Stockholms universitet

A Multiple Linear Regression Model Concerning The Swedish Board of Student Finance (CSN)

Yasmin Baghlani

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2020:20
Matematisk statistik
September 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2020:20**
http://www.math.su.se

A Multiple Linear Regression Model Concerning The Swedish Board of Student Finance (CSN)

Yasmin Baghlani*

September 2020

Abstract

This survey study is aimed to detect factors that affect the number of female students that take a loan, at the Swedish Board of Student Finance, by fitting a linear model. Data was abstracted from the National Board of Student Aid (Sweden) and included four independent variables such as education, unemployment, low income (below 60% of the median income) and high income (income above 200% of the median). Furthermore, the dependent variable of the study was the fraction of female loan borrowers from the National Board of Student Aid, in different Swedish municipalities. The main purpose of the study was to examine the effect of each of the independent variables on the fraction of loan recipients and to create a model to predict the percentage of future loan borrowers. We used assumptions of the linear regression model to get a fitted and valid multiple linear regression model, in which we looked at the outliers of the data and checked the assumptions of the regression models. As a result, we fitted a multiple linear regression model from data of 2015 and concluded that the fraction of people with an income less than 60% of the median, in each municipality, was insignificantly correlated with the fraction of loan borrowers. Regarding the significance of the independent variables, the number of educated people in each municipality was the most important variable, which had a positive relationship with the fraction of borrowers. Accordingly, a higher number of educated individuals in each municipality increased the fraction of borrowers. The second most significant independent variable was the fraction of individuals with an income higher than 200% of the median, and it was negatively correlated with the number of borrowers, indicating that the number of wealthy people in each municipality decreased the number of borrowers. The least significant variable affecting the fraction of borrowers was the unemployment rate. The higher the fraction of unemployed individuals in each municipality was, the higher the fraction of borrowers was as well. In order to make use of the regression model, the percentage of loan recipients was predicted using the fitted regression equation on the data from the next year 2016. The results suggest that the predicted values could explain 73.7% of the actual data variations.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: yasminba@kth.se. Supervisor: Ola Hössjer.

Contents

1 Introduction

Background

CSN, the National Board of Student Aid, is a Swedish government agency providing grants and loans to students. All citizen residents in Sweden have the right to receive grants or request loans for a limited period. Disbursements are regular (monthly) during the years of education. After this period, one must return the loan along with a low profit when starting a job after graduation. To qualify for receiving a loan from CSN, each student is only assessed individually, regardless of what municipality they live in or how much their family income is.

Purpose of Study

The purpose of this study was to examine factors affecting the percentage of female student loan borrowers from CSN in 2015, in different Swedish municipalities. The fraction of female borrowers was used as the dependent variable and the following four factors were examined as independent variables: The fraction of educated individuals, the fraction of unemployed individuals the fraction of individuals with an income less than 60% of the median. and the fraction of individuals with an income above 200% of the median, in each municipality.

Study aim

This study is aimed to find a model based on the multiple linear regression approach. To be able to get the best-fitted linear regression model that can reliably be used to predict future data on the number of female borrowers, we need to examine some assumptions that will be mentioned in the next section 'Problem Statement' and modify the model in order to avoid violations of it.

Problem Statement

To reach the aim of this study, we take a look at the issues that need to be solved. The study was carried out in order to answer the following questions:

1. Test the following five assumptions, which need to be satisfied when using the linear regression model, to produce a best-fitting model for the data:

   - The regression function is a linear combination of the independent variables.
   - There is no collinearity between independent variables.

- There are no outliers in the data.
- Error terms have a constant variance.
- Residuals follow the normal distribution.

2. To find a best-fitted regression model and use this model in order to predict the percentage of next year's female borrowers from CSN.

Limitations

The population surveyed in each municipality included individuals ranging in age from 18 to 65 years. The data was divided into two groups of male and female borrowers, to evaluate it. This yielded similar results for males and females in the initial analysis. To avoid duplication of analysis, we therefore decided only to study female borrowers. The number of people who are considered unemployed in this research is primarily due to the number of people registered with the Swedish Public Employment Service. There might be more unemployed, in any municipality, whose names are not registered there. As a result, this unregistered unemployment is not included in the analyses of this thesis. The variable 'Educated' was selected only from those who had at least one year of academic education at the university.

Outline of this thesis

This thesis is organized as follows: In Section 2 we outline some theory of the simple and multiple linear regression models. This is followed by a presentation of the student loan data set in Section 3, and a validation of the linear regression model for this data set in Section 4. The selected model is presented in Section 5, and after that Section 6 concludes. Finally, some of the tables and graphs are gathered in the appendix.

2 Theory

In this section, we describe the theory behind the statistical methods used in this research. Most of the theoretical concepts and methods in this section are taken from [**1**] and [**2**].

2.1 Linear Regression Models

Regression analysis is a powerful statistical method for analyzing data. The method is used for examining the linear relationship between a dependent variable and one or more independent variables. Two types of linear regression exist, simple linear regression, and multiple linear regression.

In a *simple linear regression model*, we can show the effect of one independent variable X on a dependent variable Y. If we instead have two or more independent variables $\mathbf{X} = (X_1, X_2, X_3, \ldots, X_p)$ to predict the outcome of a dependent variable Y, the model is referred to as *multiple linear regression.*

In this thesis regression analysis was used with simple and multiple linear regression models to show the effect of the independent variables (namely the fraction of *educated individuals*, *unemployed individuals*, *individuals with Income* $<$ 60% of median and *individuals with Income* $>$ 200% of median) on the dependent variable, the fraction of female borrowers (Women.Loan).

2.2 Simple Linear Regression

The simple linear regression model [**1**] looks like the following:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where the random error term ϵ_i is assumed to be independent and normally distributed with a mean of 0 and variance of σ^2. The parameter β_0 is the intercept and β_1 the slope of the line in the model. In simple linear regression analysis, β_1 is one of the most important quantities. If the value of β_1 is close to 0, there is no or little relationship between the independent and dependent variables. If the value of β_1 is large (positive or negative), on the other hand, it indicates a strong relationship. Since β_0 and β_1 are unknown, we need to estimate these parameters in order to fit the line (equation 1) to our data. For this purpose, we used the Least squares method, which will be explained further in section 2.4. First, we will take a look at the Pearson correlation coefficient (r), since it used in the formula for the Least squares method.

2.3 Pearson Correlation Coefficient (r)

The Pearson correlation coefficient (r) is a measure indicating the degree of correlation between two variables. Depending on different conditions, the Pearson correlation coefficient [**3**] is represented either by ρ or r. We use ρ if it measured in a population, and r when it measured in a sample. In this thesis we use sample data, so the correlation coefficient can be defined as follows:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_x} \cdot \frac{y_i - \bar{y}}{S_y}, \tag{2}$$

where the sample size is as n, whereas $\bar{x}$ and $\bar{y}$ are the mean values of $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ respectively in the sample and S_x, S_y are the standard deviations of these two data sets:

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad , \quad S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}. \tag{3}$$

The correlation coefficient value is in a range $-1 \leq r \leq +1$. Different values of r give different results. If the correlation between the two variables is $r = 0$, then there is no relationship between them. If $r = 1$ we have maximum (perfect) positive correlation, whereas if $r = -1$ the relationship between the two variables is maximal (perfect) and negatively correlated.

2.4 Least Square Method

To estimate the unknown parameters β_j, where $j = 0, 1$ in the (simple) linear regression model, we used the (ordinary) least squares method in order to compute estimates $\hat{\beta}_j$ of these two parameters. This is done by minimizing the difference between the observed variables and the regression line. In the least squares method the residual e_i is defined for each of the observations $i = 1, \ldots, n$. A residual measures the distance between an observed value and the fitted line, and it can be obtained as follows [**9**]:

$$e_i = y_i - \hat{y_i}, \tag{4}$$

where y_i is the observed value and $\hat{y_i}$ the predicted value. If the fitted regression line passes through the observed value, the residual is zero at that point. We square the residuals in the least square method to estimate $\hat{\beta}_j$.

$$\mathrm{SSR} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2. \tag{5}$$

The best-fitted parameter $\hat{\beta}_j$ is obtained when the sum of squared residuals (SSR) is minimized. This is obtained by finding the partial derivatives of

SSR with respect to both $\hat{\beta}_0$ and $\hat{\beta}_1$ and set them equal to zero. We get the regression estimated equation:

$$\hat{\beta}_0 = \arg\min_{\beta_0} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta} x_i)^2 = \arg\min_{\beta_0} \sum_{i=1}^{n} e_i^2, \tag{6}$$

$$\hat{\beta}_1 = \arg\min_{\beta_0} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta} x_i)^2 = \arg\min_{\beta_0} \sum_{i=1}^{n} e_i^2. \tag{7}$$

By solving these two equations, it can be seen that $\hat{\beta}_0$ and $\hat{\beta}_1$ are defined for ordinary least squares as follows:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{8}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{Cov(x,y)}{var(x)} = r \frac{S_y}{S_x}, \tag{9}$$

where r is the Pearson correlation coefficient, in the calculation of $\hat{\beta}_1$. We see from equation (9) that $\hat{\beta}_1$ is equal to the sample covariance between x and y divided by the variance of x, such that the higher covariance between x and y is, the higher the slope will be.

2.5 Hypothesis Testing

In the last section, we described how to estimate the parameters in the (simple) linear regression model with the (ordinary) least-squares method. For the regression model we have the opportunity to use a statistical hypothesis to test, e.g., if β_1, is significantly different from zero [which discussed in section 2.2.] A *statistical hypothesis* is a claim of a theory [**1**]. This will be applied for hypothesis testing, which is a way to assess the validity of the claim, the *null hypothesis*, against a counterclaim, the *alternative hypothesis*, using sample data. To calculate the validity of the hypothesis a *test statistic* will be used. A *linear hypothesis* is used when the test intends to find out if there is a linear relationship between the regression parameters, for instance that β_1 is significantly different from zero. In the case of multiple linear regression, we have the option to test if one or j independent variables (j=*1,...,p*) have any effect on the dependent variables, equivalent to test that β_j is significantly different from zero.

The two statistical hypotheses we will consider for the multiple linear regression model are:

- Null hypothesis (H_0): The null hypothesis, H_0, says that there is no statistical significance between X_j and Y

$$H_0 : \beta_j = 0 \tag{10}$$

- The alternative hypothesis (H_1): The alternative hypothesis, H_1 or H_a, is the opposite situation of the null hypothesis. By rejecting the null hypothesis, we use the alternative hypothesis. It assumes that the response Y is affected by X_j, and there is a relationship between the two variables.

$$H_1 : \beta_j \neq 0 \tag{11}$$

Hypothesis testing determines whether to accept or reject a claim H_0 about a model, based on the study sample by comparing the level of significance with the P-value.

P-value and Significance Level

The P-value is the probability of getting the test results at least as extreme as the results observed during the test, with the assumption that the null hypothesis is correct. By using the P-value, we can find out what the probability is that the result was obtained under the null hypothesis. We can compare the P-value to the *significance level* α. The significance level is defined as the pre-chosen probability of rejecting the null hypothesis when it is true. If the P-value is less or equal to the significant level α the result is statistically significant, and we can reject the null hypothesis H_0 and accept alternative hypothesis H_1. If instead of P-value is bigger then α, it gives a non-significantly result, and we accept the null hypothesis.

