**BSc report APPLIED MATHEMATICS**

# "Principal Component Analysis of Education-Related Data Sets"
# (Dutch title: "Principale-componentenanalyse van educatie gerelateerde datasets")


Thao Nguyen
4704924

**Delft University of Technology**


**Supervisors**

C. Vuik, E.D. Wobbes, E. Fleur

**Other committee member**

K.P. Hart

July, 2020 Delft

# Abstract

Principal Component Analysis (PCA) is a mathematical instrument beneficial for its dimension reduction whilst keeping the most important data. Due to its advantages, PCA is chosen to handle a substantial amount of data. In this thesis two questions are answered: what variables influence a pupil's attainment test score using linear regression and whether PCA provides better linear regression models? The data used in this thesis is provided by DUO, the Dutch Executive Agency for Education. The data contains information about pupils who completed the attainment test in 2008-2013. This thesis starts with a brief description of the data set used for the research and some background information about PCA. Before linear regression can be used, the data is preprocessed. Creating a linear model with all variables resulted in the largest absolute coefficients for teachers' secondary school recommendations. When PCA is applied, it gives great insight into which variables are (likely) dependent on each other: dependent not only in the sense of linear dependency but also the influences on each other in general. Furthermore, PCA also indicates which variables are most likely to have a significant impact. When the data set is free of linearly dependent variables, PCA may give worse fitted models. However, the models are better than models with randomly chosen variables.

# Acknowledgements

This thesis is the result of the effort and support of several people to whom I am extremely thankful. First and foremost, I would like to express my appreciation to my supervisors: Lisa Wobbes, Kees Vuik and Erik Fleur. It has been a privilege to work with you all! Thank you for guiding me during this journey and making time for me in your busy schedules.

I would also like to thank the Data Science department of DUO (the Dutch Executive Agency for Education). They have welcomed me with open arms during my internship. While I was not able to meet them in this extra-ordinary times of the COVID-19 pandemic, I will look back at my internship as a fond memory. Special thanks to Marc Meurs and Bo de Lange for their help and helpful advice.

Furthermore, I would like to thank Klaas Pieter Hart for making time to be the last member of my committee besides my supervisors, as well as providing feedback for the improvement of my thesis and spotting sloppy, overlooked mistakes.

Next, I would like to recognize the guidance that I received from my tutor Karen Aardal. During my time as a bachelor student, her door was always open for me whenever I needed some advice or help.

Finally, I would also like to extend my gratitude to my parents and to my boyfriend and friends for providing me with unfailing support and continuous encouragement.

Thanks a lot!

*Thao Nguyen*
*Delft, 2020*

# Contents

# Mathematical notation and list of symbols/abbreviations

## Mathematical notation

In this thesis most notation is similar to the article Principal component analysis by H. Abdi and L. Williams (2010) [1]. To make a distinction between matrices, vectors and elements matrices are denoted with upper case bold letters, vectors with lower case bold and elements with lower case italic. If elements, vectors and matrices are from the same matrix, the same letter is used (for example $\mathbf{A}$, $\mathbf{a}$, $a$). Elements will usually get 2 indices as subscripts, for example $a_{ij}$. The first index $i$ will stand for the matrix row the element is in and the second index $j$ for the column. When the norm $||\cdot||$ is not specified, the norm is the Euclidean norm. So if $\mathbf{x}$ is a vector of length I, then $||\mathbf{x}|| = \sqrt{\sum_{j=1}^{I} x_j^2}$.

## List of symbols and abbreviations

| | |
|---|---|
| DUO | 'Dienst Uitvoering Onderwijs', Dutch Executive Agency for Education |
| MAE | Mean Absolute Error |
| MSE | Mean Squared Error |
| OLS | Ordinary Least Squares |
| PCA | Principal Component Analysis |
| RMSE | Root Mean Squared Error |
| SSE | Sum of Squared Errors (of prediction) |
| SVD | Singular Value Decomposition |
| WPO | 'Wet op het Primair Onderwijs', Dutch Primary Education Act |

# 1 Introduction

"Educations is the foundation upon which we build our future."-Christine Gregoire

These days a (modern) society without an education system is unthinkable. Without education a society cannot flourish and grow, perhaps not even maintain itself. There is hence no doubt how important the work of the Dutch Executive Agency for Education (DUO) is.

Nowadays, a lot of data can easily be gathered. DUO also possesses large amounts of education-related data. To make an accurate interpretation of such data, techniques must be used that reduce the dimensionality whilst still keeping the main parts of the information.

In this thesis the technique Principal Component Analysis (PCA) is used to do so. PCA is a mathematical instrument, mainly based on the linear algebraic concept Singular Value Decomposition. By finding the directions in the data with the greatest variation and projecting the data onto these directions, the dimensionality can be reduced and important data remains.

Using PCA we wish to evaluate data from the elementary education in the Netherlands and eventually find the variables that influence an elementary student's performance the most. Since 2014 Dutch primary students are obligated to take a leavers attainment test in their last year of primary school. Before that, the tests were optional but common. Based on the pupil's test score and teacher's recommendation they start secondary education in the most suitable 'stream' (VMBO, HAVO or VWO).

There are different leavers attainment tests primary school can let their students take. One of them is the 'Centrale Eindtoets'. As this test is most popular and made available by the Dutch government, the Centrale Eindtoets has been chosen as the measure for pupil's performance for this research.

To see how strongly variables relate to the test score, linear models are proposed. Based on linear regression models the performance of PCA is evaluated by comparing models with and without the use of PCA. However, like most data sets it has to undergo some changes beforehand: errors, missing values and categorical variables should be addressed. Hence, the data has to be preprocessed before regression analysis and PCA can be applied.

This thesis starts with a description of the data. Then PCA is introduced in chapter 3. As noticed previously, the research can not start without preprocessing the data. This is worked out in chapter 4. Next, linear regression is introduced and linear models are made without PCA (chapter 5) and with PCA (chapter 6). In chapter 7 models with and without PCA are compared. Finally, by using hypothesis testing and p-values the variables are examined if they significantly influence the attainment test score.

# 2 Data set

In this thesis data from the Dutch (special) primary education is used. The data is described by 81 variables and includes a total number of 744771 pupils, whose attainment test (Centrale Eindtoets) scores are available, from the years 2008-2013. In this chapter, a description of the available variables is given. Each variable name is similar to the ones of DUO. Since some variables are not of interest, a selection has been made out of the 81 variables. As the given data set was not free of duplicates and errors in the attainment scores (some values were outside the domain), these corresponding observations were filtered out first. Moreover, for pupils who have done the attainment test twice in successively years only their last attainment test and personal information are taken into account. Furthermore, before this description the data has been partly altered (see section 4.2). Hence, this description is different from a description of the original DUO data.

