- | --- |
| Categorical variables | **Level 1** | **Level 2** |
| **HardorPaper** | Hardback: 69 | Paperback: 201 |

As shown in the Table above there is only one categorical variable with two levels. There are no missing values.

To be able to see the exact influence of categorical variables, two dummy columns are created from HardorPaper: Hard & Paper.

From the PubYear column a new column Age is created which is subtracting the PubYear from 2021.

The continuous variables are standardized as below:

1. NumPages: Scale of the variable is used
2. Height: Scale of the variable is used
3. Thick: Scale of the variable is used

## Data Analysis

This section the variables are analyzed to see wether there is multicollinearity between the variables. Next the two models are created: one without standardized variables and the other with standardized variables.

### Multicollinnearity

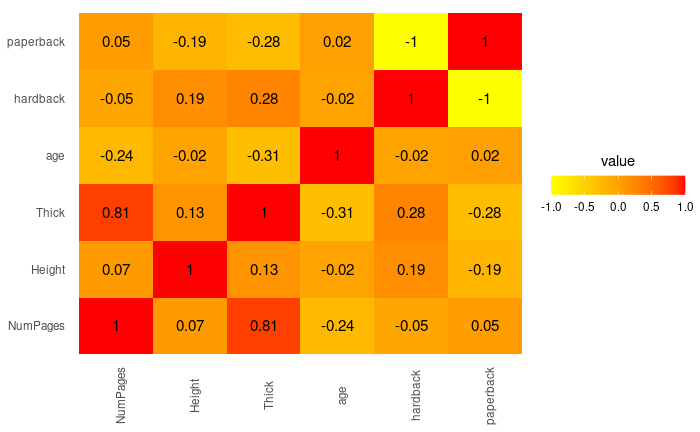


Figure 10 Correlation Heatmap

As shown in Figure 10, Thick and NumPages are highly correlated. This means that one of the variables has to be removed from the model.

### Standardization

**Table 7. Comparison standardized model and not standardized model**

|  |  |  |
| --- | --- | --- |
|  | DV: AmazonPrice | |
|  | (1) | (2) |
|  | | |
| NumPages | 0.005 (0.004) |  |
| Height | 4.797\*\*\* (0.774) |  |
| age | -0.066 (0.060) |  |
| scale(NumPages) |  | 0.894 (0.673) |
| scale(Height) |  | 4.123\*\*\* (0.665) |
| scale(age) |  | -0.738 (0.670) |
| hardback | -1.481 (1.521) | -1.481 (1.521) |
| Constant | -26.556\*\*\* (6.464) | 12.956\*\*\* (0.756) |
|  | | |
| Observations | 270 | 270 |
| R2 | 0.144 | 0.144 |
| Adjusted R2 | 0.131 | 0.131 |
| Residual Std. Error (df = 265) | 10.663 | 10.663 |
| F Statistic (df = 4; 265) | 11.105\*\*\* | 11.105\*\*\* |
|  | | |
| *Significance:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | |

Table 7 shows differences in the model before and after transformation. NumPag pages have a greater impact but no significance. Height remain significant and has slightly lower influence. R-square remained same.

### Transformation

**Table 8. Comparison model before and after transformation**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
|  | *DV: Dependent variable:* | | |
|  |  | | |
|  | AmazonPrice | | log(AmazonPrice) |
|  | (1) | (2) | (3) |
|  | | | |
| log(NumPages) | 5.672\*\*\* (2.068) |  |  |
| log(Height) | 33.337\*\*\* (6.112) |  |  |
| log(Thick) | -6.996\*\* (2.935) |  |  |
| log(age) | -1.845 (1.506) |  |  |
| scale(poly(NumPages, 2))1 |  | 5.911\*\*\* (1.273) |  |
| scale(poly(NumPages, 2))2 |  | -1.475\* (0.843) |  |
| scale(poly(Height, 2))1 |  | 4.179\*\*\* (0.637) |  |
| scale(poly(Height, 2))2 |  | 2.011\*\*\* (0.658) |  |
| scale(poly(Thick, 2))1 |  | -5.749\*\*\* (1.346) |  |
| scale(poly(Thick, 2))2 |  | -0.211 (0.817) |  |
| scale(age) |  | -1.282\* (0.658) |  |
| NumPages |  |  | 0.001\*\*\* (0.0003) |
| Height |  |  | 0.241\*\*\* (0.031) |
| Thick |  |  | -0.408\*\*\* (0.153) |
| age |  |  | -0.003 (0.002) |
| hardback | 1.381 (1.839) | 2.214 (1.782) | 0.193\*\*\* (0.073) |
| Constant | -85.704\*\*\* (17.235) | 12.012\*\*\* (0.769) | 0.397 (0.263) |
|  | | | |
| Observations | 270 | 270 | 270 |
| R2 | 0.145 | 0.230 | 0.252 |
| Adjusted R2 | 0.129 | 0.206 | 0.238 |
| Residual Std. Error | 10.672 (df = 264) | 10.188 (df = 261) | 0.429 (df = 264) |
| F Statistic | 8.983\*\*\* (df = 5; 264) | 9.745\*\*\* (df = 8; 261) | 17.787\*\*\* (df = 5; 264) |
|  | | | |
| *Significance:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | |

**Table 9. Without multicollinearity**

|  |  |  |  |
| --- | --- | --- | --- |
|  | *DV: Dependent variable:* | | |
|  |  | | |
|  | AmazonPrice | | log(AmazonPrice) |
|  | (1) | (2) | (3) |
|  | | | |
| log(NumPages) | 1.553 (1.147) |  |  |
| log(Height) | 34.423\*\*\* (6.149) |  |  |
| log(age) | -1.226 (1.497) |  |  |
| scale(poly(NumPages, 2))1 |  | 1.268\* (0.674) |  |
| scale(poly(NumPages, 2))2 |  | -1.131\* (0.656) |  |
| scale(poly(Height, 2))1 |  | 4.130\*\*\* (0.656) |  |
| scale(poly(Height, 2))2 |  | 1.940\*\*\* (0.674) |  |
| scale(age) |  | -0.696 (0.664) |  |
| NumPages |  |  | 0.0005\*\*\* (0.0002) |
| Height |  |  | 0.239\*\*\* (0.031) |
| age |  |  | -0.002 (0.002) |
| hardback | -1.016 (1.553) | -2.052 (1.529) | 0.085 (0.062) |
| Constant | -64.400\*\*\* (14.867) | 13.102\*\*\* (0.749) | 0.284 (0.263) |
|  | | | |
| Observations | 270 | 270 | 270 |
| R2 | 0.127 | 0.175 | 0.232 |
| Adjusted R2 | 0.114 | 0.156 | 0.220 |
| Residual Std. Error | 10.765 (df = 265) | 10.506 (df = 263) | 0.434 (df = 265) |
| F Statistic | 9.637\*\*\* (df = 4; 265) | 9.284\*\*\* (df = 6; 263) | 20.004\*\*\* (df = 4; 265) |
|  | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | |

## Regression results & Conclusion

Because the exponential model fits the data better than logarithmic and polynomial transformation, the final model is made based on polynomial transformation:

**Table 10. Graduate Earning**

|  |  |  |  |
| --- | --- | --- | --- |
|  | DV: log(AmazonPrice) | | |
|  | (1) | (2) | (3) |
|  | | | |
| scale(NumPages) | 0.077\*\*\* (0.027) | 0.127\*\*\* (0.046) |  |
| scale(Height) | 0.206\*\*\* (0.027) | 0.218\*\*\* (0.027) |  |
| scale(age) | -0.018 (0.027) | -0.028 (0.028) |  |
| NumPages |  |  | 0.0005\*\*\* (0.0002) |
| Height |  |  | 0.239\*\*\* (0.031) |
| age |  |  | -0.002 (0.002) |
| hardback | 0.085 (0.062) |  | 0.085 (0.062) |
| scale(Thick) |  | -0.068 (0.047) |  |
| Constant | 2.354\*\*\* (0.031) | 2.375\*\*\* (0.026) | 0.284 (0.263) |
|  | | | |
| Observations | 270 | 270 | 270 |
| R2 | 0.232 | 0.232 | 0.232 |
| Adjusted R2 | 0.220 | 0.221 | 0.220 |
| Residual Std. Error (df = 265) | 0.434 | 0.434 | 0.434 |
| F Statistic (df = 4; 265) | 20.004\*\*\* | 20.059\*\*\* | 20.004\*\*\* |
|  | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | |

Table above presents the results of the three regression models: 1) Exponential standardized without multicollinear variable and 2) exponential standardized with multicollinear variable and 3) not standardized without multicollinear variable.

The second model fits the data better than the other two models. Numpages and Height are positively related with AmazonPrice variable. The variables NumPages and Hight have the strongest influence: if Height increases by 1, AmazonPrice increases by 0.218 and if Numpages increase by 1 AmazonPrice increases by 0.127.

The R square in all three models is not high enough to represent a sufficient part of the data sample. All three models have a R square of 22%.

# Assignment 3. Economic Journal Subscription

## Data Description

The dataset Economics Journal Subscription contains data regarding subscriptions to economic journals in US libraries for the year 2000. The data is provided by HAN as the final assignment of the lesson Introduction to Modeling by HAN.

De dataset contains the following variables:

1. **Abbreviation**: the ID of the titles (can also been seen as de SKU)
2. **Title**: Journal title
3. **Publisher**: factor with publisher name
4. **Society**: Is the journal published by a scholarly society
5. **Price**: the price of the journal
6. **Pages**: number of pages
7. **Charpp**: total number of characters per page
8. **Citations**: total number of citations
9. **Foundingyear**: year of first publication
10. **Subs**: the total number of subscribers
11. **Field**: factor with field description.

The goal of this assignment is to create an appropriate model with the number of subscribers as the dependent variable (subs) and the variables below as the independent variable:

1. Price
2. Citations
3. Charpp
4. Foundingyear

## Data preparation

This section describes the steps that are taken to clean and prepare the data for the analysis.