Test Statistic

In a hypothesis test, a test statistic is calculated from sample data and used to determine whether to reject the null hypothesis. It compares the data from the sample with the results expected under the null hypothesis. Such as it measures the degree of agreement between the data and the null hypothesis, to determine whether to reject the null hypothesis or not. Since the true distribution of the test statistics is unknown, we have the sampling distribution of the test statistic under the null hypothesis known as the *null distribution.* For instance, for a t-test, the test statistic has a t-distribution under the null hypothesis $\beta_j = 0$ that independent variable X_j has no association with the response variable Y. The test statistic is used to calculate the P-value. If the test statistic becomes too extreme (too small/large depending on the alternative hypothesis), the test's P-value becomes small enough to reject the null hypothesis. In this case, the data shows strong evidence against the assumptions of the null hypothesis.

T-Test

Assuming σ is unknown, the test statistic [**12**] used to test the null hypothesis $\beta_j = 0$, (10), is a t-statistic T. When using the linear regression t-test to sample data, we need to know the standard error of the estimate of the slope, the slope of the regression line assumed under the null hypothesis, and the degrees of freedom. The t statistic compares the difference between the estimated value of the slope and its assumed null value, with the standard error and follows, under the null hypothesis, a t-distribution. The t-statistic is defined as follows for the null hypothesis that the effect parameter of X_j is β_j:

$$T = \tfrac{\hat{\beta}_j - \beta j}{\mathrm{SE}_{\hat{\beta}_j}} \sim t(n-2). \tag{12}$$

The standard error of the $\hat{\beta}_j$ can be calculated as follows:

$$SE_{\hat{\beta}_j} = \frac{\sqrt{\mathrm{SSR}/(n-2)}}{\sqrt{\sum\limits_{i=1}^{n}(X_{ji}-\bar{X}_j)^2}}, \tag{13}$$

where SSR is defined in equation (5), X_{ji} is the value of X_j for observation i and $\bar{X}_j = \sum_{i=1}^{n} X_{ji}/n$.

For a chosen significance level (α), we can find the rejection region that corresponds to that value under the null distribution. If the null hypothesis is true:

$$P(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1 - \alpha. \tag{14}$$

The null hypothesis rejected if the P-value is less then α.

2.6 Multiple Regression Model

Let's now look at *Multiple Linear Regression* [**13**] that we have touched upon before, an extension of simple linear regression, were we have a vector $(X_{1i}, \ldots, X_{pi})$ of independent variables, instead of a single independent variable X_i, for every observation i. Just like in *Simple Regression* we have n observations, each with p different independent variables. For each observed value Y of the response, it will be predicted as a linear function of the different independent variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i, \; i = 1, 2, 3, \ldots, n, \tag{15}$$

where the error terms $\epsilon_i \sim N(0, \sigma^2)$ are independent and identically distributed (i.i.d).

As an example we have the statistical method using all the four independent variables, used in this thesis, $\boldsymbol{X}_i = (1, X_{1i}, X_{2i}, X_{3i}, X_{4i})$ at the same time, to predict the outcome of a dependent variable Y_i. This is defined as following:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i \tag{16}$$

where Y_i is the dependent variable (Women.Loan), X_{1i} the first independent variable (Unemployed), X_{2i} the second independent variable (Educated), X_{3i} the third independent variable (Income $<$ 60%), X_{4i} the fourth independent variable (Income $>$ 200%), β_0 the intercept, β_j the effect parameter of independent variable $j = 1, 2, 3, 4$, and $n = 290$ the number of observations.

It is also possible to write the multiple linear regression model (15) in matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{17}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ is the vector of regression parameters, $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)^T$ the design matrix, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ the response vector, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ a vector of error terms. The least squares estimator of the parameter vector is then written as:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{18}$$

2.7 Linear Regression Assumptions

Linear regression has many applications in everyday life. One of the advantages of this method is the ability to predict response variables for the future. To be able to get the best-fitted multiple linear regression model for the data that is reliable, we need to examine some underlying assumptions before fitting a model to data [**2**]. All of these assumptions must be satisfied so that we can fit the model with confidence, and if the assumptions are not satisfied, we must account for the model violations by adjusting the model.

Here are five assumptions that are needed to be satisfied, to produce a well-fitting model:

1. The regression function is a linear combination of the independent variables.
2. There is no collinearity between independent variables.
3. There are no outliers in the data.

4. Error terms have a constant variance.
5. Residuals follow the normal distribution.

Test the Assumptions

In this next part of the theory, we are looking further into the assumptions regarding the linear regression model that we just mentioned in the list above.

2.7.1 1-Linearity: The regression function is a linear combination of the independent variables.

The relationship between the dependent variable and the independent variable(s) should be linear and additive [**1**]. Since an additive relationship tells us that the effect of X_j on Y is independent of other variables. If the relationship is non-linear, the regression algorithm of the linear model will be inefficient, since it will not capture the actual trend or the response variable mathematically.

With simple linear regression, we can show the effect of an independent variable on a dependent variable and identify the linear relationship between them. We can investigate the relationship between two numerical variables by using a scatter plot, with the dependent variable along the vertical axis, and the independent variable along the horizontal axis.

2.7.2 2-Multicollinearity: There is no collinearity between variables.

Collinearity is known as the condition when two independent variables are highly linearly related. When an independent variable is a linear function of two or more independent variables, it referred to as *multicollinearity.* A high degree of multicollinearity among the independent variables in a regression model will lead to problems with fitting the model. The results will be difficult to interpret since the relationship between each independent variable and the dependent variable will not be estimated independently. This leads to a decline in the accuracy of the estimated coefficients, and the statistical power of the model is reduced. If independent variables are exactly linearly related, it is called perfect multicollinearity.

Variance Inflation Factor (VIF)

To detect multicollinearity, we can use VIF values for each independent variable [**1**]. Mathematically the VIF value, for an independent variable X_j is equal to the variance of $\hat{\beta}_j$ for the given model (with collinearity) divided by the variance of $\hat{\beta}_j$ for a model where X_j is uncorrelated with the other independent variables. VIF for independent variable X_j is obtained by dividing one by the tolerance:

$$\text{VIF}_j = \tfrac{1}{1-R_j^2} = \tfrac{1}{\text{Tolerance}_j}, \tag{19}$$

here R_j^2 is the coefficient of determination when regressing X_j against the other independent variables. The VIF value should be as small as possible, VIF$<$ 10 and Tolerance$>$ 0.1 are acceptable. If the VIF value is higher than 10, there is high collinearity between X_j and the other independent variables and further investigation is needed.

2.7.3 3-Outliers: There are no outliers in the data

In linear regression, an outlier is an observation with a large residual. In other words, it is a data point that differs fundamentally from the other observed data. One or more outliers will cause significant differences in the regression analysis. An outlier can indicate a sample abnormality, data entry error, or some other problem. One problem of the existence of the outlier is that it can affect the mean square error (MSE) of the parameter estimates, which we use in most parts of the analysis [**10**].

An observation is called influential when omitting that observation will fundamentally change the estimates of the regression coefficients. Influence can be a product of Outlyingness and *Leverage*, which will be explained in the next section. Leverage measures 'unusualness' of the p independent variables for an observation, and it is one out of several different methods we will look at now, through which we will identify influential observations.

Leverage

How much an independent variable deviates from its mean is the measurement of Leverage. In other words, leverage measures how far an observation on the independent variables is from the mean of the independent variables. A high leverage could have unusual effects on the estimation of regression coefficients [**16**]. A *point with high Leverage* is an observation with an extreme value on an independent variable. The Leverage formula in the special case of simple linear regression ($p = 1$) is:

$$h_{ii} = \tfrac{1}{n} + \tfrac{(x_i - \bar{x}^2)}{\sum\limits_{j=1}^{n}(x_j - \bar{x})^2}, \tag{20}$$

where n is the number of observations. More generally, for multiple linear regression with p independent variables it is possible to extend this definition of h_{ii}, as the ith diagonal element of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

When we have observations with Leverage $h_{ii} > (\frac{p+1}{n})$, they should be carefully examined. Observations with high Leverage will have Leverage scores. Some possible rules of thumb for declaring an observation i as a leverage point are the following lower bounds of h_{ii}: $(2 \times \frac{p+1}{n})$ if the sample size exceeds 30 or $(3 \times \frac{p+1}{n})$ if the sample size is lower than 30 [**15**]. In this analysis, we use the rule $h_{ii} > (2 \times \frac{p+1}{n})$.

Studentized deleted residuals

Studentized deleted residuals [**10**] is a way to detect outliers. The method refits the regression model with $n - 1$ observations after deleting an observation from the model, one at the time. The observed response values to the fitted values based on the models with the i^{th} observation deleted, will then be compared. The plot of the studentized deleted residuals include the deleted residual divided by its estimated standard deviation.

By using the studentized deleted residuals, we can identify possible outlier observations in the model. A studentized residual is:

$$t_i = \frac{\text{Deleted residual of observation } i}{\text{An estimate of its standard deviation}}, \tag{21}$$

which is equivalent to:

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{\text{MSE}_i(1-h_{ii})}}, \tag{22}$$

where $d_i = y_i - \hat{y}_{(-i)}$ is the residual of the deleted observation, $s(d_i)$ is the standard error and h_{ii} is the leverage of that observation, whereas $\text{MSE}_i = \text{SSR}_i/(n-p)$ is an estimate of the variance σ^2 of the error terms when observation i is deleted from the sum of squares [**8**]. The observation is removed in order to determine how the model behaves without this potential outlier. If an observation has a studentized deleted residual of absolute value higher than 3, it may be an outlier, and we must consider it in the next steps of the regression analysis.

Standardized deleted residuals

This method [**9**] is very similar to the studentized residual. The difference is that here, MSE is based on all observations. The studentized residuals method is more effective than the standardized residuals method. The latter residuals are defined as:

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sqrt{\text{MSE}(1-h_{ii})}}, \tag{23}$$

where e_i is the ordinary residual, defined in (4), $s(e_i)$ the standard error of that residual and $\text{MSE} = \text{SSR}/(n-p-1) = \hat{\sigma}^2$ an estimate of the variance σ^2 of the error terms, with SSR the sum of squared residuals (5).

Cook's Distance

Cook's Distance [**18**] is a general criterion for identifying influential data. This criterion combines residue information and leverage information. The minimum value of Cook's Distance D_i of observation i is zero, whereas higher values indicate that observation i influences the prediction of the response values of all the other observations. The Cook's distance of observation i is defined as:

$$D_i = r_i^2 \cdot \tfrac{h_{ii}}{p+1}, \tag{24}$$

where r_i is the standardized residual in (23), and h_{ii} the leverage, defined for $p = 1$ in (20) and more generally below that equation. A value of D_i larger than $> \frac{4}{n}$ signifies an influential observation and this should be accounted for in the next steps of regression analysis.

2.7.4 4-Homoscedasticity: Error terms have a constant variance.

Homoscedasticity means that we have the same variance ($Var(\epsilon_i) = \sigma^2$) of the error terms of the linear regression model, regardless of the values of the independent variables [**2**]. If the error term does not have a constant variance, we call it heteroscedasticity and have $Var(\epsilon_i) = \sigma_i^2$.

Breusch-Pagan and Koenker Score test

We use the Breusch-Pagan and Koenker score to tests for homoscedasticity in a linear regression model. By using these two methods, we can demonstrate whether the variance of the errors of the regression model is dependent on the values of the independent variables or if heteroscedasticity is present [**14**]. The test statistic LM of the Breusch-Pagan test is

$$\mathrm{LM} = nR^2, \tag{25}$$

where n is the number of observations, and R^2 the coefficient of determination when the squared residuals e_i^2 in (4) are regressed against a set of q independent variables that model possible heteroscedasticity. Under the null hypothesis of homoscedasticity, the distribution of LM is chi-square with q degrees of freedom.