1. **CFINR:**[1] unique number of length 9 assigned to each pupil
   - Variable set is complete
   - Variable type = nominal variable
2. **GESLACHT:** gender of the pupil (m/v/o)
   - Variable set is complete
   - Domain = {M, V, O}
   - M = male, V = female, O = unknown
   - Percentages: M = 49.8%, V = 50.2%, O = 0.0003%
   - Variable type = nominal variable
3. **LAND_GEB:** country of birth, represented by a 4 digit code
   - Variable set is complete
   - Domain = 179 different countries
   - Most frequent: the Netherlands (96.3%)
   - Second and third most frequent: China (0.3%), Germany (0.2%)
   - Variable type = nominal variable
4. **LAND_OUDER1:** country of birth of one of the parents, represented by a 4 digit code
   - Variable set is complete
   - Domain = 236 different countries
   - Most frequent: the Netherlands (82.3%)
   - Second and third most frequent: Morocco (3.4%), Turkey (3.0%)
   - Variable type = nominal variable
5. **LAND_OUDER2:** country of birth of the other parent, represented by a 4 digit code
   - Variable set for 99,22% complete
   - Domain = 230 different countries
   - Most frequent: the Netherlands (81.8%)
   - Second and third most frequent: Morocco (3.6%), Turkey (3.3%)
   - Variable type = nominal variable
6. **DAT_VEST_NL:** date of settlement in the Netherlands
   - Applicable to 5.17% of the pupils
   - Domain = 4351 different dates
   - Variable type = ordinal variable

[1] **CFINR** is used as an index rather than a variable.

7. **NATIO1:** first nationality of pupil
   - Variable set is complete
   - Domain = 137 different nationalities
   - Most frequent: Dutch (97.6%)
   - Second and third most frequent: Unknown (0.67%), Turkish (0.44%)
   - Variable type = nominal variable
8. **NATIO2:** second nationality of pupil
   - Applicable to 11.9% of the pupils
   - Domain = 175 different second nationalities
   - Most frequent: Moroccan (27.2%)
   - Second and third most frequent: Turkish (25.6%), German (6.0%)
   - Variable type = nominal variable
9. **PC4_LEERL:** numerical part of Dutch zip code of pupil
   - Variable set for 99.3% complete
   - Domain = 3930 different zip codes
   - Most frequent: 5045 (Tilburg)
   - Variable type = nominal variable
10. **BRIN:** special number representing the school of the pupil
    - Variable set is complete
    - Domain = 6210 different schools
    - Variable type = nominal variable
11. **ACHTERGR:** indication (non-)Dutch cultural background (or not applicable)
    - Variable set is complete
    - Domain = {0, 1, 2}
    - 0 = inapplicable, 1 = Dutch cultural background, 2 = non-Dutch cultural blackground
    - Percentages: 0 = 1.5%, 1 = 87.7%, 2 = 10.8%
    - Variable type = nominal variable
12. **VOORS_MND:** total number of months a pupil attended preschool education (age 2-3)
    - Applicable to 0.45% of the pupils
    - Most frequent: 18 months (49.0%)
    - Variable type = integer variable
13. **ADVIES_VO:**[2] recommendation of teacher for secondary education level
    - Variable set for 77.0% complete
    - Domain = 34 numerical codes corresponding to secondary education levels
    - Most frequent = VWO (17.8%)
    - Second and third most frequent: HAVO (17.3%), VMBO TL (15.6%)
    - Variable type = ordinal variable
14. **GEWICHT:**[3] weight intended to allocate extra funding to primary schools for the funding of students who need extra attention considering the educational level of the parent(s)
    - Variable set is complete
    - Most frequent: at least one parent followed two years of general secondary education or higher (86.4%)
    - Variable type = ordinal variable

[2]Before 2014 the recommendation was given after the attainment test. However, the teachers should not consider the attainment test score when giving their recommendation and it may therefore be taken into account in this research.

[3]Not always filled in correctly by schools.

15. **DAT_TOETS:** date of attainment test
    - Variable set is complete
    - Most frequent: 3 February 2011 (9.8%)
    - Variable type = ordinal variable

16. **UITSLAG:** result of attainment test
    - Variable set is complete
    - Domain = {501, 502, ..., 550}
    - Most frequent: 538 (4.2%)
    - Second and third most frequent: 540 (4.1%), 544 (4.1%)
    - Variable type = interval-scaled variable

17. **LEERJAAR:** grade of pupil when doing the attainment test
    - Variable set is complete
    - Domain = {0, 1, ..., 8}
    - Most frequent: 8 (99.6%)
    - Second most frequent: 7 (0.3%)
    - Variable type = interval-scaled variable

18. **VBJ_BO:**[4] number of years of residence at (regular) primary education
    - Variable set is complete
    - Domain = {0, 1, ..., 10}
    - Most frequent: 8 (65.3%)
    - Second and third most frequent: 9 (11.8%), 7 (8.8%)
    - Variable type = integer variable

19. **VBJ_SBO:**[4] number of years of residence at special primary education
    - Variable set is complete
    - Domain = {0, 1, ..., 9}
    - Most frequnt: 0 (99.787%)
    - Variable type = integer variable

20. **VBJ_SO:**[4] number of years of residence at special education
    - Variable set is complete
    - Domain = {0, 1, ..., 9}
    - Most frequent: 0 (99.895%)
    - Variable type = integer variable

21. **VBJ_VSO:**[4] number of years of residence at reformed secondary education
    - Variable set is complete
    - Domain = {0, 1, 2, 3, 5}
    - Most frequent: 0 (99.998%)
    - Variable type = integer variable

22. **VBJ_INS:** number of years of residence at the current institution
    - Variable set is complete
    - Domain = {1, 2, ..., 10}
    - Most frequent: 8 (61.0%)
    - Second and third most frequent: 9 (10.7%), 7 (7.9%)
    - Variable type = integer variable

[4] Variables about previous schools and school years are not reliable, because information before 2008 was not always available.

23. **LJR_INS_1E:** grade of pupil in its first year at the institution
    - Variable set is complete
    - Domain = {1, 2, ..., 8}
    - Most frequent: 1 (72.3%)
    - Second and third most frequent: 2 (7.6%), 3 (3.8%)
    - Variable type = interval-scaled variable

24. **TYPE_PO:**
    - Variable set is complete
    - Domain = {BO, SBO}
    - (S)BO = (special) primary education
    - Percentages: BO = 99.8%, SBO = 0.2%
    - Variable type = nominal variable

25. **GROEPSGR:**[5] group size of the pupil's class
    - Variable set is complete
    - Most frequent: 26 (7.3%)
    - Second and third most frequent: 25 (7.2%), 24 (6.8%)
    - Variable type = integer variable

26. **VROEG_MND:** total number of months a pupil attended preschool education (age 4-6)
    - Applicable to 0.2% of the pupils
    - Most frequent: 23 (4.9%)
    - Second and third most frequent: 22 (3.7%), 85 (questionable, 3.6%)
    - Variable type = integer variable
    - Usability of this field is limited (e.g. end date preschool education is not always available and is set as attainment test date or information is not filled in at all)

27. **NVS:**[4] number of different schools the pupil has attended
    - Variable set is complete
    - Most frequent: 1 (91.7%)
    - Second and third most frequent: 2 (7.7%), 3 (0.5%)
    - Variable type = integer variable

28. **DENOMINATIE:** denomination of the institution
    - Variable set is complete
    - Most frequent: Roman Catholic (36.6%)
    - Second and third most frequent: public (30.1%), Protestant Christian (24.0%)
    - Variable type = nominal variable

29. **GENERATIE:** indicator to which generation of immigrant the pupil belongs, based on country of birth of the pupils and their parents
    - Variable set is complete
    - Domain = 0, 1, 2, 3 (3 if ethnicity is unknown)
    - Percentages: 0 = 78.0%, 1 = 2.4%, 2 = 19.4%, 3 = 0.2%
    - Variable type = integer variable

[5] Not always filled in correctly, about 95% correct.