A new data frame is created with the variables that will be used in the analysis (Subs, Price, Citations, Charpp, Founding year). Here below the summary of the newly created data frame:

**Table 11. Variables Summary**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | **Min** | **Median** | **Mean** | **Max** | **Observations** | **Missing values** |
| **price** | 36.0 | 287.5 | 429.7 | 2120.0 | 300 | 0 |
| **charpp** | 1820 | 3030 | 3261 | 6859 | 300 | 0 |
| **citations** | 21.0 | 293.0 | 647.1 | 7943.0 | 300 | 0 |
| **foundingyear** | 1844 | 1970 | 1959 | 1996 | 300 | 0 |
| **subs** | 2.0 | 120.0 | 197.2 | 972.0 | 300 | 0 |

There are 300 observations in the dataset with no missing values. The values differ proportionally which means that the variables must be standardized. Otherwise, certain variables, for example charpp, will have a bigger impact on the model than other variables with smaller mean, median, min and max (such us the subs variable). All the variables are a continuous variable, based on this information the variables are standardized as below:

1. Price: scaled
2. Charpp: scaled
3. Citations: scaled

The foundingyear variable contains date values (in year format, e.g., 1998). To be able to include this column into the model a new column is created which subtracts foundingyear variable from the current year, which is in 2000 according to the dataset. For example, if the foundingyear value of a given observation is 1963 the age column will be 37 (2000-1963).

## Data Analysis

This section describes the steps that are taken to analyze the variables before fitting them into a model.

Firstly, an initial regression is carried out without removing the outliers and transformation. Secondly each variable is analyzed with different transformations and the best fitting transformation is chosen for the end model. Lastly the outliers are removed from the data frame and regression analyses is carried out for the data frame without the outliers.

### Initial Regression

During this phase four models are created to see the relationship between subs and every other independent variable individually.

**Table 12. Regression analysis of individual independent variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| The number of Subscription by Economic Journal | | | | |
|  | DV: subs | | | |
|  | Citations | Price | Age | Charpp |
|  | | | | |
| citations | 0.097\*\*\* (0.013) |  |  |  |
| price |  | -0.165\*\*\* (0.039) |  |  |
| age |  |  | 1.006\*\* (0.475) |  |
| charpp |  |  |  | 0.014 (0.020) |
| Constant | 134.500\*\*\* (16.730) | 268.046\*\*\* (23.177) | 155.428\*\*\* (25.675) | 150.966\*\* (67.891) |
|  | | | | |
| Observations | 150 | 150 | 150 | 150 |
| R2 | 0.263 | 0.106 | 0.029 | 0.003 |
| Adjusted R2 | 0.258 | 0.099 | 0.023 | -0.003 |
| Residual Std. Error (df = 148) | 175.619 | 193.521 | 201.584 | 204.279 |
| F Statistic (df = 1; 148) | 52.914\*\*\* | 17.463\*\*\* | 4.490\*\* | 0.493 |
|  | | | | |
| *Significance* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | |

When independent variables are not controlled:

* Citations have a positive relationship (if citations increase by 1, subs increase by 0.97)
* Price is negatively correlated (if price goes up by 1, subs decrease by 0.16)
* Age is positively correlated (if age increases by 1, subs increase by 1)
* Charpp seems to have no significant relationship

**Table 13 . Regression results with controlled variables**

|  |  |
| --- | --- |
| The number of Subscription by Economic Journal | |
|  | DV: subs |
|  | |
| price | -0.187\*\*\* (0.034) |
| citations | 0.102\*\*\* (0.012) |
| charpp | 0.012 (0.016) |
| age | -0.051 (0.398) |
| Constant | 176.098\*\*\* (58.271) |
|  | |
| Observations | 150 |
| R2 | **0.397** |
| Adjusted R2 | 0.381 |
| Residual Std. Error | 160.485 (df = 145) |
| F Statistic | 23.899\*\*\* (df = 4; 145) |
|  | |
| *Significance:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |

Table 13 above shows the results of the regression analysis that is applied on the dataset without transformation. Price and subs have a negative relationship, meaning that if the price goes up by 1, the number of subscriptions decreases by 0.19 (if all the other variables are kept on average). Citations have a positive relationship, thus the more the citation a journal has, the more people are subscribed to the journal.

Other variables do not have a significant relation to the number of subs.

The R square tells that the model with no transformation can explain only 40% of the variation in the dependent variable, which is not sufficient.

### Identifying nonlinear relationships

This sections describes the analysis that are carried out to find the best fitting model depending on different transformations. Each variable is analyzed with different transformation (logarithmic, exponential and polynomial) and the transformation with the highest R-square is chosen to be used in the end model.

**Price**

Chart, scatter chart

Description automatically generated

**Figure 20 Trend between Subs and Price**

Figure 10 shows that there is a nonlinear relationship between price and subs. After trying different transformation, the best results in R-square is with logarithmic transformation.

**Citations**

Chart

Description automatically generated

**Figure 11 Trend between Subs and Citations**

As shown in Figure 11 there is a nonlinear relationship between citations and price. Polynomial transformation has the highest R-square.

**Charpp**

Chart, scatter chart

Description automatically generated

**Figure 12 Trend between Subs and Charpp**

Figure 12 shows that there is nonlinear relationship between Subs and Charpp, however even trying all different transformation there are no significant results. Meaning that this variable is not correlated with subs.

**Age**

Chart, scatter chart

Description automatically generated

**Figure 13 Trend between Subs and Age**

Between age and subs there is nonlinear relationship (Figure 13). Polynomial transformation has the highest R-Square.