Koenker Score test

The studentized version of Breusch-Pagan is the Koenker Score test; this test holds its null size better than other tests. If the test statistic has a P-value less than 5%, then the null hypothesis of homoscedasticity rejected, and heteroscedasticity assumed. In this case, the violation appears, and we should transform the data and also rebuild the model as described in the section of Transformation [**14**].

2.7.5 5-Normality: Residuals follow the normal distribution.

One of the most critical assumptions in the least-squares method is that the error terms should be normally distributed with $E(\epsilon_i) = 0$ and a variance $Var(\epsilon_i) = \sigma^2$. Using the following two methods, we examined the normality of the model [**2**].

Shapiro Wilk Test

To test the normality assumption of the error terms, we can use the Shapiro Wilk test. The test assesses the null hypothesis indicating that the collection $e_1, \ldots, e_n$ of residuals (4) from a fit of the multiple linear regression model comes from a normally distributed population. The test statistic is:

$$W = \frac{(\sum_{i=1}^{n} a_i e_{(i)})^2}{\sum_{i=1}^{n} (e_i - \bar{e})^2}, \tag{26}$$

where $e_{(i)}$ is the ith smallest residual from the fit of the regression model, $\bar{e} = (e_1 + \ldots + e_n)/n = 0$ when an intercept is included in the model, $(a_1, \ldots, a_n) = m^T V^{-1}/C$, where $m = (m_1, \ldots, m_n)^T$ and V is the mean vector and covariance matrix of the order statistics of an i.i.d. sample of size n from standard normal distribution, whereas $C = \sqrt{m^T V^{-1} V^{-1} m}$. The null distribution of W is determined by Monte Carlo. To have a normal model, the P-value the Shapiro Wilk test should be larger than 5%.

Kolmogorov Smirnov Test

The Kolmogorov Smirnov test is used to determine whether observations from a sample of size n follow a given distribution F. We may apply this test to the standardized residuals $r_1, \ldots, r_n$, defined in (23), and with F the standard normal distribution $N(0, 1)$. The empirical distribution function F_n of the standardized residuals is defined as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(r_i), \tag{27}$$

where the indicator function is:

$$I_{(-\infty,x]}(r_i) = \begin{cases} 1, & r_i \leq x, \\ 0, & r_i > x. \end{cases} \tag{28}$$

If sup_x is the supremum of the set of real numbers x, we have for a given cumulative distribution function $F(x)$, the Kolmogorov-Smirnov statistic such as:

$$D_\infty = \underset{x}{sup}[F_n(x) - F(x)]. \tag{29}$$

2.8 Transformation of Data

In the situation that we have a violation of one or more assumptions treated in Section 2.7, we might have to transform data. If the response variable grows exponentially as a function of some linear combination of the independent variables, we need to transform the dependent variable with the logarithm in order to change the model to increase linearly instead of exponentially. In that case we can use linear regression to specify a model ([**11**]).

$$\log Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i. \tag{30}$$

Inverse Logarithm(Anti-Log)

We fit the new regression model (30) with the log.Women.Loan variable in the Transformation section. Then we also need to transform the data back at the end of the calculation, in order to resume the original scale [**11**].

2.9 Variable Selection for Model Building

In the previous section, we have assumed that the independent variables included in the model are of importance. We have focused on the theory about the violation of the assumptions of linear regression and techniques to guarantee that the functional form of the model was correct. We will, in this section, acknowledge the method of choosing those variables that explain the data most easily. There are a few methods to select significant independent variables for a multiple regression model. Here we will present two different approaches, the *Entry method*, and *Stepwise procedures*.

2.9.1 Entry Method

The method, also known as the *standard method*, is a procedure for independent variables selection where all variables in the model are included in one single step.

2.9.2 Stepwise Procedures

We will introduce three stepwise procedures, where the algorithms are based on *AIC* (*Akaike's Information Criterion*). AIC is a score used to test how well a model fits the data, without over-fitting. AIC score is defined as

$$AIC = 2p - 2\log(\hat{L}), \tag{31}$$

where $\hat{L}$ is the maximized likelihood for the fitted model and p the number of estimated regression parameters in the model.

Description of the stepwise procedures:

- Forward Selection: In the forward selection method, we start with no variables in the model and then add one independent variable at a time to the model, in each step producing the model with smallest AIC, until no additional independent variables lead to a further reduction of the AIC criterion. .

- Backward Elimination: In the Backward Elimination method, all independent variables are first entered into the model and then removed from the model individually by the criteria of the largest decrease of the AIC-value. The algorithm stops when such a removal does not lead to a better fit.

- Stepwise Regression: In the stepwise model, each independent variable is added to the model step by step and then removed if it is non-significant.

The *stepwise regression* method is a combination of the backward elimination and forward selection techniques. We use the entry method in the first model when we examine the data, and then later we use a stepwise selection method to create a new model.

2.10 Goodness of fit of the model

2.10.1 Coefficient of Determination (R^2)

The Goodness of fit of the model can determined in various ways, such as testing whether the residuals of the regression fit have the prescribed properties [**3**]. Two other quantities that we can use to decide how well a model fits data are the Coefficient of Determination R^2 and the adjusted R^2. For the fitted model to explain the dependent variable, it must have a sufficiently large R^2. This quantity R^2 corresponds to the fraction of the variance of the dependent variable that is explained by the independent variables. As the explanatory factor is closer to one, the more successful is the model in predicting the dependent variable, and the fitting power of the model is enhanced. The values of the coefficient of determination fall within a range $0 \leq R^2 \leq 1$. For simple linear regression R^2 is essentially a squared value of the coefficient of correlation r of Section 2.3, and it demonstrates the percentage of variation in Y caused by the single independent variable X. A similar interpretation of R^2 is possible for multiple linear regression, with R^2 the square of a multiple correlation coefficient between the dependent variable and all p independent variables. A higher R^2 value corresponds to a more optimal model.

We can define R^2 as follows:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}, \tag{32}$$

where SSR is the residual sum of squares in (5), SSE is the explained sum of squares and SST = SSE + SSR the total sum of squares.

2.10.2 Adjusted (R^2)

The R^2_{adj} indicates how much the dependent variable varies due to the independent variables. An addition of new independent variables always leads to that the coefficient of determination R^2 increases, possibly causing a false increment of the ratio due to overfitting, when the added independent has no effect. The R^2_{adj} method corresponds to a criterion which does not create an incorrect increase when additional non-significant independent variables are added to the model. This method modifies the non-adjusted coefficient of determination by considering variations due to overfitting, explained by the independent variables affecting the dependent variable. Moreover, the small difference between R^2 and R^2_{adj} is an adjustment for the number of degrees of freedom $n - p - 1$ of the residuals. The exact definition of the adjusted coefficient of determination is

$$R^2_{\text{adj}} = 1 - \frac{\frac{\text{SSR}}{\text{n-p-1}}}{\frac{\text{SST}}{\text{n-1}}} = 1 - \frac{\text{SSR}}{\text{SST}}\frac{\text{(n-1)}}{\text{(n-p-1)}} = 1 - \frac{(1-R^2)\text{(n-1)}}{\text{n-p-1}}, \tag{33}$$

where p is the number of regression parameters and n is the number of observations.

2.10.3 Prediction

We use the fitted regression model in order to predict the values of the dependent variable that correspond to new values of the independent variables. By plotting the predicted value against the observed value for the dependent variable, we look at the value of R^2_{adj} and R^2 to decide if they indicate a relatively acceptable fit of the model.

Suppose the log transformed model (30) is used for our data set of female student loan borrowers. After predicting the log response, by taking the inverse logarithm of the obtained predictions, we may predict the percentage of the future female borrowers in all municipalities.

3 Data Description

IBM SPSS [17] software is used to perform the statistical analysis in this thesis. The data is provided by CSN [4] , Statistiska Centralbyrån SCB [5], and the Swedish Public Employment Service [6] . We used data from year 2015 for the regression analysis, and to predict the future fraction of borrowers, we used data from year 2016.

Table 1: The descriptive statistics of the data in the year 2015

Variable Type	Variable Name	Number of observations n	Minimum Statistic	Maximum Statistic	Mean Statistic	Std. Deviation Statistic
Dependent	Women.Loan	290	1,19	8,90	2,3621	0,92689
Independent	Unemployed	290	2,30	15,20	7,7359	2,83898
	Educated	290	11,37	46,79	19,4747	6,06225
	Income $< 60\%$	290	5,50	29,00	14,5497	4,10610
	Income $> 200\%$	290	1,20	36,60	5,7076	3,92927

The number of observations in this study is based on the number of Swedish municipalities ($n = 290$) and there is no missing data. The dependent variable Women.Loan refers to the fraction of women, in each municipality, with a loan from CSN, whereas the four independent variables correspond to the fraction of women within each municipality that are unemployed, are educated, have a low Income ($< 60\%$) and have a high Income ($> 200\%$) respectively.

In Table 1, we introduce all the variables (dependent and independent) with a descriptive statistics summary. For further information, see Table 19 and 20 in the appendix. There the five municipalities with highest and lowest incomes, along with their values, are presented in Table 19. In Table 20 the five municipalities with highest and lowest middle (median) income are displayed.

The reason for including the variable 'Educated' is that although there is no reason to believe that all educated individuals necessarily received loans, there might still be a positive association between the fraction of educated individuals and the fraction of borrowers. Regarding this variable 'Educated' we considered individuals with at least one year of academic study as educated.

According to the information available at the SCB and the Swedish Public Employment Service, students are not considered unemployed [7], and therefore one might expect that the fraction of unemployed (the value of 'Unemployed') is associated with the fraction of women with a loan from CSN. We considered 'Income $> 200\%$' (at least twice as much as the 100% median income) and 'Income $< 60\%$' (less than 60% of the median income) as two separate independent variables. The reason was that many students are typically enrolled in full-time studies and therefore cannot have a high income. For this reason, one might expect a positive (negative) association between Income $< 60\%$ (Income $> 200\%$) and the fraction of borrowers. However, if many individuals study part of full time, these two associations may also go in the other direction.

4 Checking model assumptions of the regression model and transformation

As described in Section 2.7, we will examine the five mentioned assumptions as follows in this section:

4.1 Linearity

Simple Linear Regression

The first assumption we examined, in order to get the best-fitted regression model, was the linearity. We used simple linear regression to develop four models with the dependent variable 'Women.Loan' against each one of the four independent variables, to see if the regression models were linear or non-linear.

Scatterplots of the four independent variables against the dependent variable 'Women.Loan' are shown in Figures 11 to 14 on Pages 44 to 45 in the appendix, were they indicate that the models are linear. The four fitted simple linear regression models were the following:

$$\begin{aligned}
&\text{Model 1: } \hat{Y}_i = 2.39 - 0.004(\text{Unemployed})_i,\\
&\text{Model 2: } \hat{Y}_i = 0.36 + 0.1(\text{Educated})_i,\\
&\text{Model 3: } \hat{Y}_i = 2.77 - 0.03(\text{Incom} < 60\%)_i,\\
&\text{Model 4: } \hat{Y}_i = 2.02 + 0.06(\text{Income} > 200\%)_i,
\end{aligned}$$

where $Y_i = \text{Women.Loan}_i$ corresponds to the fraction of female borrowers in municipality i.