30. **GEBJAAR:** year of birth
    - Variable set is complete
    - Domain = {1994, 1995, ..., 2008}
    - Most frequent: 2000 (20.5%)
    - Second and third most frequent: 2001 (20.0%), 1999 (19.5%)
    - Variable type = interval-scaled variable

31. **AFSTAND:** distance between pupil's home address and school in meters
    - Variable set is complete
    - Most frequent: 0 (71.3%)
    - Second and third most frequent: 999 999 (0.9%), 949 (0.06%)
    - Variable type = ratio-scaled variable
    - Not accurate as only the numerical part of the zip code is used. If the distance cannot be determined, '999 999' is filled in.

As the attainment test score (**UITSLAG**) is the target variable, a histogram is made to visualize the variable:

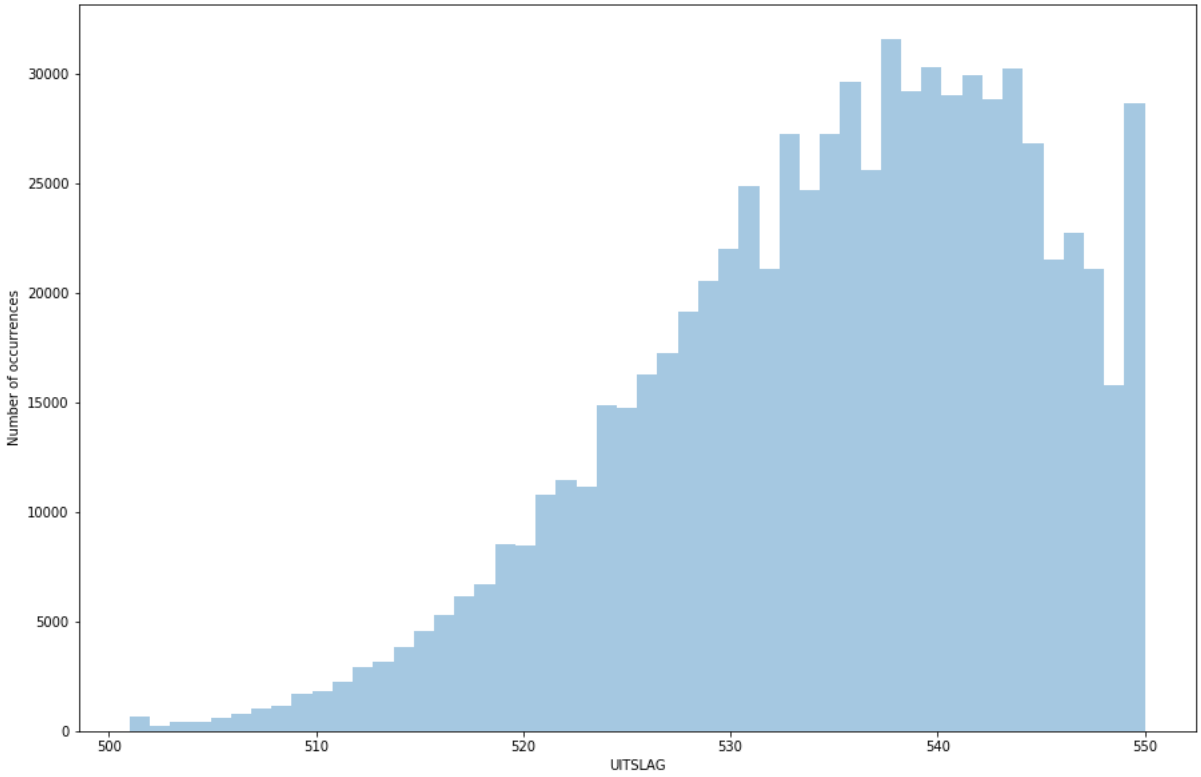



Figure 1: Number of occurences per attainment test score in the years 2008-2013.

# 3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a multivariate method. Its central idea is to reduce the dimension of the data set while preserving the main parts of the information. This is done by representing the data set with new and fewer variables, which are linear combinations of the original variables. Therefore, PCA is able to extract the most important information from the data set whilst compress the size of the data set. This section is mainly based on the article 'Principal component analysis' by H. Abdi and L. Williams (2010) [1].

## 3.1 PCA steps

In a nutshell, PCA consists of the following steps:

1. Construction of a data matrix
2. Centering and standardization of the data matrix
3. Computation of the eigenvalues and eigenvectors of the correlation matrix
4. Analysis of the found eigenvalues and eigenvectors
5. Computations of the transformed values and further analysis

In the remainder of this paragraph, each step is described in more detail.

**Step 1: Construction of a data matrix**

To get a clear overview of our data a data matrix is constructed. Each row consists of the variable values of one observation. The data matrix is denoted by $\mathbf{S}$ and will hence be an $I \times J$ matrix, where $I$ denotes the number of observations and $J$ denotes the number of variables.

$$\mathbf{S} = \begin{pmatrix} s_{11} & \dots & s_{1J} \\ \vdots & \ddots & \vdots \\ s_{I1} & \dots & s_{IJ} \end{pmatrix}$$

**Step 2: Centering and standardization of the data matrix**

It is common to preprocess the data table before the analysis. First, the columns are centered such that each column has a mean equal to 0. This is done by substracting the mean of a column from the column's elements. So the matrix has the following entries:

$$\begin{pmatrix} s_{11} - \bar{S}_1 & \dots & s_{1J} - \bar{S}_J \\ \vdots & \ddots & \vdots \\ s_{I1} - \bar{S}_1 & \dots & s_{IJ} - \bar{S}_J \end{pmatrix}$$

where $\bar{S}_k$ denotes the mean of column $k$, in other words $\bar{S}_k = \frac{1}{I} \sum_{i=1}^{I} s_{ik}$.
Moreover, dividing the matrix by $\sqrt{I}$ (or $\sqrt{I-1}$) results in a matrix denoted by $\mathbf{D}$. Doing the analysis with matrix $\mathbf{D}$ gives the so-called *covariance* PCA. Looking at the matrix $\mathbf{D^T D}$ and noting that covariance $cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ this feels intuitive as $\mathbf{D^T D}$ has the following entries:

- On the $j$'th diagonal entry for every $1 \leq j \leq J$: $\frac{1}{I}\sum_{i=1}^{I}(s_{ij} - \bar{S}_j)^2$

- On the other $jk$'th entry: $\frac{1}{I}\sum_{i=1}^{I}(s_{ij} - \bar{S}_j)(s_{ik} - \bar{S}_k)$

Matrix $D^T D$ is therefore called the covariance matrix.