## Regression results & Conclusion

### Regression results after transformation

**Table 15. Comparison of the model before and after the transformation**

|  |  |  |
| --- | --- | --- |
| The number of Subscription by Economic Journal | | |
|  | DV: subs | |
|  | Initial regression | Regression after transformation |
|  | | |
| price | -0.187\*\*\* (0.034) |  |
| citations | 0.102\*\*\* (0.012) |  |
| charpp | 0.012 (0.016) |  |
| age | -0.051 (0.398) |  |
| scale(log(price)) |  | -89.057\*\*\* (10.781) |
| scale(poly(citations, 2))1 |  | 87.778\*\*\* (10.255) |
| scale(poly(citations, 2))2 |  | -65.059\*\*\* (9.977) |
| scale(charpp) |  | -3.824 (9.750) |
| scale(poly(age, 2))1 |  | 2.256 (10.175) |
| scale(poly(age, 2))2 |  | -52.957\*\*\* (10.946) |
| Constant | 176.098\*\*\* (58.271) | 197.187\*\*\* (9.502) |
|  | | |
| Observations | 150 | 150 |
| R2 | 0.397 | 0.687 |
| Adjusted R2 | 0.381 | 0.674 |
| Residual Std. Error | 160.485 (df = 145) | 116.372 (df = 143) |
| F Statistic | 23.899\*\*\* (df = 4; 145) | 52.428\*\*\* (df = 6; 143) |
|  | | |
| *Significance:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | |

After the transformation price and citations remain significantly correlated to subs. The coefficients of price has changed drastically from -0.187 to –89. Citations impact has also changed. Interestingly in the second polynomial coefficients of age a has a significant relationship with subs, meaning that I(Age^2) have a significance.

R-square increased from 40% to 70% which makes the transformed model a better fitting model for the dataset.

### Regression results after removing outliers

**Table 16. Regression results with no outliers compared to transformed model**

|  |  |  |
| --- | --- | --- |
| The number of Subscription by Economic Journal | | |
|  | DV: subs | |
|  | Model No Outliers | Transformed model |
|  | | |
| scale(log(price)) | -68.504\*\*\* (8.558) | -89.057\*\*\* (10.781) |
| scale(poly(citations, 2))1 | 96.649\*\*\* (8.405) | 87.778\*\*\* (10.255) |
| scale(poly(citations, 2))2 | -13.002 (8.353) | -65.059\*\*\* (9.977) |
| scale(charpp) | 9.391 (8.125) | -3.824 (9.750) |
| scale(poly(age, 2))1 | -4.054 (8.178) | 2.256 (10.175) |
| scale(poly(age, 2))2 | -53.100\*\*\* (8.747) | -52.957\*\*\* (10.946) |
| Constant | 173.662\*\*\* (7.767) | 197.187\*\*\* (9.502) |
|  | | |
| Observations | 136 | 150 |
| R2 | 0.736 | 0.687 |
| Adjusted R2 | 0.724 | 0.674 |
| Residual Std. Error | 90.572 (df = 129) | 116.372 (df = 143) |
| F Statistic | 59.943\*\*\* (df = 6; 129) | 52.428\*\*\* (df = 6; 143) |
|  | | |
| *Significance:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | |

Table 16 above compares the results of the transformed model with the model without outliers. The are slight changes in the coefficients of the variables price and citations but overall, these variables remain significant. R-square has increased from 67% to 72% in the model without Outliers.

So it can be concluded that price and sub are negatively correlated. If the price increases by 1, the subs decrease by 68. If citations increase by 1 the subs increase by 96, thus there is positive relationship. Age and subs are negatively correlated, meaning if the I(Age^2) increase by 1 subs decreases by 52.

# Assignment 4. E-commerce Customers

## Data Description

## Data preparation

## Data Analysis

## Regression results & Conclusion

# Assignment 5. Crime Rate in Philadelphia

## Data Description

This section focuses on the dataset of Philadelphia\_Crime\_Rate, which contains the number of crimes per 1000 people for different cities in the state of Philadelphia.

The dataset contains the variables below:

1. Name: person's name
2. Country: name of cities
3. HousePrice: how expensive the house is
4. CrimeRate: the crime rate for a given city
5. MilesPhila: Distance from the city center
6. PopChg: changes in the population

The goal of this assignment is to detect whether there are outliers and if so remove them appropriately and to create a model that predicts the number of crimes (CrimeRate) based on the variables below:

1. HousePrice
2. MilesPhila
3. PopChg

## Data preparation

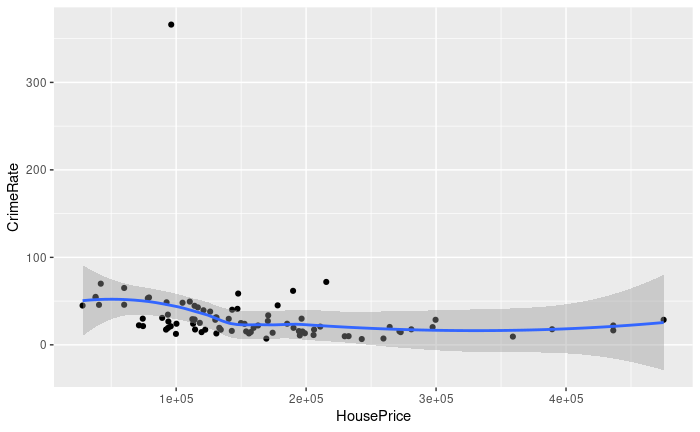
This section describes the steps that are taken to clean and prepare the data for the analysis.