Table 2: This table gives a summary of four distribution diagrams of simple linear regressions of independent variables against the 'Women.Loan' dependent variable. The standardized coefficients refer to parameter estimates where each independent variable has been rescaled to a variance of 1.

Coefficient Summary								
		Unstandardized	Coefficients	Standardized Coefficients				
Model		β	Std.Error	β	R^2	R^2_{adj}	t	P-value
1	(Intercept)	2,393	0,158				15,099	0,000
	Unemployed	-0,004	0,019	-0,012	0,000	-0,003	-0,208	0,836
2	(Intercept)	0,363	0,136				2,665	0,008
	Educated	0,103	0,007	0,671	0,451	0,449	15,376	0,000
3	(Intercept)	2,769	0,200				13,880	0,000
	Income < 60%	-0,028	0,013	-0,124	0,015	0,012	-2,120	0,035
4	(Intercept)	2,021	0,093				21,702	0,000
	Income > 200%	0,060	0,013	0,253	0,064	0,061	4,445	0,000

Table 2 summarizes the four simple linear regression models in which only one independent variable is included. The variables 'Educated', 'Income < 60%', and 'Income > 200%' were statistically significant at the 5% level while the 'Unemployed' model had a P-value of 0.836 $> 5\%$, indicating that the relationship between the variables of 'Women.Loan' and 'Unemployed' were non-significant.

4.2 No Collinearity

4.2.1 Pearson Correlation

In this section we investigate the correlations between all pairs of variables, the dependent variable as well as the four independent variables .

Table 3: The Pearson correlation demonstrating a significant correlation between 'Educated', 'Income $< 60\%$', 'Income $> 200\%$' and 'Women.Loan', and a non-significant correlation between 'Unemployed' and 'Women.Loan'

Correlations^c						
		Women.Loan	Unemployed	Educated	Income $< 60\%$	Income $> 200\%$
Women.Loan	Pearson Correlation	1	-0,012	0,671**	-0,124*	0,253 **
	P-value (2-tailed)		0,836	0,000	0,035	0,000
Unemployed	Pearson Correlation	-0,012	1	-0,388**	0,657**	-0,527**
	P-value (2-tailed)	0,836		0,000	0,000	0,000
Educated	Pearson Correlation	0,671**	-0,388**	1	-0,588**	0,782**
	P-value (2-tailed)	0,000	0,000		0,000	0,000
Income $< 60\%$	Pearson Correlation	-0,124*	0,657**	-0,588**	1	-0,693**
	P-value (2-tailed)	0,035	0,000	0,000		0,000
Income $> 200\%$	Pearson Correlation	0,253**	-0,527**	0,782**	-0,693**	1
	P-value (2-tailed)	0,000	0,000	0,000	0,000	

**. Correlation is significant at the 0.01 level(2-tailed)
*. Correlation is significant at the 0.05 level (2-tailed).
a correlation is statistically significant if its P-value (2-tailed) < 0.05.
c. Based on $n = 290$ observations.

The correlations between the variables are presented in Table 3. It shows the significance between the four independent variables (namely 'Unemployed', 'Educated', 'Income $< 60\%$' and 'Income > 200) and the variable 'Women.Loan'. The only variable that had a non-significant correlation with the dependent variable was 'Unemployed' with correlation $r = -0.012$ and a P-value of 0.836.

The highest correlation coefficient for 'Women.Loan' was related to the variable 'Educated' with $r = 0.671$. The positive and significant correlation coefficient indicated that the municipalities with more educated females received more loans than other municipalities.

4.2.2 Multicollinearity

In the previous section, we used the Pearson correlation coefficient to study the relationship between pairwise variables. We will now investigate multicollinearity, which occurs when two or more independent variables are highly correlated. We have to do this before the variable selection; thus, stepwise regression does not perform as well with multicollinearity present. Multicollinearity will be measured in terms of the VIF value.

Table 4: Results of fitting a multiple linear regression model in order to check multicollinearity in data using the VIF and tolerance indices, where VIF $<$ 10 and Tolerance $>$ 0.1 are acceptable. For the standardized coefficients the independent variables have been rescaled so that their variance is 1.

Coefficients^a							
	Unstandardized Coefficients		Standardized Coefficients			Collinearity Stat	
Modell	β	Std. Error	β	t	P-value	Tolerance	VIF
(Intercept)	-1,293	0,246		-5,265	0,000		
Unemployed	0 ,018	0,015	0,055	1,182	0,238	0,552	1,811
Educated	0,189	0,009	1,235	22,209	0,000	0,380	2,629
Income $<$ 60%	0,041	0,012	0,180	3,297	0,001	0,394	2,535
Income $>$ 200%	-0,132	0,015	-0,559	-8,880	0,000	0,297	3,364
a.Dependent Variable: Women.Loan							
Method:Entry							

Table 4 shows that all the variables imported to the model have VIF less than 10; therefore, the problem of multicollinearity in the model is not severe. In this table, if we look at the 'Unemployed' variable, we observe that tolerance is 0.552; this means that if we run a multiple regression with the 'Unemployed' variable as a dependent variable in Table 4, then we have $R^2 = 0.448$. This corresponds to a value $1 - 0.448 = 0.552$ of $1 - R^2$, which is the same as the tolerance of this variable shown in Table 4. As a result, we realized that there is no severe collinearity between the variables in the model.

Multiple Regression

Furthermore, multiple regression between the dependent and independents variables, listed in Table 4, shows that the 'Unemployed' variable was the only non-significant variable in the model with a P-value of $,238 > 5\%$, whereas the other variables were statistically significant at significance level 0.05.

The fitted multiple regression model formulated as follows:

$$\hat{Y}_i = -1,293 + 0,018(\text{Unemployed})_i + 0,189(\text{Educated})_i + 0,041(\text{Income} < 60\%)_i - 0,132(\text{Income} > 200\%)_i \tag{34}$$

The adjusted coefficient of determination of the model in Table 5 is $R^2_{adj} = 66\%$, and the coefficient of determination is $R^2 = 67\%$. Because this value is close to 1 and the difference between these values is meager, it demonstrates that the model fits data reasonably well.

Table 5: Goodness-of-fit summary, with information about correlation, coefficient of determination and adjusted coefficient of determination of the multiple regression model from Table 4

Model Summary^b				
Model	R	R^2	R^2_{adj}	Std. Error of the Estimate
1	0.815^a	0,665	0,660	0,54040

a. Predictors: (Intercept), Income > 200%, Unemployed, Income < 60%, Educated
b. Dependent Variable: Women.Loan

Until now, we have examined the linearity in simple linear regression, correlation, multicollinearity, and multiple regression between dependent and independents variables. In the next section, we will investigate observations from all municipalities to test if there are any outliers in the data.

4.3 Diagnostic of outliers

In this section, we investigated the presence of outliers by using these three methods:

1. Outlier
2. Leverage
3. Cook's distance

Note that we did not delete any detected outliers from the observations in this section.

Before we begin to study outliers, Table 6 presents residual statistics. It summarizes the nature of the residuals and predicted values from the model of Table 4, which give a better understanding of the distribution of values that the model predicts. Moreover, by studying Table 6 we can predict the existence of outliers by observing the range of the standardized deleted residuals r_i.

Table 6: Residual Statistics: The minimum, maximum and standard deviation of the residuals, from the model fit of Table 4

Parameters	Min	Max	Mean	Std. deviation	n
Predicted Value	0,864	7,110	2,362	0,756	290
Deleted residual	-1,317	2,410	0,000	0,537	290
Std. Predictive Value	-1,982	6,283	0,000	1,00	290
Std. deleted residual	-2,437	4,460	0,000	0,993	290

Dependent Variable: Women.Loan (percentage of borrowers in each municipality)

In order to find whether we have outliers or leverage data points, we look at the standardized, deleted residuals r_i. Because the value of the maximum 4.46 of the standardized residuals is larger than 3.29, we know that we have an outlier in the data.

4.3.1 Outliers

In order to find outliers, we use a histogram of the studentized deleted residuals t_i, plotted together with a standard normal density curve. As shown in Figure 1, about five observations were outside central region of the standard normal density curve.

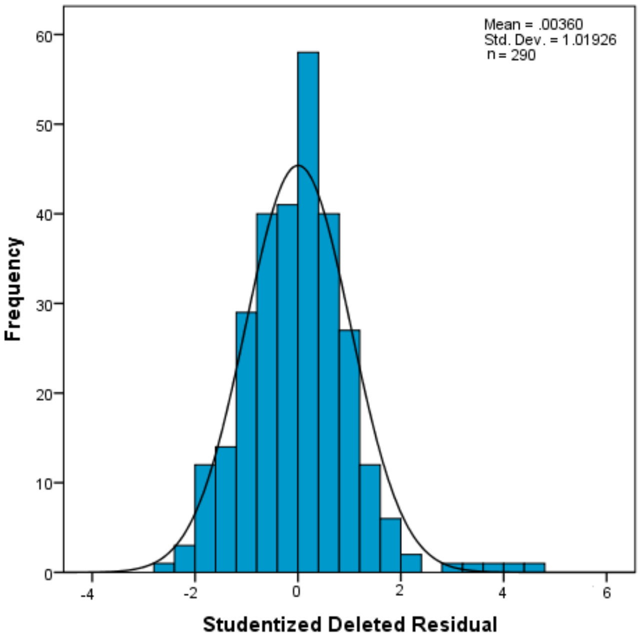

Figure 1: A relatively normal distribution of the studentized deleted residuals t_i, with SD≈ 1 and Mean≈ 0 is observed. At the right tail of the distribution, several outliers are found.

Table 7: Casewise diagnostics for the municipalities with a studentized deleted residual absolute value higher than 2

Casewise-Diagnostics[a]					
Case Number	Municipality	Studentized Deleted Residuals	Women.Loan	Predicted Value	Residual
30	Heby	2,093	2,65	1,531	1,115
31	Tierp	2,313	2,90	1,668	1,236
32	Uppsala	4,059	7,32	5,252	2,071
51	Linköping	3,442	6,51	4,718	1,788
84	Kalmar	2,801	5,34	3,856	1,482
97	Staffanstorp	-2,216	1,85	3,029	-1,180
104	Lomma	-2,339	2,15	3,378	-1,229
118	Lund	3.699	8.90	7,110	1,785
126	Simrishamn	-2,038	1,50	2,596	-1,093
188	Hammarö	-2,487	2,25	3,563	-1,317
274	Umeå	4,793	7,76	5,351	2,410

a. Dependent Variable: Women.Loan

Table 7 demonstrates the number of studentized deleted residuals with with an absolute value higher than 2. Based on the municipalities, we have eleven observations with their absolute studentized deleted residuals higher than 2. Four of the municipalities (i.e., Umeå, Uppsala, Lund, and Linköping) had their studentized deleted residual values higher than 3, indicating the high degree of their outlyingness.

4.3.2 Leverage

We examined the leverage values in order to find data with high potential effects. Figure 2 demonstrates the leverage values h_{ii} against the studentized deleted residuals. The municipality Danderyd had the highest value for the leverage and likely had an effect on the regression coefficient estimates.