Next to centering, it is usual to standardize each column to scaled unit norm. To obtain this each column $j$ of matrix $\mathbf{D}$ is divided by the scaled norm, i.e. $\sqrt{\frac{\sum_{i=1}^{I}(s_{ij}-\bar{S}_j)^2}{I}}$. Using the newly obtained matrix $\mathbf{X}$ for the analysis gives the so-called *correlation* PCA. The matrix $\mathbf{X^T X}$ namely looks as follows:

- All the diagonal entries are equal to 1

- An off-diagonal entry $jk$ is equal to: $\frac{\sum_{i=1}^{I}(s_{ij}-\bar{S}_j)(s_{ik}-\bar{S}_k)}{\sqrt{\sum_{i=1}^{I}(s_{ij}-\bar{S}_j)^2}\sqrt{\sum_{i=1}^{I}(s_{ik}-\bar{S}_k)^2}}$. This is the correlationcoefficient of columns $j$ and $k$.

To see this, one should realize that $\mathbf{X^T X}$ $jk$'th entry can be seen as the inner product between column $j$ and column $k$. Each $jk$'th entry of the covariance matrix will be divided by the scaled norm of column $j$, as well as the scaled norm of column $k$. As both scaled norms have division by $\sqrt{I}$, the $\frac{1}{I}$ of the covariance term drops. The $jk$'th term from above is hence found. For the diagonal insert the same index $i$ for both $j$ and $k$. The numerator and denominator will both be equal to $\sum_{i=1}^{I}(s_{ij} - \bar{S}_j)^2$. Thus the diagonal entry results in the value 1.
The matrix $\mathbf{X^T X}$ is therefore called the *correlation* matrix.

**Step 3: Computation of the eigenvalues and eigenvectors of the correlation matrix**

To obtain the principal components the Singular Value Decomposition (SVD) of matrix $\mathbf{X}$ must be found. Singular Value Decomposition (SVD) can be seen as the diagonalization for rectangular matrices. It decomposes a rectangular matrix $\mathbf{A}$ into three simpler matrices $\mathbf{P}, \mathbf{\Delta}, \mathbf{Q^T}$, where

- $\mathbf{P}$ represents for the (normalized) eigenvectors of matrix $\mathbf{AA^T}$. $\mathbf{P}$'s columns are also known as the *left singular vectors* of $\mathbf{A}$.

- $\mathbf{Q}$ represents for the (normalized) eigenvectors of the matrix $\mathbf{A^T A}$. $\mathbf{Q}$'s columns are also known as the *right singular vectors* of $\mathbf{A}$.

- $\mathbf{\Delta}$ represents the diagonal matrix of *singular values*. It can be found by taking the square root of the eigenvalues of matrices $\mathbf{AA^T}$ and $\mathbf{A^T A}$ (as they share the same eigenvalues).

Suppose $\mathbf{X}$'s SVD is given by $\mathbf{X = P\Delta Q^T}$. The eigenvectors in $\mathbf{Q}$ are the principal components. The matrix $\mathbf{Q}$ in this decomposition is the projection matrix that projects the original values onto the principal components. The principal components are the vectors that determine the new feature space. Note that in this case both $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal matrices. The principal components are a new orthogonal basis, computed in such a way that the variance is maximized in each basis vector in the same order as their corresponding eigenvalue size.

**Step 4: Analysis of the found eigenvalues and eigenvectors**

The ordering of the principal components is equivalent to the value order of their corresponding eigenvalue, i.e. the eigenvector of the correlation matrix corresponding to the largest eigenvalue

is the first principal component. To reduce the dimension of our data set a selection of the principal components is made. The number of components that are kept is discussed in the next subsection.

**Step 5: Computations of the transformed values and further analysis**

As matrix $\mathbf{Q}$ can be viewed as the projection matrix, the matrix with the transformed values can be found by multiplying matrix $\mathbf{X}$ with $\mathbf{Q}$. As $\mathbf{X} = \mathbf{P\Delta Q^T}$ and $\mathbf{Q^T Q} = \mathbf{I}$,

$$\mathbf{XQ} = \mathbf{P\Delta Q^T Q} = \mathbf{P\Delta}$$

This can be intuitively explained as follows: as $\mathbf{Q}$ is the matrix with the coefficients of the principal components, multiplying by $\mathbf{Q}$ gives the projection of the original matrices on the principal components. When, for example, the decision is made to keep 2 principal components, the first principal component becomes the horizontal axis. This results in a better overview of the data and in the correlation PCA case is equivalent to rotating the original data points.

## 3.2 Choosing the number of components

Not all variables provide important insight into the data set. Therefore, some components will be neglected. In this subsection, it is explained how to select the components that should be kept.

The first solution is called the *scree* or *elbow* test. The idea is to plot all eigenvalues, ordered with respect to their size. In the resulting scree-plot, a point is assigned as the 'elbow'. Before the elbow, the slope between the eigenvalues should be steep, whilst after the elbow the slope becomes flat. Only the principal components corresponding to eigenvalues before the elbow are taken into account. As different people do not necessarily assign the same point as the elbow, this procedure is not objective.

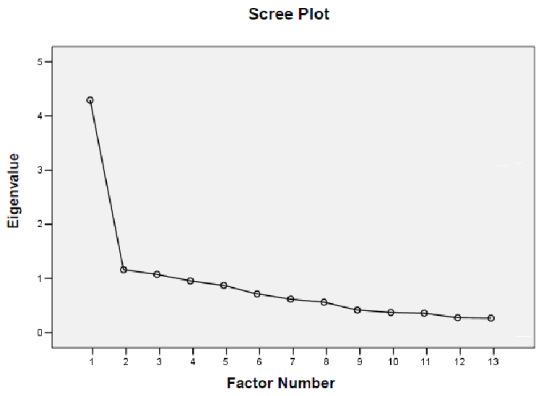



Figure 2: Example of a scree-plot with a clear elbow. Source: [8]

The second solution is to take the components with eigenvalues greater than the average eigenvalue. As the trace of a matrix is equal to the sum of its eigenvalues, when doing correlation PCA this means that only the components with an eigenvalue greater than 1 will remain.

This procedure is not ideal since important information may be neglected.

In the book 'Principal Component Analysis' Ian Jolliffe [13] describes another way: 'Cumulative Percentage of Total Variation'. In short, one decides what percentage one wants to keep of the total variation. The number of principal components should be the smallest number with which the chosen percentage of the total variation is exceeded.

# 4 Preprocessing data

In the last chapter centering and standardization of the data (matrix) have been mentioned, but more things can be troubling. When a variable is not described with numeric values, it is not possible to center or standardize in the first place. Also missing data can become a problem when applying PCA. In this chapter solutions are given to deal with such problems.

## 4.1 Type of data pre-processing

In a data collection the following issues may arise:

- The data is not collected perfectly and errors may occur
- The data is incomplete
- Some variables do not have numerical values or have categorical values and should be transformed

#### Errors in data

As the data is provided by the Dutch Executive Agency for Education (DUO) the decision has been made to assume the majority of the data set is correct. However, when the values of one variable are very unlikely, the decision might be made to delete the whole variable or make a new variable with more suitable values. If one is not sure whether to keep or delete a variable, one can analyze a data set including the dubious variable and on a data set without.