A new data frame is created with the variables that will be used in the analysis (CrimeRate, HousePrice, MilesPhila, PopChg). Here below is the summary of the newly created data frame:

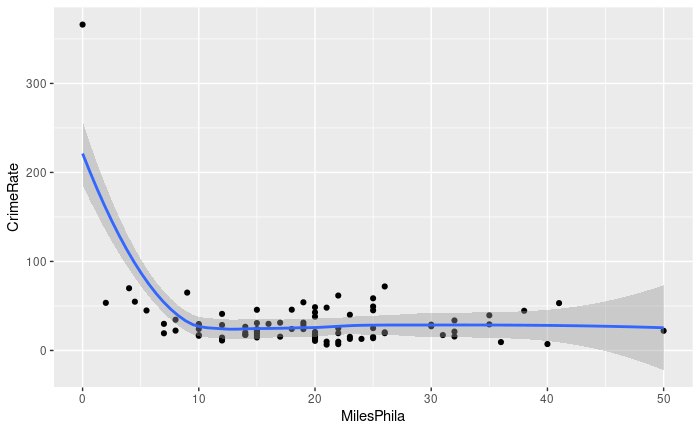
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | **Min** | **Median** | **Mean** | **Max** | **Observations** | **Missing values** |
| **HousePrice** | 28000 | 142811 | 161653 | 475112 | 85 | 0 |
| **CrimeRate** | 6.60 | 24.10 | 32.04 | 366.10 | 85 | 0 |
| **MilesPhila** | 0.00 | 20.00 | 19.52 | 50.00 | 85 | 0 |
| **PopChg** | - 9.200 | 1.600 | 2.449 | 26.900 | 85 | 1 |

## Data Analysis

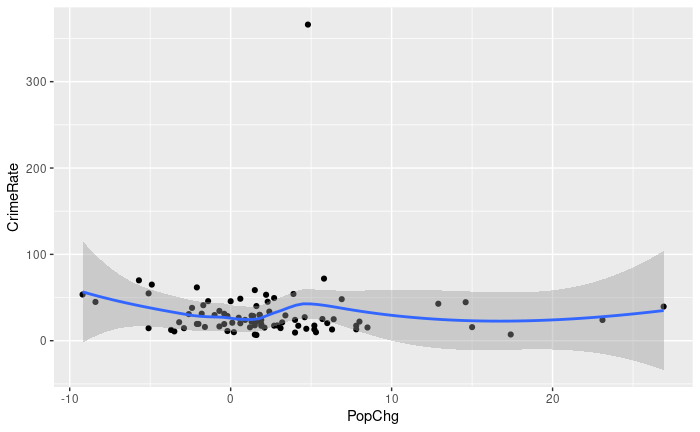
This section describes the methods that have been appllied to detect and remove the outliers



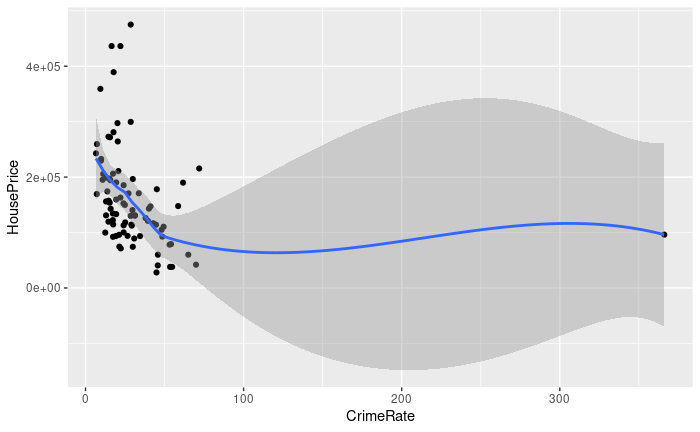
As shown in the figure above HousePrice seems to have one outliers



As shown in the figure there is one outlier in MilesPhila.



PopChg has one outlier as well.



CrimeRate has one outlier.

Initial Regression before removing outliers

Table 17

|  |  |
| --- | --- |
|  | |
|  | *Dependent variable:* |
|  |  |
|  | CrimeRate |
|  | |
| HousePrice | -0.0001\*\* (0.0001) |
| MilesPhila | -1.342\*\* (0.540) |
| PopChg | 1.201 (0.853) |
| Constant | 71.910\*\*\* (12.261) |
|  | |
| Observations | 84 |
| R2 | 0.131 |
| Adjusted R2 | 0.098 |
| Residual Std. Error | 38.068 (df = 80) |
| F Statistic | 4.011\*\* (df = 3; 80) |
|  | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |

## Regression results & Conclusion

In this section the final model is compared with the initial model with outliers

Table 18

|  |  |  |
| --- | --- | --- |
|  | | |
|  | *Dependent variable:* | |
|  |  | |
|  | CrimeRate | |
|  | (1) | (2) |
|  | | |
| HousePrice | -0.0001\*\*\* (0.00002) | -0.0001\*\* (0.0001) |
| MilesPhila | -0.031 (0.237) | -1.342\*\* (0.540) |
| PopChg | -0.259 (0.382) | 1.201 (0.853) |
| Constant | 41.671\*\*\* (5.016) | 71.910\*\*\* (12.261) |
|  | | |
| Observations | 81 | 84 |
| R2 | 0.209 | 0.131 |
| Adjusted R2 | 0.178 | 0.098 |
| Residual Std. Error | 14.266 (df = 77) | 38.068 (df = 80) |
| F Statistic | 6.770\*\*\* (df = 3; 77) | 4.011\*\* (df = 3; 80) |
|  | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | |

# Assignment 6. Graduate Earnings

## Data Description

In this section the analysis focuses on the dataset Graduate Earnings that contains data about the earnings of alumni from different schools in US. The alumni were asked what their earnings have been in the 5 years after graduation.

The dataset contains the columns below:

1. **School**: the name of the school
2. **Public**: whether the school is Public or Private (0 or 1)
3. **Location**: the location of the school
4. **Earn**: how much money an alumni makes after years
5. **SAT**: their SAT results
6. **Price**: tuition fee
7. **Need**\_fraction: How much of their tuition fee was paid for through a scholarship

The goal of this assignment is to present a model with Earn being the dependent variable, and the variables below as the independent variables:

1. Public
2. SAT
3. Price
4. Need fractions
5. Earn

## Data preparation

This section describes the steps that are taken to clean and prepare the data for the analysis.

A new data frame is created with the variables that will be used in the analysis (Earn, Public, SAT, Price, need\_fractions). Here below the summary of the newly created data frame:

**Table 19. Variables Summary**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | Min | Median | Mean | Max | Observations | Missing values |
| Earn | 28300 | 44900 | 45769 | 79700 | 600 | 0 |
| Public | 0.0000 | 0.0000 | 0.3774 | 1.0000 | 600 | 25 |
| SAT | 810 | 1120 | 1141 | 1550 | 600 | 87 |
| Price | 16500 | 44050 | 42249 | 70400 | 600 | 0 |
| Need\_fraction | 0.0700 | 0.5600 | 0.5729 | 1.0000 | 600 | 16 |

As shown in Table 19, there are 600 observations in the dataset with missing values in Public (25), SAT (87) and need\_fraction (16). The values differ proportionally which means that the variables must be standardized. Otherwise, certain variables, for example Earn and Price, will have a bigger impact on the model than other variables with smaller mean, median, min and max (such us the SAT, Public and need\_fractions variables). All the variables, excluding Public, are a continuous variable, based on this information the variables are standardized as below:

1. Earn: scaled
2. SAT: scaled
3. Price: scaled
4. Need\_fraction: scaled

Public variable is a categorical since it is binary, meaning if an observation is a Public school, it will 1 and if it is a private school, it will be 0. However, with the presence of missing values in this column it is not possible to create a dummy variable yet.

### Missing values

The data frame is copied into a new data frame so that missing values can be replaced in a separate data frame. This makes it easier to compare the regression results between the model with missing values and the model without missing values.

The variables Public, SAT and need\_fraction have missing values. Before replacing the missing values 3 new dummy columns are created that represent the missing values of the variable:

1. **Public\_missing:** if the column Public has a missing value, then the value is 1, else it is 0.
2. **SAT\_missing:** if the column SAT has a missing value, then the value is 1, else it is 0.
3. **Need\_fraction\_missing:** if the column need\_fraction has a missing value, then the value is 1, else it is 0.

Before replacing the missing values, a regression model is applied, in which the 3 dummy columns above are used as the dependent variable to see whether there are any associations between missing values and other variables.

## Data Analysis

In this section the newly created dummy columns that are representing the missing values in other columns are analyzed to see whether there are association between the missing values and other variables.

**Public\_missing**

**Table 20. Regression with missing values (Public\_missing)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | DV: Public\_missing | | | | | | | |
|  | Earn | Price | Need\_fraction | All variables | | All missing | |
|  | | | | | | | | |
| Earn | 0.0000\*\* (0.00000) |  |  | 0.000 (0.000) | |  | |
| Price |  | 0.00000 (0.00000) |  | 0.000 (0.000) | |  | |
| need\_fraction |  |  | -0.022 (0.049) | 0.000 (0.000) | |  | |
| Public |  |  |  | 0.000 (0.000) | |  | |
| SAT\_missing |  |  |  |  | 0.059\*\* (0.023) | |
| Need\_fraction\_missing |  |  |  |  | -0.042 (0.050) | |
| Constant | -0.070 (0.055) | 0.014 (0.023) | 0.055\* (0.029) | 0.000 (0.000) | | 0.034\*\*\* (0.009) | |
|  | | | | | | | | |
| Observations | 600 | 600 | 584 | 559 | | 600 | |
| R2 | 0.007 | 0.003 | 0.0003 |  | | 0.012 | |
| Adjusted R2 | 0.005 | 0.001 | -0.001 |  | | 0.009 | |
| Residual Std. Error | 0.199 (df = 598) | 0.200 (df = 598) | 0.203 (df = 582) | 0.000 (df = 554) | | 0.199 (df = 597) | |
| F Statistic | 4.188\*\* (df = 1; 598) | 1.655 (df = 1; 598) | 0.199 (df = 1; 582) |  | | 3.584\*\* (df = 2; 597) | |
|  | | | | | | | | |
| *Significance:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | | | | | |

Table 11 shows the results of the regression analysis with Public\_missing being the depended variable. The variable is the set into five different models. Overall it can be concluded that the missing values in Public columns is only associated with earn and SAT\_missing. However the R-square of all five models are too small to use this information during missing values replacement.