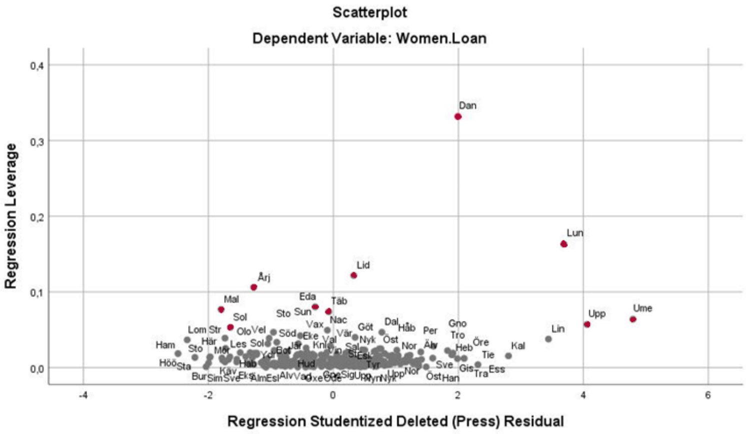

Figure 2: The scatterplot of leverage values against the studentized deleted residuals. Gray points represent the municipalities. The highest leverage is for the Danderyd municipality.

In general, points whose leverages that are greater than $2(p+1)/n$ must be considered with caution. The leverage values we need consider in this study are higher than $2(4+1)/290 = 0.034$. Table 8 shows leverage values for all ten observations with values higher than 0.034. Therefore, they need to be further investigated.

Table 8: Municipalities with leverage values larger than 0.034.

Case Number	Municipality	Statistic
15	Danderyd	0,332
118	Lund	0,164
22	Lidingö	0,122
192	Årjäng	0,106
185	Eda	0,080
117	Malmö	0,077
14	Täby	0,074
274	Umeå	0,064
32	Uppsala	0,057
21	Solna	0,053

4.3.3 Cook's distance

In order to identifying highly influential data, Figure 3 demonstrates a plot of Cook's values against the studentized deleted residuals. The Lund, Danderyd, Umeå and Uppsala municipalities showed the highest values for the Cook's distance and in Table 8 they were also among the municipalities that showed the most considerable influence on the regression model.

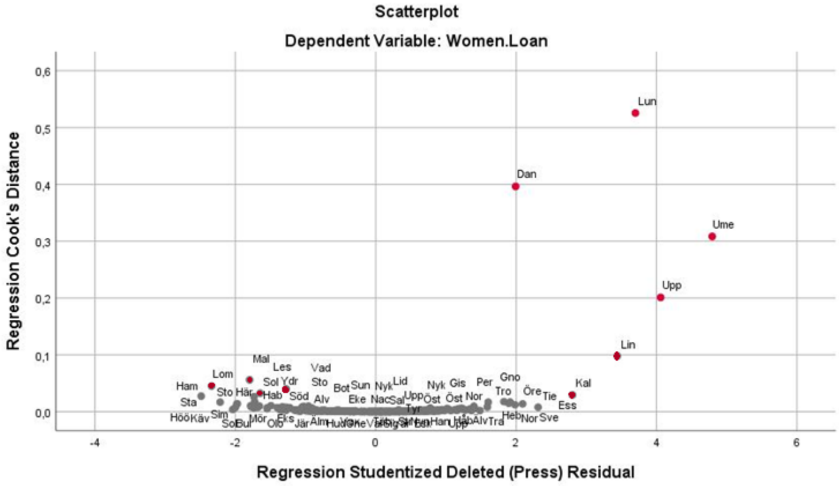

Figure 3: A scatter plot of Cook's distance values against the studentized deleted residuals. Gray points in the plot represent the municipalities. The highest values of the Cook's distance are observed for the Lund, Danderyd, Umeå and Uppsala municipalities.

The threshold for Cook's distance is $4/n$; therefore, points with values greater than $4/290 = 0,014$ along the vertical axis for this data set are is considered to be highly influential.

Table 9: Cook's distances. A list of observations with influence values higher than $0,014$.

Case Number	Municipality	Cook's distance
118	Lund	0,525
15	Danderyd	0,396
274	Umeå	0,308
32	Uppsala	0,201
51	Linköping	0,098
117	Malmö	0,056
104	Lomma	0,045
192	Årjäng	0,040
21	Solna	0,032
84	Kalmar	0,029

Table 9 demonstrates Cook's values for the top ten observations. As can be noticed, each one of these ten observations has a Cook's distance greater than 0.014; hence, they should be further investigated.

4.3.4 Combination of residual data, leverage and Cook's distance

We combined all the outliers we identify, with all three previous methods listed in Table 10. This was done in order to compare all the outliers we found.

Table 10: Outlier statistics: Comparison of influential level of the outlying municipalities in a combined table based on all three outlier detection methods.

Case Number	Municipality	Stud. Deleted Residual	Cook's.Distance	Centered Leverage Value
14	Täby	.	.	0,074
15	Danderyd	.	0,396	0,332
21	Solna	.	0,032	0,053
22	Lidingö	.	.	0,122
30	Heby	2,093	.	.
31	Tierp	2,313	.	.
32	Uppsala	4,059	0,201	0,057
51	Linköping	3,442	0,098	.
84	Kalmar	2,801	0,029	.
97	Staffanstorp	-2,216	.	.
104	Lomma	-2,339	0,045	.
117	Malmö	.	0,056	0,077
118	Lund	3,699	0,525	0,164
185	Eda	.	.	0,080
188	Hammarö	-2,487	.	.
192	Årjäng	.	0,040	0,106
274	Umeå	4,793	0,308	0,064

The conclusion of Table 10 is that the Umeå, Lund, and Uppsala municipalities are considered to be outliers by all three methods. Moreover, according to Cook's and leverage indicators, the Danderyd and

Lund municipalities are highly influential. We must therefore examine these observations more carefully in the transformation Section 4.6 to see if there exists an outlier in the data even after we transform the dependent variable.

4.4 Homoscedasticity

The method used to study the variation of variance is to inspect the standardized residuals r_i and plot them the predicted values. Figure 4 shows a plot of the standardized values of the residuals against the predicted values. The variance increases somewhat with the increase of predicted values.

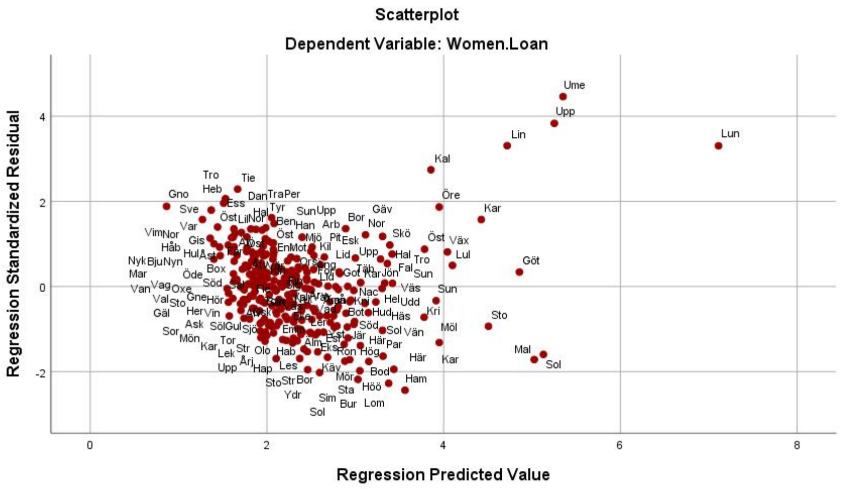

Figure 4: Scatter plot of standardized regression residual r_i against predicted values. (Red hollow squares represent municipalities)

4.4.1 Breusch-Pagan and Koenker tests

We use the Breusch-Pagan and Koenker tests to calculate the probability that the error term ϵ_i has constant variance. If the P-value is less than 5%, then we say that the homoscedasticity assumption has been violated and the model has heteroscedasticity or if it is homoscedastic.

Table 11: Breusch-Pagan and Koenker tests, testing if the model is Homoscedasticity or Heteroscedasticity

Breusch-Pagan and Koenker test statistics and P-values		
Test	LM	P-value
BP	160,797	0,000
Koenker	81,153	0,000
If P-value < 5%, reject the null hypothesis of homoscedasticity.		

Table 11 shows that the Breusch-Pagan and Koenker tests both give a P-value of $0,000 < 5\%$, which means that the error terms show signs of heteroscedasticity and do not have a constant variance. We need to re-examine this assumption later in the transformation Section 4.6.

4.5 Normality

One of the properties of the the most commonly used linear regression model is that the error terms follow the normal distribution. Table 12 shows descriptive statistics of the residuals. By looking at the table, we obtain important information about the unstandardized residuals.

The standard normal distribution with a symmetrical shape has a skewness and kurtosis equal to zero, but the residuals have skewness of 0.627 and a kurtosis 2.018, which indicates that the residuals are not normal.

Table 12: Descriptives statistics showing information about the unstandardized residuals of the model fit from Table 4

			Descriptive statistic	Std. Error
Unstandardized Residual	Mean		0,0000000	0,03151323
	95% Confidence Interval for Mean	Lower Bound	-0,0620245	
		Upper Bound	0,0620245	
	5% Trimmed Mean		-0,0152979	
	Median		0,0119347	
	Variance		0,288	
	Std. Deviation		0,53665094	
	Minimum		-1,31713	
	Maximum		2,41025	
	Range		3,72738	
	Interquartile Range		0,69278	
	Skewness		0,627	0,143
	Kurtosis		2,018	0,285

4.5.1 Kolmogorov-Smirnov and Shapiro-Wilk

We can also use the Kolmogorov-Smirnov and Shapiro-Wilk tests to verify the normality of the residuals. Table 13 demonstrates these two normality tests of the residuals.

Table 13: Using the Kolmogorov-Smirnov and Shapiro-Wilk tests to test the normality of the residuals, obtained from the model fit of Table 4.

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	P-value	Statistic	df	P-value
Unstandardized Residual	0,047	290	0,200*	0,971	290	0,000
*. This is a lower bound of the true significance.						

The result in Table 13 demonstrates that the Shapiro-Wilk test with a P-value of $0,000 < 5\%$ rejects normality of the residuals, whereas the Kolmogorov-Smirnov test with a P-value of $0,200 > 5\%$ did not reject the null hypothesis of normally distributed residuals.

We can, for further study of the normality, look at the normality of the residuals in Figure 5, the histogram of the unstandardized residuals based on the frequency of the residuals that fall into different bins. This plot shows that the histogram of the residuals has a normal shape and the residuals are approximately normally distributed with an approximate mean of zero and an approximately standard deviation of 0.54. However, a few data points at the right tail of the distribution were outside the concerned range and appear to be outliers.