#### Missing data

There are several ways to handle missing data. For some variables like second nationality, they are not only applicable to all pupils. All missing values could be filled in with a zero element. In the case of the variable second nationality, an 'unknown' element 0000 already exists and can be filled in for the unknown values.

In other cases, it is more suitable to fill in the average value or the most frequent value [3]. For example, if a variable has three different values and value $a$ occurs 80% of the time, it may be reasonable to fill in the missing values with $a$. When the different values are more evenly distributed, this approach seems less justified. In the case of an evenly distributed continuous variable, it can be chosen to replace the missing values with the average value. Replacement with the average value or most frequent value is an easy solution but has a disadvantage as it will reduce the variance of the variable and affects the results of the PCA.

#### Transforming data

A more difficult problem is how to deal with categorical variables. Categorical variables are variables with label values including binary, nominal and ordinal variables. Some examples: gender (with values: male, female), pet (e.g. dog, cat, duck). Each value represents a different category. Sometimes the categories are labeled with numerical values rather than names. It is however not possible to take sums or products as they do not make sense. Moreover, in the world of mathematics numbers have a natural ordering. This does not have to apply for all categorical variables. Doing calculations with integer encoding may, therefore, result in bad performance or unexpected results (predictions halfway two categories). [4] As the numbers do not make sense

from a mathematical perspective, categorical variables should be changed first. This will be done with the use of the Dummy Coding method.

The Dummy Coding method works as follows: if a variable takes on $k$ different values, $k$ dummy variables are made for each value. So each dummy variable is a different category within the variable. If a pupil's variable belongs to that category, the value of the dummy variable is equal to 1. If not, the value is 0. In other words, the dummy variable is an indicator function for that category. To reduce our data set size and to ensure that the variables are independent of each other, one dummy variable is taken out and can be obtained as the variable for which the value of the other dummy variables is 0.

For example, suppose the data set looks as follows:
In this example the only categorical variable is the gender of a pupil. The variable has three

| Name | Gender | Test score |
|---|---|---|
| Nico | Male | 547 |
| Simon | Male | 520 |
| Mindy | Female | 538 |
| Sammy | Unknown | 550 |

values: 'Male', 'Female' and 'Unknown'. For each value a column for their indicator function is made. This results in the following table: Now, the columns Gender and Gender_Unknown can

| Name | Gender | Test score | Gender_Male | Gender_Female | Gender_Unknown |
|---|---|---|---|---|---|
| Nico | Male | 547 | 1 | 0 | 0 |
| Simon | Male | 520 | 1 | 0 | 0 |
| Mindy | Female | 538 | 0 | 1 | 0 |
| Sammy | Unknown | 550 | 0 | 0 | 1 |

be deleted as they can be deduced from the column Gender_Male and Gender_Female. Therefore the final table is:

| Name | Test score | Gender_Male | Gender_Female |
|---|---|---|---|
| Nico | 547 | 1 | 0 |
| Simon | 520 | 1 | 0 |
| Mindy | 538 | 0 | 1 |
| Sammy | 550 | 0 | 0 |

## 4.2 Application on the data set

### Incorrect data

In the original data the following variable is included:
**LFT_1_JAN:** age of pupil on the first of January in the year of the pupil's attainment test

- Variable set is complete
- Domain = {3, 4, ..., 13}
- Most frequent: 10 (63.2%)
- Second and third most frequent: 11 (33.6%), 12 (1.9%)

- Variable type = integer variable
- Note: after some inspection the data is incorrect as no 3 year old has taken the attainment test. Also, most pupils should be 12 years old

As the variable **LFT_1_JAN** does not seem to be correct, the decision has been made to make a new variable **LFT_TOETS** by subtracting the year of the attainment test from a pupil's year of birth. As both years are available for all pupils, a complete variable set is attained with the following characteristics:
**LFT_TOETS:** age of pupil in the year of the primary school leavers attainment test

- Domain = {4, 5, ..., 15}
- Most frequent: 12 (62.6%)
- Second and third most frequent: 13 (33.1%), 14 (1.9%)
- Variable type: integer variable

As most pupils finish each grade in a year, it is likable most pupils are 12 years old when doing the attainment test. Moreover, normally no child can start primary school at the age of 3. Hence the new variable seems more logical and is added to the data set. Since this variable makes it also possible to compare all pupils without looking at the year of their attainment test, this variable will be used instead of the date of the test **DAT_TOETS**.

### Missing data

Before the data description some missing values have been replaced with the 'unknown' value for the following variables:

- **LAND_GEB** country of birth of pupil
- **LAND_OUDER1** country of birth one of the parents
- **NATIO1** (first) nationality of pupil

Also, the following changes were made before the data description:
Because preschool and early childhood activities are not widely known among parents, they usually give preference to the nursery school [7]. So the assumption has been made that pupils without available information for (one of) those variables are considered as non-participants of preschool or early childhood activities. In other words, the number of participation months is set to 0. The last variable that is completed before the data description is **GEWICHT**. The missing values of that variable are filled by the most frequent value: 0.

After the data description, there are 5 variables left with an incomplete data set: country of birth of the second parent (**LAND_OUDER2**), date of settlement in the Netherlands of the pupil (**DAT_VEST_NL**), second nationality (**NATIO2**), zip code of pupil (**PC4_LEERL**) and recommendation for secondary education levels (**ADVIES_VO**). As the date of settlement and second nationality do not apply to most pupils, the pupils without information are assumed to have only 1 nationality and are born in the Netherlands, respectively.
For the other variables, their most frequent value is substituted for the missing values. For the last two variables this has been done for the categories they will be subcategorized with in the next subsection.
For the distance to school variable (**AFSTAND**) 999999 is filled in when the distance could not be determined. As this value is not accurate, it is substituted by the average of the other values (808 m). This value should probably be larger as pupils living in neighboring countries originally had the value 999999 as well.

**Transforming data**

The following variables are categorical:

1. **CFINR**
2. **GESLACHT**
3. **LAND_GEB**
4. **LAND_OUDER1**
5. **LAND_OUDER2**
6. **DAT_VEST_NL**
7. **NATIO1**
8. **NATIO2**
9. **PC4_LEERL**
10. **BRIN**
11. **ACHTERGR**
12. **ADVIES_VO**
13. **GEWICHT**
14. **DAT_TOETS** (deleted after introduction of **LFT_TOETS**)
15. **TYPE_PO**
16. **DENOMINATIE**

**(1)**: As the variable **CFINR** is the unique number representing a pupil, the variable will be used as the unique index for a pupil.

**(2, 11, 13, 15)**: For variables with a small domain the Dummy Coding method is used directly. This has been done for variables **GESLACHT**, **ACHTERGR**, **GEWICHT** and **TYPE_PO**. Each dummy variable is named in the following way: suppose the dummy variable is the indicator for the value '**t**' in category **S**, then the dummy variable is called **S_t**.