**SAT\_missing**

**Table 21. Regression with missing values (SAT\_missing)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | |
|  | DV: SAT\_missing | | | | | | |
|  | Earn | | Price | Need\_fraction | Public | All variables | All missing |
|  | | | | | | | |
| Earn | 0.00000 (0.00000) | |  |  |  | -0.00000 (0.00000) |  |
| Price |  | | -0.00000 (0.00000) |  |  | -0.00000 (0.00000) |  |
| need\_fraction |  | |  | -0.156\* (0.085) |  | -0.172 (0.111) |  |
| Public |  | |  |  | 0.031 (0.030) | -0.021 (0.075) |  |
| Public\_miss |  | |  |  |  |  | 0.182\*\* (0.072) |
| needf\_miss |  | |  |  |  |  | -0.013 (0.089) |
| Constant | 0.125 (0.097) | | 0.179\*\*\* (0.041) | 0.235\*\*\* (0.051) | 0.126\*\*\* (0.018) | 0.313\* (0.172) | 0.138\*\*\* (0.015) |
|  | | | | | | | |
| Observations | | 600 | 600 | 584 | 575 | 559 | 600 |
| R2 | 0.0001 | | 0.001 | 0.006 | 0.002 | 0.007 | 0.011 |
| Adjusted R2 | -0.002 | | -0.0004 | 0.004 | 0.0002 | -0.001 | 0.007 |
| Residual Std. Error | 0.353 (df = 598) | | 0.352 (df = 598) | 0.352 (df = 582) | 0.345 (df = 573) | 0.345 (df = 554) | 0.351 (df = 597) |
| F Statistic | 0.042 (df = 1; 598) | | 0.769 (df = 1; 598) | 3.366\* (df = 1; 582) | 1.093 (df = 1; 573) | 0.919 (df = 4; 554) | 3.251\*\* (df = 2; 597) |
|  | | | | | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | | | | |

Based on the information in Table 12 it can be concluded that there are no significant relationship between SAT\_missing and other variables. Only the variables need\_fraction and Public\_missing have correlation with SAT\_missing however the R-square of all models are not sufficient enough to include this association in to missing value replacement.

**Need\_Fraction\_missing**

**Table 22. Regression with missing values (need\_fraction\_missing)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |
|  | DV: Need\_fraction\_missing | | | | | |
|  | Earn | Price | Need\_fraction | Public | All variables | All missing |
|  | | | | | | |
| Earn | -0.00000 (0.00000) |  |  |  | 0.000 (0.000) |  |
| Price |  | 0.00000 (0.00000) |  |  | 0.000 (0.000) |  |
| need\_fraction |  |  | 0.000 (0.000) |  | 0.000 (0.000) |  |
| Public |  |  |  | -0.030\*\* (0.014) | 0.000 (0.000) |  |
| Public\_miss |  |  |  |  |  | -0.027 (0.033) |
| SAT\_miss |  |  |  |  |  | -0.003 (0.019) |
| Constant | 0.044 (0.045) | 0.003 (0.019) | 0.000 (0.000) | 0.039\*\*\* (0.009) | 0.000 (0.000) | 0.028\*\*\* (0.007) |
|  | | | | | | |
| Observations | 600 | 600 | 584 | 575 | 559 | 600 |
| R2 | 0.0003 | 0.003 |  | 0.008 |  | 0.001 |
| Adjusted R2 | -0.001 | 0.001 |  | 0.006 |  | -0.002 |
| Residual Std. Error | 0.161 (df = 598) | 0.161 (df = 598) | 0.000 (df = 582) | 0.164 (df = 573) | 0.000 (df = 554) | 0.161 (df = 597) |
| F Statistic | 0.151 (df = 1; 598) | 1.851 (df = 1; 598) |  | 4.481\*\* (df = 1; 573) |  | 0.366 (df = 2; 597) |
|  | | | | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | | | |

Overall conclusion of the regression that is applied on the missing values (in Table 20, Table 21 and Table 22) shows that even if there are significant relationships between the missing values and other variables, R-square overall is very low which is making it difficult to predict there are missing values and if they are associated with other variables.

Because the regression analysis on the missing values did not bring any accurate estimation, general methods below are applied for the replacement of the missing values:

* **Public**: since this is a binary variable (1 and 0) it is not suitable to replace the missing values with mean or any other mathematical calculation. There are 358 private schools (0) and 217 Public schools (1). Which leaves 25 missing values. These are replaced with the MOD of Public, meaning that the most common value, 0, is replaced with the 25 missing values.
* **SAT**: there are 87 missing values, which are replaced with the mean of SAT (without the NA’s being included during the mean calculation)
* **Need\_fraction:** 16 missing values are replaced with the mean of need\_fraction (without the NA’s being included during the mean calculation).