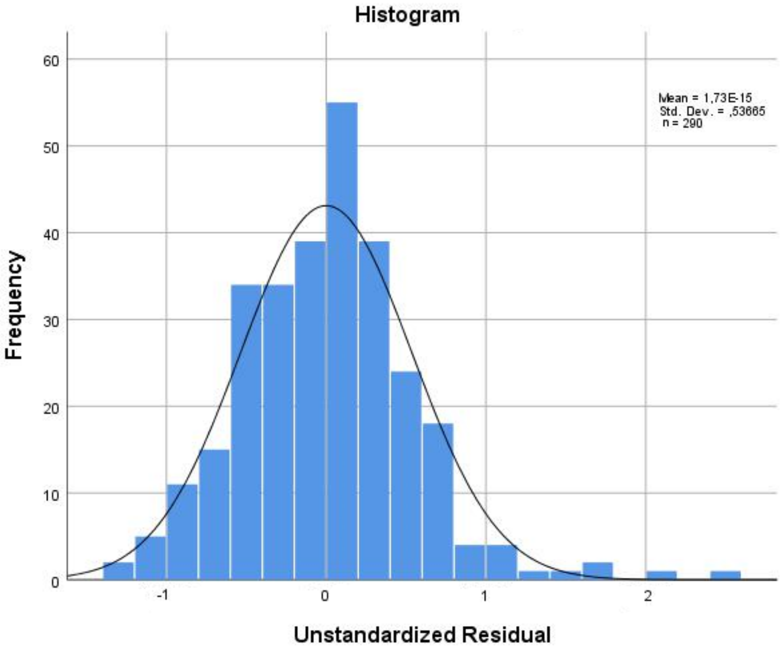

Figure 5: Bell-shaped histogram of the non-standardized residual values, from the model fit of Table 4, and a normal density curve. The figure indicates that the error terms are almost normal with $SD \approx 0,54$ and mean $\approx 0,00$

Figure 6 shows the expected values of the normal distribution plotted against the ordered unstandardized residuals, with a reference line for normality plotted. The figure shows that, at the end of the line of the plot, the observations Umeå, Uppsala, Linköping, and Kalmar were far from the normal reference line, thus revealing non-normality of the residuals.

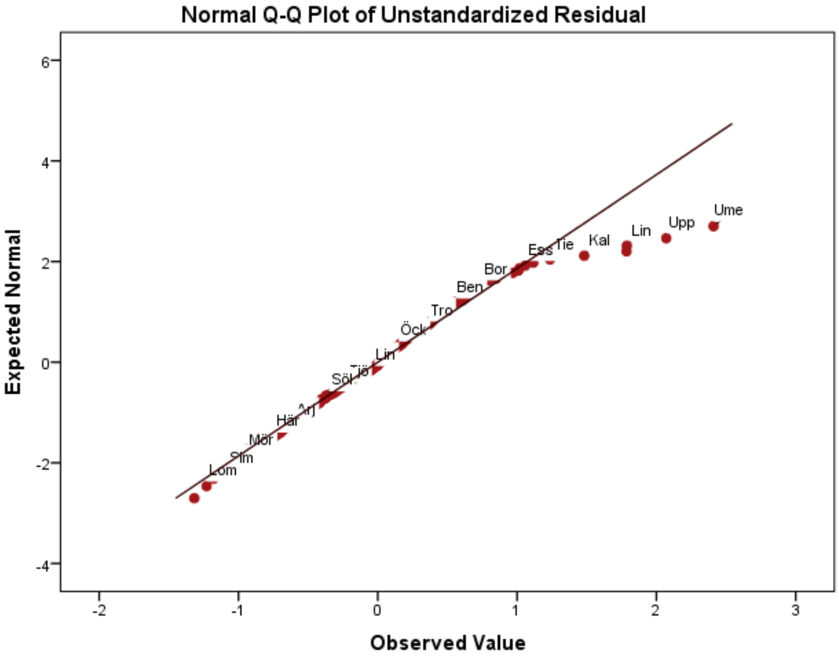

Figure 6: The normal Q-Q plot for unstandardized residuals: The observations should mostly be placed along a straight line to be normal. Probabilistic outliers at the end of the plot are distant from the the straight line.

4.6 Transformation

In this section, we re-examined the assumption of violations discussed in the previous section. We realized that the effect of the 'Unemployed' variable was non-significant in the full multiple regression model. Moreover, in Section 4.3, some observations showed more leverage than others. Additionally, the homoscedasticity assumption in Section 4.4 and the normality assumption of the error terms in Section 4.5 were violated. Thus, to modify the model, we used the log-linear transformation and transformed the dependent variable by a logarithm and re-developed a new model.

Table 14 shows the new multiple regression model of the transformed dependent variable against independent variables with a non-significant result for the independent variable 'Income $<$ 60%' with a P-value of $0,529$.

Table 14: The multiple regression coefficient summary of ’log.Women.Loan’, based on a logarithmic transformation of the dependent variable. For the standardized coefficients the independent variables have been rescaled to have a variance of 1.

Coefficient Summary						
				standardized		
Method:Enter		Unstandardised	Coefficients	Coefficients		
Model		β	Std.Error	β	t	P-value
1	Intercept	-0,105	0,039		-2,724	0,007
	Unemployed	0,008	0,002	0,163	3,188	0,002
	Educated	0,024	0,001	1,105	17,946	0,000
	Income < 60%	0,001	0,002	0,038	0,630	0,529
	Income > 200%	-0,015	0,002	-0,459	-6,591	0,000

Table 14 can be summarized in terms of the following regression formula:

$$\log \hat{Y}_i = -0,105 + 0,008(\text{Unemployed})_i + 0,024(\text{Educated})_i + 0,001(\text{Income} < 60\%)i - 0,015(\text{Income} > 200\%)_i, \quad (35)$$

where Y_i = Women.Loan$_i$ is the fraction of female borrowers in municipality i.

In the previous section, we had three assumptions which were violated, now we re-examine them with the new model in order to see if their violation is fixed or not in the new model.

4.6.1 Transformation step 1: There are no outliers in the data

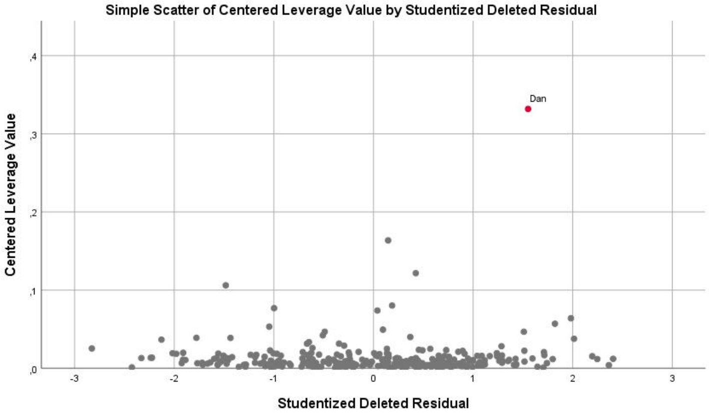

Figure 7: Transformation: The scatterplot of leverage values against standardized deleted residual values with the transformed model of ’log.Women.Loan’. Gray points represent municipalities. The most considerable amount of leverage is observed for the Danderyd municipality.

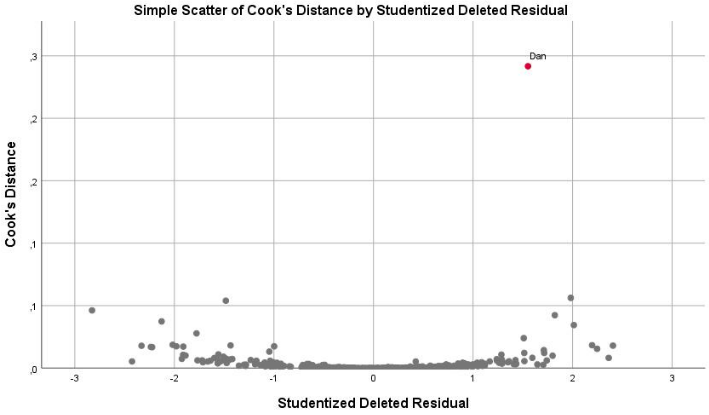

Figure 8: Transformation: The scatter plot of Cook's distance values against the standardized deleted residuals for the transformed model of 'log.Women.Loan'. Gray points represent municipalities. The largest value of the Cook's distance is observed for the Danderyd municipality.

By fitting the log model, it was found that the *Danderyd* observation was an outlier (see Figures 7 and 8).The *Danderyd* municipality had the highest income ('Income > 200%'), and the women receiving a loan there were relatively small in number, so we excluded this observation as an outlier from the data.

4.6.2 Transformation step 2: Error terms have a constant variance

In order to check if the problem of violations of the assumption of homoscedasticity is solved, with the new model, we perform the Breusch-Pagan and Koenker tests again.

Table 15: Breusch-Pagan and Koenker tests, for testing whether the error terms have a constant constant variance (homoscedasticity).

Breusch-Pagan and Koenker test statistics and P-values		
	LM	P-value
BP	4,685	0,321
Koenker	5,820	0,213
If P-value is less than 5%, the homoscedasticity assumption is rejected.		

As a result in Table 15, we have a Breusch-Pagan P-value of $,321 > 5\%$ and Koenker's score P-value of $,213 > 5\%$. Because both of the tests are significant at the 5% level, the error terms are assumed to have a constant variance and the homoscedasticity model is chosen. Figure 15, on page 46, shows a scatterplot of this model.

4.6.3 Transformation step 3: Residuals follow a normal distribution

The last violated assumption was the normality of the error terms, which is closely related to normality of the residuals. Hence we re-examined whether the normality problem of the residuals was solved after using the new model.

Table 16: Transformation: Test of the normality of the residuals with the Kolmogorov-Smirnov and Shapiro-Wilk tests for the fitted *log.Women.Loan* model

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	P-value	Statistic	df	P-value
Unstandardised Residual	0,059	289	0,017	0,990	289	0,056

The results in Table 16 demonstrate that the Kolmogorov-Smirnov test with a P-value of $0.02 < 5\%$ and the Shapiro-Wilk test with a P-value of $0.06 > 5\%$ lead to different conclusions at significance level 5%, although none of the two tests reject normality of the residuals at significance level 1%. These results reveal that the obtained model was desirable concerning goodness of fit. As observed in Figure 9, the standardized residuals approximately follow a normal distribution.

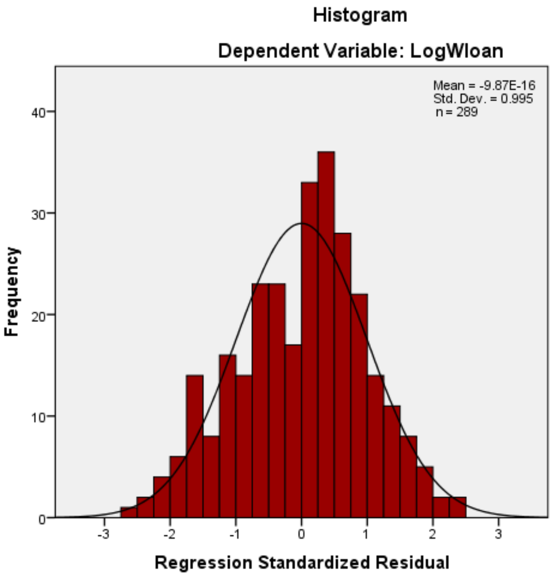

Figure 9: The histogram demonstrates approximate normality of the standardized residuals. The data almost follows a normal distribution with $SD \approx 1$ and mean ≈ 0.

5 Final result and Discussion

5.1 The Best-Fitted Model

In section 4, we examined all five assumptions investigated in this thesis in order to get the best-fitted model. Some of the assumptions were violated, but we modified them in Section 4.6. As a result, we have a linear regression model, the parameter estimates of which result in the best-fitted model.

The regression function is linear, and there is no collinearity between the independent variables. The error terms have an approximately constant variance, and the residuals follow a normal distribution; one of the observations is deleted because it is an outlier, and in the new regression model we use 289 observations instead of 290.

Now, we can, with better confidence, search for the best-fitted submodel with the logarithmic dependent variable, all independent variables, and 289 observations.

5.1.1 Multiple Linear Regression

We use the stepwise regression model in order to find the best-fitted regression model. The regression coefficients in Table 17 shows that among the four measured independent variables, three variables (namely 'Educated', 'Income $>$ 200%' and 'Unemployed') were associated with the logarithm of the percentage of female borrowers.