**(3, 4, 5, 7, 8)**: Variables for which one value occurs noticeably more frequent than other values, it might be a good idea to only look at whether the most frequent value is applicable or not. This method is applied to variables **LAND_GEB**, **LAND_OUDER1**, **LAND_OUDER2**, **NATIO1** and **NATIO2**. For the first 4 variables, the most frequent value is the Netherlands or Dutch. The dummy variables are chosen to be named similar to their belonging original variable and adding '**_NL**' at the end. For **NATIO2** the most frequent variable is not having a second nationality. The dummy variable is called '**NATIO2_NVT**', based on the Dutch phrase for inapplicable ('niet van toepassing').

**(6)**: Like for **DAT_TOETS** it is possible to make an integer variable replacing **DAT_VEST_NL**. In this research, this is done by recreating the variable **MND_TOT_VEST_NL**. The variable represents the number of months a pupil has supposedly spend outside the Netherlands until his or her settlement. For pupils without a value for **DAT_VEST_NL** is assumed that they lived in the Netherlands their whole life. Their **MND_TOT_VEST_NL** value is 0. For pupils with a value (so those who are not born in the Netherlands) is assumed that their birth date is January 1st of their birth year.
The variable **MND_TOT_VEST_NL** has the following characteristics:

- Mean = 3.45 months
- Most frequent: 0 months (94.8%)
- Maximum = 171 months

For the remaining variables, there are too many different values to directly apply the Dummy Coding method. Therefore, they are sorted into bigger categories.
**(9)**: For the zip code variable (**PC4_LEERL**) the Dutch provinces have been used as categories. As foreign zip codes have been given certain values, they also got their own dummy variable. Moreover, the missing values have been categorized as part of the most common province (namely 'Zuid-Holland').

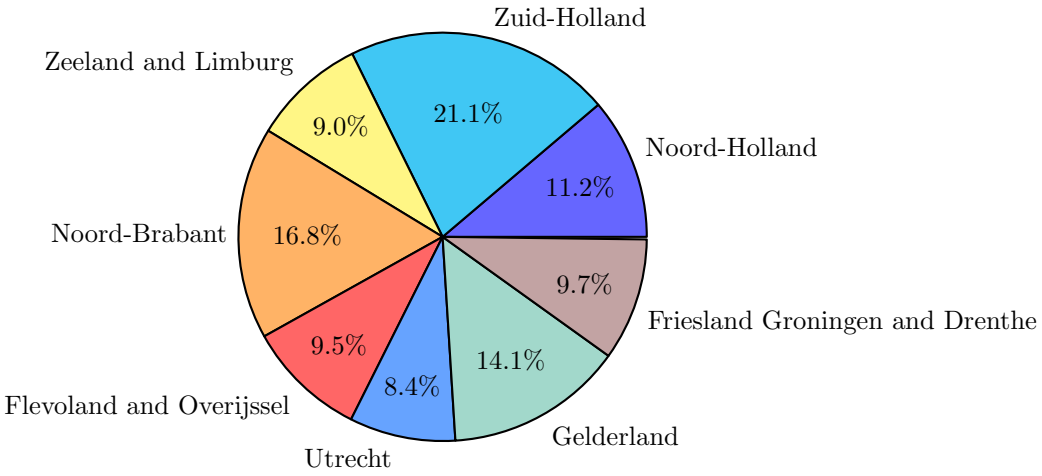



Figure 3: Percentages of pupils who completed an attainment test in 2008-2013 by province.

The dummy variables for the provinces are called **PC4_PROVINCE**, where 'PROVINCE' is substituted by the interested province. If a pupil has value 0 for all the dummy variables, the pupil's residence is located outside the Netherlands.

**(10)**: As there are too many schools to take into account separately, the decision has been made to look at the school sizes during the pupils' last year of primary school. Since the school sizes were sorted per reference date (variable **DAT_PEIL**, which is different for each year), the variable is (temporarily) added to be able to include the sizes. The school size variable is called **SCHOOLGROOTTE**.
A few characteristics and pie charts describing the school sizes in 2008-2013 are given below.

- Data is complete
- Mean (over the period 2008-2013) = 219.5 pupils
- Mean per year = 219.7 (2008), 221.0 (2009), 220.6 (2010), 219.3 (2011), 218.5 (2012), 217.9 (2013)
- Number of schools per year = 7165 (2008), 7182 (2009), 7157 (2010), 7112 (2011), 7040 (2012), 6954 (2013)
- Variable type = integer variable