## Regression results & Conclusion

**Table 23. Regression results comparison before and after missing value replacement**

|  |  |  |
| --- | --- | --- |
|  | | |
|  | *Dependent variable:* | |
|  |  | |
|  | Earn | |
|  | Without NA | With NA |
|  | | |
| Public | 5,887.311\*\*\* (991.799) | 7,208.328\*\*\* (1,257.612) |
| SAT | 12.374\*\*\* (2.345) | 13.568\*\*\* (2.718) |
| Price | 0.246\*\*\* (0.031) | 0.271\*\*\* (0.041) |
| need\_fraction | -9,602.235\*\*\* (1,727.194) | -6,859.553\*\*\* (2,000.894) |
| Constant | 24,632.250\*\*\* (3,129.200) | 20,032.000\*\*\* (3,524.918) |
|  | | |
| Observations | **600** | **482** |
| R2 | 0.378 | 0.394 |
| Adjusted R2 | **0.374** | **0.389** |
| Residual Std. Error | 5,423.911 (df = 595) | 5,426.209 (df = 477) |
| F Statistic | 90.317\*\*\* (df = 4; 595) | 77.458\*\*\* (df = 4; 477) |
|  | | |
| *Significance:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | |

As shown in Table 24 the model with missing values (where the missing values are removed) has a slightly higher R-square, 39% compare to the model with missing values which is at 38%. The sample size differs of course between the models. So the model without missing values can be applied to a bigger population with 1% less accuracy. This results are before the transformation of the values.

**Table 24. Regression results comparison before and after missing value replacement by logarithmic model**

|  |  |  |
| --- | --- | --- |
|  | | |
|  | DV: Earn | |
|  | Transformed without NA | Transformed with NA |
|  | | |
| Public | 7,004.166\*\*\* (1,035.411) | 8,957.447\*\*\* (1,329.193) |
| scale(log(SAT)) | 1,713.893\*\*\* (286.296) | 1,856.848\*\*\* (351.746) |
| scale(log(Price)) | 4,299.057\*\*\* (509.862) | 5,088.719\*\*\* (686.301) |
| scale(log(need\_fraction)) | -1,619.435\*\*\* (279.099) | -1,257.585\*\*\* (323.223) |
| Constant | 43,236.160\*\*\* (436.026) | 42,341.790\*\*\* (560.497) |
|  | | |
| Observations | 600 | 482 |
| R2 | 0.367 | 0.389 |
| Adjusted R2 | 0.363 | 0.384 |
| Residual Std. Error | 5,471.086 (df = 595) | 5,445.680 (df = 477) |
| F Statistic | 86.212\*\*\* (df = 4; 595) | 76.054\*\*\* (df = 4; 477) |
|  | | |
| *Significance:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | |

Table 24 shows the results of the regression analysis with the data frame that has no missing values with the one with missing values, where logarithmic transformation is applied.

**Table 25. Regression results comparison before and after missing value replacement by polynomial model vs logarithmic.**

|  |  |  |
| --- | --- | --- |
|  | | |
|  | DV: Earn | |
|  | Without NA | With NA |
|  | | |
| Public | 6,065.962\*\*\* (1,081.181) |  |
| scale(poly(SAT, 2))1 | 1,750.692\*\*\* (303.523) |  |
| scale(poly(SAT, 2))2 | 1,103.960\*\*\* (233.128) |  |
| scale(poly(Price, 2))1 | 3,781.069\*\*\* (511.145) |  |
| scale(poly(Price, 2))2 | -514.718\* (298.491) |  |
| scale(poly(need\_fraction, 2))1 | -1,881.116\*\*\* (297.560) |  |
| scale(poly(need\_fraction, 2))2 | -45.795 (227.508) |  |
| scale(log(SAT)) |  | 1,856.848\*\*\* (351.746) |
| scale(log(Price)) |  | 5,088.719\*\*\* (686.301) |
| scale(log(need\_fraction)) |  | -1,257.585\*\*\* (323.223) |
| Constant | 43,575.480\*\*\* (447.617) | 42,341.790\*\*\* (560.497) |
|  | | |
| Observations | 600 | 482 |
| R2 | 0.401 | 0.389 |
| Adjusted R2 | 0.394 | 0.384 |
| Residual Std. Error | 5,336.220 (df = 592) | 5,445.680 (df = 477) |
| F Statistic | 56.565\*\*\* (df = 7; 592) | 76.054\*\*\* (df = 4; 477) |
|  | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | |

Table 25 shows the results of the regression analysis with the data frame that has no missing values with the one with missing values, where logarithmic transformation is applied on the data frame with missing values and polynomial transformation is applied on the data frame with no missing values.

In both models the variables SAT, Price, need\_fraction are significant. The model with polynomial transformation has however slightly higher R-square, meaning that the model can be applied 39% of the population accurately compare the model with missing values, which has a R-square of 38%.

Because the missing values are replaced either with the mean or the mod of the column that is missing the values, it might not represent the reality accurately since the missing values are replaced with calculated values. Because there is a small difference in the R-squere between the two models, it is like 1% difference, it is preferable to use the model with missing values, which will remove the missing values rather then filling it in with estimations.