Table 17: The regression model found with the stepwise method, with the independent variable Income $<$ 60% non-significant.

Coefficients[a]								
				Standardized				
Method:Stepwise		Unstandardized	Coefficients	Coefficients				
Model		β	Std.Error	β	t	P-value	Tolerance	VIF
Step: 1	Intercept	0,063	0,020		3,184	0,002		
	Educated	0,015	0,001	0,666	15,132	0,000	1.000	1.000
Step: 2	Intercept	-0,011	0,019		-0,586	0,558		
	Educated	0,025	0,001	1,124	17,928	0,000	0,380	2,631
	Income $>$ 200%	-0,022	0,002	-0,582	-9,279	0,000	0,380	2,631
Step: 3	Intercept	-0,083	0,027		-3,037	0,003		
	Educated	0,024	0,001	1,103	17,873	0,000	0,377	2,655
	Income $>$ 200%	-0,18	0,003	-0,477	-7,024	0,000	0,311	3,218
	Unemployed	0,008	0,002	0,163	3,610	0,000	0,703	1,423

a. Dependent Variable: log.Women.Loan

Table 17 displays those nested submodels whose regression coefficients are significant (P-value $<$ 0.05 for the last included independent variable). This indicates that the third and largest submodel is able the predict the dependent variable the best. The regression coefficients listed in Table 17, shows that of the four independent variables measured, the three variables

'Educated', 'Income > 200%', and 'Unemployed' of the third submodel are included in order to predict the percentage of borrowers.

The fitted model must have a high R^2 in order to be able to explain the dependent variable well. The closer the coefficient of explanation is to one, the more the quality of the fit of the model increases, and the better the prediction of the dependent variable behaves as a function of the independent variables.

Table 18 demonstrated that in each step, the independent variables entered into the model by their level of importance, and the adjusted R^2 for the best fitted model is 0.587. This indicates that the model fits well with 58.7% of the changes in the logarithm of the percentage of female borrowers, explained by the three independent variables that entered into the model.

Table 18: Summary, for the models of Table 17, in terms of the coefficient of determination.

Model Summary[d]				
Model	R	R^2	R^2_{adj}	Std. Error of the Estimate
Step: 1	$0,666^a$	0,444	0,442	0,09837
Step: 2	$0,757^b$	0,572	0,569	0,08639
Step: 3	$0,769^c$	0,591	0,587	0,08463

a. Predictors: (Intercept), Educated
b. Predictors: (Intercept), Educated, Income > 200%
c. Predictors: (Intercept), Educated, Income > 200%, Unemployed
d. Dependent Variable: log.Women.Loan

The judgments about the strength and role of each of the three variables in explaining the dependent variable can be obtained from the standardized regression coefficients (β). Because these values are standardized, they allow for comparison in order to find the relative contribution of each independent variable.

According to the obtained standardized coefficients, the 'Educated' variable with the estimated beta coefficient of $\hat{\beta}_j = 1.103$ had the most strong direct relationship to the logarithm of the percentage of borrowers in each municipality, which means that this variable can be considered as the most informative independent variable regarding the percentage of borrowers. Thus, while keeping other independent variables constant, a higher percentage of the 'Educated' variable in each municipality leads to an increased percentage of borrowers.

Following this variable, the 'Income > 200%' variable with a standardized estimated regression coefficient $\hat{\beta}_j = -0.477$ had the second highest association with the percentage of borrowers. This indicates that, while keeping the rest of the independent variables constant, if the percentage of wealthy individuals in each municipality is enhanced the percentage of borrowers will be reduced.

The third most significant variable in the regression model was 'Unemployed' with $\hat{\beta}_j = 0.163$. If the other variables are kept constant, the percentage of borrowers is expected to increase as the 'Unemployed' value increases in each municipality.

The regression equation, which can be used for estimating the percentage of borrowers, in each municipality is as follows:

$$\text{log.}\widehat{\text{Women}}\text{.Loan}_i = -0,083 + 0,024(\text{Educated})_i - 0,018(\text{Income} > 200\%)_i + 0,008(\text{Unemployed})_i \tag{36}$$

We can use it to predict the percentage of borrowers in each municipality.

5.1.2 Prediction of loan recipients in 2016

In order to evaluate the regression model, the percentage of Women.Loan recipients from each municipality in 2016 is predicted using the regression model in equation (36).

By entering the values of the independent variables, including the percentage of 'Educated', the percentage of 'Income> 200%', and the percentage of 'Unemployed' in each municipality in the formula, we can predict the log-percent of borrowers in 2016. Then we exponentiate and convert these numbers into percentage of borrowers.

The plot of the actual values against the predicted values is demonstrated in Figure 10. The coefficient of determination values in this model are $R^2 = 0.737$ and $R^2_{\text{adj}} = 0.736$, which is indicative of a relatively acceptable predictive ability of the model.

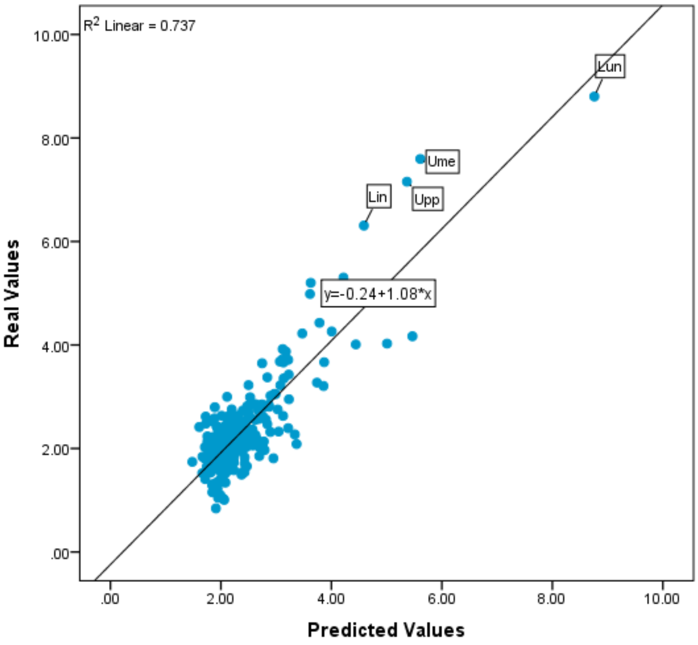

Figure 10: The distribution chart of the observed values of the percentage of women borrowers in 2016 for 289 municipalities against the corresponding predicted values in per cent. The continuous line represents the fitted line of the regression, and blue spots are the municipalities.

Table 21 on page 43 shows the values of the independent variables, the actual values of the percentage of female borrowers, and the predicted values for ten of the municipalities.

6 Conclusion

Given the importance of estimating the percentage of borrowers in each municipality for better planning for the coming years, the purpose of this study was to estimate the percentage of female borrowers using the multiple linear regression model. Also, the relative importance of different variables affecting the percentage of female borrowers was studied.

For this purpose, we used data from the year 2015 in each Swedish municipality, including information of four independent variables that correspond to the fraction of women that are educated ('Educated'), are unemployed ('Unemployed'), have a low income ('Income < 60%) and have a high income ('Income > 200%') respectively. The correlation coefficients were used to examine the pairwise relationship between the variables. The three variables of Education, Income > 200%, and Income < 60% had a significant relationship with the percentage of female borrowers.

The distribution of the dependent variable against each of the independent variables was investigated using scatterplots. In the first step, multiple linear regression with all independent variables was fitted. The results indicated that the 'Unemployed' variable was non-significant in the model. Investigation of the assumptions of the regression model also showed that some of the municipalities were outliers and more influential than others.

The variance of the residuals was higher for values of the predicted percentage of female borrowers above five per cent than for the values less than five. The normality assumption of the residuals was also violated. Due to the problems mentioned earlier, the independent variables were fitted instead to the logarithm of the dependent variable. By fitting this model, it was found that the 'Income < 60%' variable was non-significant, and moreover we found that the observation Danderyd was found to be an outlier that was removed from data.

The new model was fitted using stepwise regression. According to the results of this procedure, the most important variable in the regression model was 'Education', revealing that a higher number of educated individuals in each municipality gives a higher number of borrowers, which is not surprising.

The second most significant independent variable, after the 'Education' variable, was the 'Income > 200%' variable, having a negative correlation with the percentage of borrowers. Because the higher the percentage of wealthy individuals in each municipality is, a larger fraction of these individuals naturally do not need loans. This observation reinforces that the result is reasonable and correct.

Finally, in order to confirm the fitted model, we used data from 2016 in order to predict the percentage of female borrowers in each municipality. The results showed that the model was able to predict the percentage of borrowers in a satisfactory way.

7 Appendix

Table 19: The table demonstrates the five highest and lowest observations (municipalities) in each variable from the data of 2015.

	Dependent Variable:	Min	%	Max	%
Y_i	Women.Loan	Haparanda	1,19	Lund	8,90
		Askersund	1,20	Umeå	7,76
		Årjäng	1,28	Uppsala	7,32
		Kiruna	1,28	Linköping	6,51
		Filipstad	1,34	Kalmar	5,34
	Independent Variables:	Min	%	Max	%
X_{1i}	Unemployed	Danderyd	2,30	Landskrona	15,20
		Knivsta	2,50	Södertälje	15,20
		Vaxholm	2,50	Malmö	15,00
		Öckerö	2,60	Lessebo	14,30
		Vallentuna	2,60	Eskilstuna	13,90
X_{2i}	Educated	Munkfors	11,73	Lund	46,79
		Filipstad	11,80	Danderyd	41,78
		Dorotea	11,81	Solna	41,73
		Eda	12,14	Stockholm	39,03
		Årjäng	12,27	Uppsala	38,15
X_{3i}	Income < 60% of Middle Income	Täby	5,50	Årjäng	29,00
		Nykvarn	5,60	Eda	27,90
		Öckerö	5,70	Dals-Ed	24,70
		Lomma	5,70	Haparanda	23,80
		Ekerö	5,80	Malmö	23,10
X_{4i}	Income > 200% of Middle Income	Munkfors	1,20	Danderyd	36,60
		Dorotea	1,40	Lidingö	26,00
		Vilhelmina	1,40	Täby	22,90
		Hällefors	1,60	Nacka	19,70
		Norsjö	1,80	Vaxholm	18,80

Table 20: The table demonstrates the five highest and lowest observations (municipalities) for the *Middle Income* 2015 variable

Variable	Min	SEK/Year	Max	SEK/Year
Middle Income	Eda	181,05K	Danderyd	660,75K
	Årjäng	181,60K	Lidingö	454,45K
	Åsele	188,45K	Täby	367,30K
	Haparanda	190,40K	Nacka	354,50K
	Vilhelmina	190,40K	Vaxholm	349,70K

Table 21: The summary of the data for 2016, used to predict the percentage of female borrowers in each municipality this year.