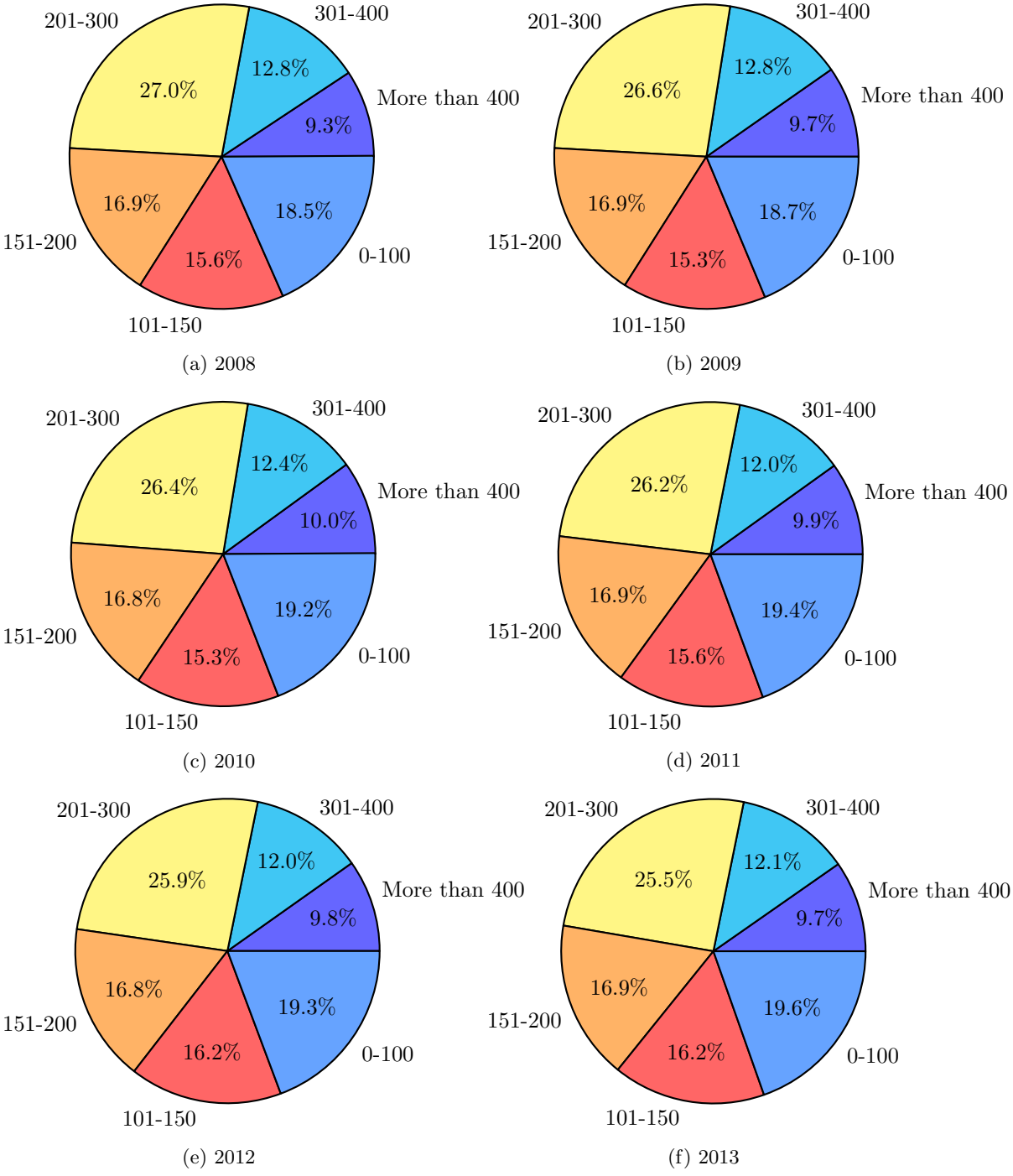



Figure 4: Classification of school sizes for different years.

**(12)**: In the Netherlands are roughly 4 streams for secondary education: 'praktijkonderwijs' (practical education), VMBO (pre-vocational secondary education), HAVO (senior general secondary education) and VWO (university preparatory education) [9]. These streams have smaller substreams within them, leading to 34 different secondary education recommendations. For the categories a finer partition than the 4 streams is chosen:

- secondary special education (VSO)
- practical education (PRO)
- VMBO B
- VMBO B up to VMBO K
- VMBO K
- VMBO K up to VMBO (G)T
- VMBO (G)T
- VMBO (G)T up to HAVO
- HAVO
- HAVO up to VWO
- VWO

If a recommendation consists of more than 2 'adjacent' categories, it is placed at the mixed recommendation starting with the lowest secondary education type. This ordering is based on information of the site 'Onderwijs in Cijfers' [14]. Pupils without a value are assigned to the most common category (VMBO (G)T).

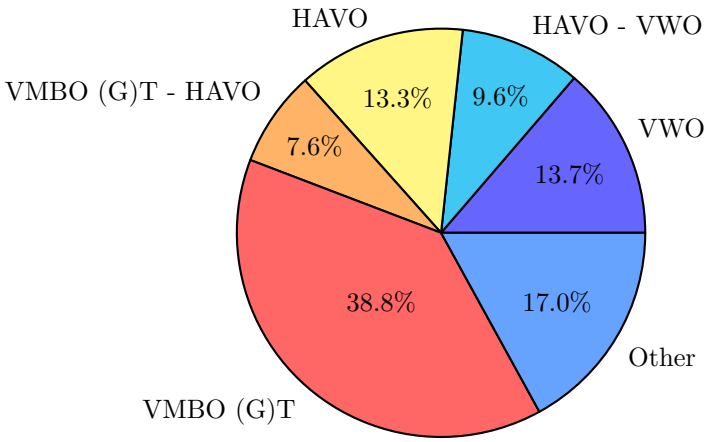



Figure 5: Percentages of pupils who completed an attainment test in 2008-2013 by received secondary school recommendation.

The dummy variables are named as follows (in the same order as the list given above): **ADVIES_VO_VSO**, **ADVIES_VO_PRO**, **ADVIES_VO_VMBOB**, **ADVIES_VO_VMBO_K**, **ADVIES_VO_VMBOK**, **ADVIES_VO_VMBOK_T**, **ADVIES_VO_VMBOGT**, **ADVIES_VO_VMBOGT_HAVO**, **ADVIES_VO_HAVO**, **ADVIES_VO_HAVOVWO**, **ADVIES_VO_VWO**. Note that secondary special education (VSO) is removed to reduce a dimension and ensure that all variables are independent.

**(16)**: For the denomination of the schools, 5 denominations are chosen to make a dummy variable of. The other denomination are represented in the 'left-out' dummy variable. This choice is based on research of CBS (Statistics Netherlands, also known as Centraal Bureau voor de Statistiek) [5]. The chosen denominations are (with the Dutch translation in the brackets):

- Non-enominational private education and public education (Algemeen Bijzonder & Openbaar)
- Roman Catholic (Rooms-Katholiek)
- Reformed/Reformational (Gereformeerd/Reformatorisch)
- Protestant Christian (Protestants-Christelijk)
- Islamic (Islamitisch)

The dummy variables are called **DENOMINATIE_OPB_ABZ**, **DENOMINATIE_RK**, **DENOMINATIE_REF_GEV**, **DENOMINATIE_PC** and **DENOMINATIE_ISL**, respectively.

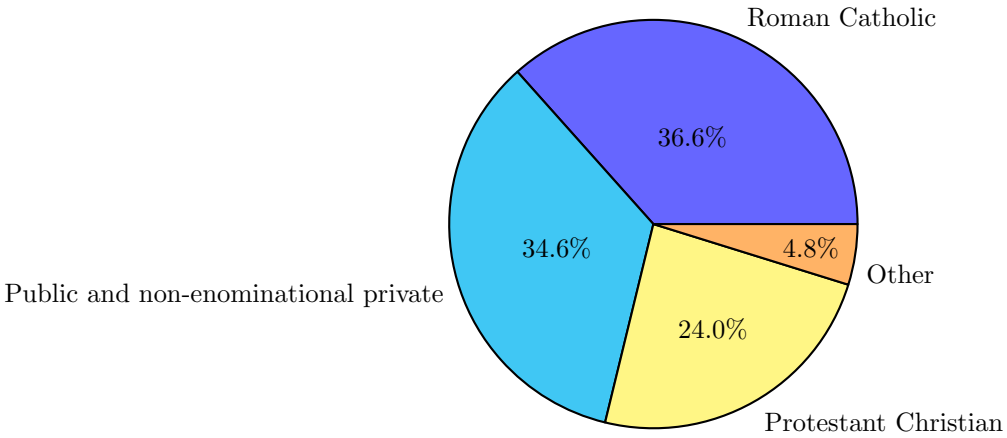



Figure 6: Percentages of pupils who completed an attainment test in 2008-2013 by school denomination.

Note that the categorie 'Other' in the pie chart includes Islamic and Reformed/Reformational.

# 5 Regression analysis

After preprocessing the data regression analysis is applied two times on the data: once with and once without PCA. Regression analysis is a method to determine how big of an impact other variables have on the variable of interest. In this case, the test score variable (**UITSLAG**) is the variable of interest (also called *response*). In regression analysis, it is assumed that the variable of interest is dependent on the other variables. The aim is to build a model to predict the outcome of the attainment test. Not only is the model useful for making predictions, also for analyzing the behavior of the data and finding the important variables [12]. As not many assumptions should be made when making a model, as a start a linear model is proposed. If in the future reasons appear to reject linear models, polynomial models are taken into account.