Case number	Municipality	Unemployed	Education	Income > 200%	Predict	Inverse-Log	Residual
1	Uppsala	6.1	24.08	8.3	4	2.53	0.32
2	Vallentuna	2.5	26.11	11.4	3.7	2.35	-0.20
3	Österåker	3.4	25.42	13.3	3.3	2.12	0.28
4	Värmdö	3.5	25.29	13.6	3.2	2.08	0.14
5	Järfälla	8.3	27.15	9.1	4.8	3.03	-0.27
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
286	Luleå	6.5	30.39	6	6	4.01	0.26
287	Piteå	6.1	21.60	4.4	4.1	2.59	0.25
288	Boden	8.8	23.34	4	4.8	3.04	-0.72
289	Haparanda	13.6	13.30	1.9	3.1	2.05	-1.03
290	Kiruna	4.4	18.33	4.5	3.2	2.08	-0.74

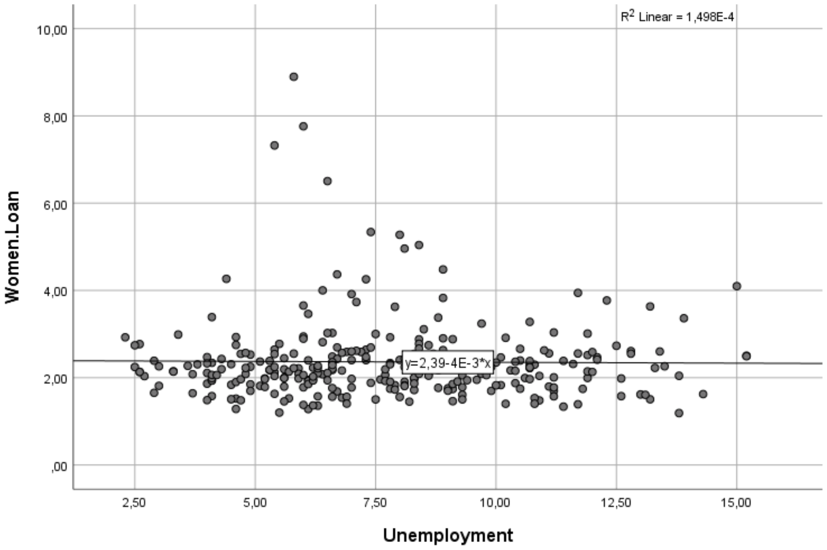

Figure 11: A scatterplot of the dependent variable 'Women.Loan' against the 'Unemployed' independent variable. The points are randomly scattered, and no significant relationship is observed with $R^2 = 1.498\text{E} - 4$.

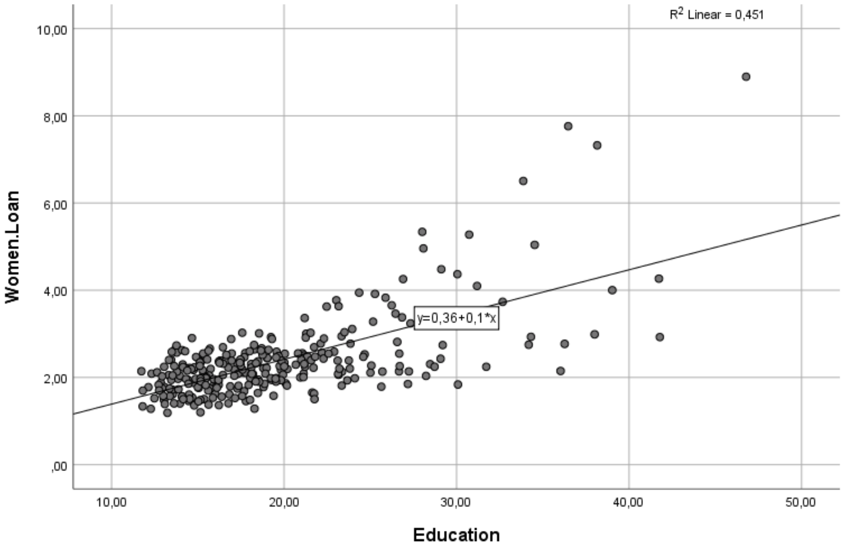

Figure 12: A scatterplot of the dependent variable 'Women.Loan' against the 'Educated' independent variable. A positive linear relationship is observed with $R^2 = 0.451$.

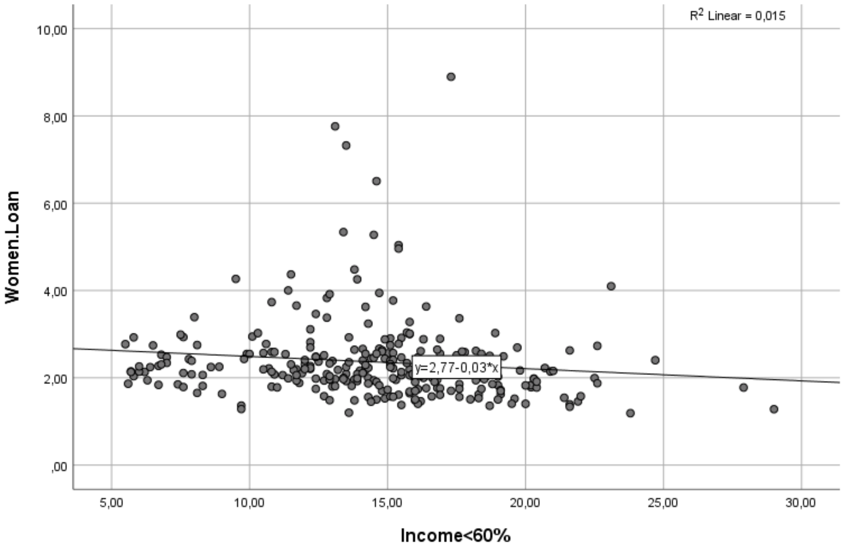

Figure 13: A scatterplot of the dependent variable 'Women.Loan' against the 'Income $< 60\%$' independent variable. A negative linear relationship is observed with $R^2 = 0.015$.

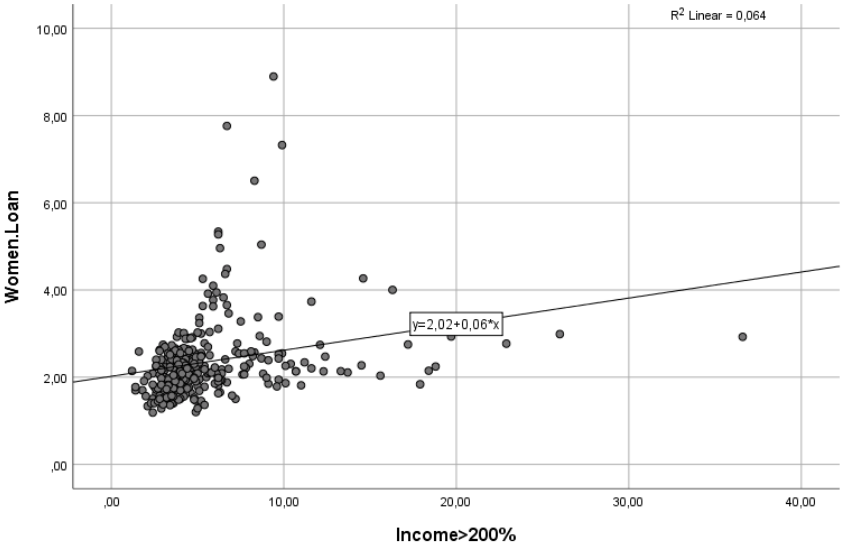

Figure 14: A scatterplot of the dependent variable 'Women.Loan' against the 'Income $> 200\%$' independent variable. A positive linear relationship is observed with $R^2 = 0.064$.

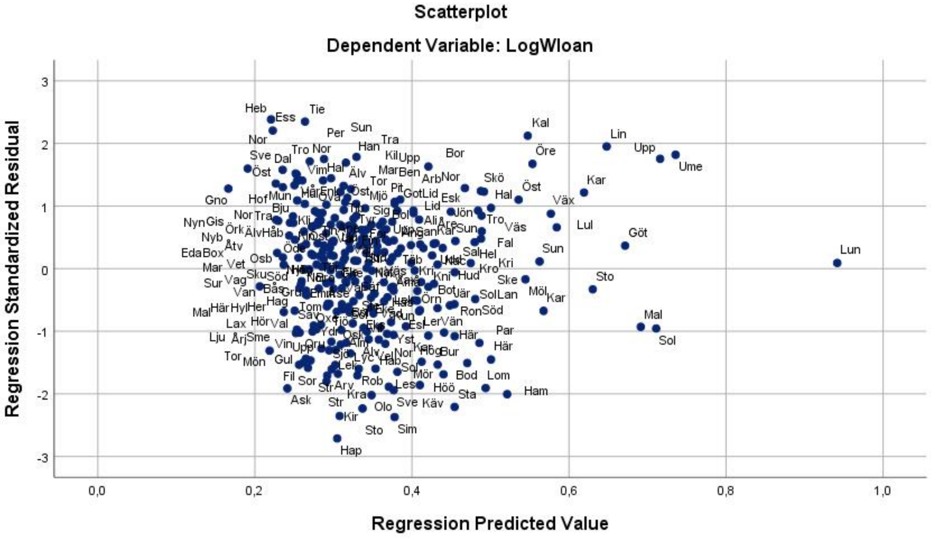

Figure 15: Homoscedasticity Scatterplot of predicted values against standardized residuals for the new model with a log-transformed dependent variable

References

[1] Sundberg, Rolf. (2016). *Linjära Statistiska Modeller*, Stockholm University

[2] Andersson, Patrick and Tyrcha, Joanna. (2016). *Notes in Econometric*, Stockholm University

[3] Frost, Jim, Ms. (2019). *Regression Analysis*, E-Book.

[4] Centrala studiestödsnämnden, www.csn.se

[5] Statistiska Centralbyrån, www.scb.se

[6] Arbetsförmedlingen, www.arbetsformedlingen.se

[7] Samuelsson, Daniel (2018.12.17) *Students are not unemployed*, www.scb.se

[8] Simon, Laura and Young, Derek. *STAT 462: Regression Methods*, The Pennsylvania State University
Online Course: [https://newonlinecourses.science.psu.edu/stat462/node/247]

[9] Simon, Laura and Young, Derek. *STAT 462: Regression Methods*, The Pennsylvania State University
Online Course: [https://newonlinecourses.science.psu.edu/stat462/node/172/]

[10] Simon, Laura and Young, Derek. *STAT 501:Regression Methods*, The Pennsylvania State University
Online Course: [https://onlinecourses.science.psu.edu/stat501/lesson/11/11.3]

[11] Alm, Sven Erick and Britton, Tom. (2011) *Stokastik sannolikhetsteori och statistikteori med tillämpningar*, Liber

[12] *T-statistic. Wikipedia article.* [https://en.wikipedia.org/wiki/T-statistic]

[13] Simon, Laura and Young, Derek.*STAT 462: Regression Methods STAT 501:Regression Methods*, The Pennsylvania State University,
Online Course: [https://online.stat.psu.edu/stat462/node/131/]

[14] Sweeten, Gary. *Advanced Statistical Analysis*, Arizona State University
Breusch-Pagan, Koenker-Score test,
E-resource: [http://www.public.asu.edu/~gasweete/crj604/slides/Lecture%208.pdf]

[15] Field, Andy. (2009). *Discovering Statistics Using SPSS*. E-Book.

[16] Dunn, Kevin. *Process Improvement Using Data Release 04388a*, E-resourse: [https://learnche.org/pid/least-squares-modelling/index]

[17] Bruin, J. (2006).*Introduction to Regression with SPSS*. UCLA: Statistical Consulting Group
Online Course: [https://stats.idre.ucla.edu/spss/seminars/introduction-to-regression-with-spss/introreg-lesson2/]

[18] Simon, Laura and Young, Derek. *STAT 462:Regression Methods*, The Pennsylvania State University
Online Course [https://online.stat.psu.edu/stat462/node/173/]