## 5.1 Linear regression

The simplest model is a linear model. In a linear model the response is assumed to be linearly dependent of the other variables. This can be represented in vector notation in the following way: let vector $\mathbf{y}$ represent the response and let $\mathbf{x}_j$'s denote the vectors of the $J$ (independent) variables on which $\mathbf{y}$ could be dependent. All vectors have length $I$ where $I$ is the number of observations. Let $\mathbf{1}$ denote the vector with only ones with also length $I$. The vector $\mathbf{y}$ can be expressed in the following way:

$$\hat{\mathbf{y}} = a_0\mathbf{1} + a_1\mathbf{x}_1 + \cdots + a_J\mathbf{x}_J$$

Here the $a_j$ terms are the coefficients or parameters of the model. Not every value of $\mathbf{y}$ will coincide with its predicted value of the linear model. The expression for $\mathbf{y}$ is actually $\mathbf{y} = a_0\mathbf{1} + a_1\mathbf{x}_1 + \cdots + a_J\mathbf{x}_J + \epsilon$ where $\epsilon$ is the error. The error term is usually neglected and used to evaluate different models. To make a distinction between $\mathbf{y}$ and its predicted value, $\hat{\mathbf{y}}$ is used for the latter.

For an even more compact notation of the model the variables $\mathbf{x_j}$'s can be put into a matrix $\mathbf{X}$ and all coefficients in a vector $\mathbf{a}$. Then

$$\hat{\mathbf{y}} = \mathbf{Xa}$$

where $\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_I \end{pmatrix}$ is the vector with all attainment test scores of the pupils, $\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix}$ is the data matrix of the variables (see Section 3) where each $j$'th column is equal to $\mathbf{x_j}$ and $\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_J \end{pmatrix}$ the vector with the coefficients.

### Simple linear regression

Simple linear regression only looks at the influence of one variable on the response. The model has the following expression:

$$\hat{\mathbf{y}} = a_0\mathbf{1} + a_1\mathbf{x} \tag{1}$$

Here $a_0$ is called the constant term or *intercept*.

There are different ways to create a model. One way is with *Ordinary Least Squares* (OLS). OLS chooses the parameters in such a way that the square of the difference between the predicted and actual values is minimized. In other words, $a_0$ and $a_1$ are found such that the following is minimized:

$$||\hat{\mathbf{y}} - \mathbf{y}||^2 = ||a_0\mathbf{1} + a_1\mathbf{x} - \mathbf{y}||^2 = \sum_{j=1}^{I}(a_0 + a_1x_j - y_j)^2 \tag{2}$$

As the difference between the predicted and actual values can be seen as the 'error', equation 2 is also called the 'sum of squared errors of prediction' (SSE). As SSE is a quadratic form of the parameters $\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$ with positive-definite Hessian [12], the SSE has a global minimum at a certain $\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix}$ (see proposition 2.54 of lecture notes Analysis 2 [6]).

### Multiple linear regression and polynomial regression

In most cases the response is not only dependent of just one variable. Therefore equation 1 is expanded to:

$$\hat{\mathbf{y}} = a_0\mathbf{1} + a_1\mathbf{x}_1 + \cdots + a_J\mathbf{x}_J \tag{3}$$

When the $\mathbf{x}_j$'s are replaced by polynomial terms of $\mathbf{x}_j$'s, the model is changed into one of polynomial regression. This can be formulated in the following way:

$$\hat{\mathbf{y}} = a_1\phi_1(\mathbf{x}_1) + \cdots + a_J\phi_J(\mathbf{x}_J) \tag{4}$$

where $\phi_j$ is the transformation function for the variable $\mathbf{x}_j$. With higher-order polynomials complex functions are better fitted. However, this also results in more complex computations and overfitting. Overfitting is when a model is too well-fitted on its original data set but does not perform well on unseen data.

### Comparing different models

When multiple models are fitted on the data set, it can be hard to decide which one is the best. There are multiple criteria for model selections, each with their own limitations. Model selection is therefore no easy task. For this research the decision has been made to look at the following goodness-of-fit tests and error measures:

- Coefficient of Determination ($R^2$)
- Adjusted $R^2$
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

A brief description of the model selection criteria:
$\mathbf{R^2}$: $R^2$ is a measure that represents the proportion of variance that can be explained by the model. $R^2$ is calculated by $\frac{\sum_{j=1}^{I}(\hat{y}_j-\bar{y})^2}{\sum_{j=1}^{I}(y_j-\bar{y})^2} = \frac{(||\hat{\mathbf{y}}-\bar{\mathbf{y}}||)^2}{(||\mathbf{y}-\bar{\mathbf{y}}||)^2}$ or its equivalent $1 - \frac{\sum_{j=1}^{I}(y_j-\hat{y}_j)^2}{\sum_{j=1}^{I}(y_j-\bar{y})^2} = 1 - \frac{(||\mathbf{y}-\hat{\mathbf{y}}||)^2}{(||\mathbf{y}-\bar{\mathbf{y}}||)^2}$, where $\bar{\mathbf{y}}$ denotes the vector with the sample mean $\bar{y}$ in all its components and $\hat{\mathbf{y}}$

represents the vector with the predicted values. One limitation of the $R^2$ is that when adding (non-relevant) dependent variables the $R^2$ increases whilst the model does not improve.
**Adjusted $\mathbf{R^2}$**: To deal with the limitation of $R^2$ described above, Adjusted $R^2$ can be used instead. It is defined as $R^2_{adj} = 1 - \frac{\sum_{j=1}^{I}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{I}(y_j - \bar{y})^2}\frac{I-1}{I-J-1} = 1 - \frac{(||\mathbf{y}-\hat{\mathbf{y}}||)^2}{(||\mathbf{y}-\bar{\mathbf{y}}||)^2}\frac{I-1}{I-J-1}$.
**Mean Absolute Error**: MAE is, as its name suggests, the average of the absolute difference between the actual and predicted values. In mathematical notation: MAE $= \frac{1}{I}\sum_{j=1}^{I}|y_j - \hat{y}_j| = \frac{||\mathbf{y}-\hat{\mathbf{y}}||_1}{I}$.
**Mean Squared Error**: MSE is, as its name suggests, the average of squared errors. Its formula is as follows: MSE $= \frac{1}{I}\sum_{j=1}^{I}(y_j - \hat{y}_j)^2 = \frac{||\mathbf{y}-\hat{\mathbf{y}}||^2}{I}$. Since the errors are squared before taken the average, large errors have more weight in MSE compared to MAE.
**Root Mean Squared Error**: RMSE is the square root of the Mean Squared Error, hence RMSE $= \sqrt{\frac{1}{I}\sum_{j=1}^{I}(y_j - \hat{y}_j)^2} = \frac{||\mathbf{y}-\hat{\mathbf{y}}||}{\sqrt{I}}$. RMSE is usually more favorable than MSE since its values are in the same measurement unit as the response.

## 5.2 Multiple linear regression applied on the data set

Now it is time to apply the theory from above on the education-related data set. The response in this case is the attainment test score and after the preprocess it has a total of 60 attributes (independent variables that could influence the response). First the data set is split into a training and test set. 80% of the pupils are randomly selected for the training set. On this set multiple linear regression is applied, leading to the coefficients of table 1 and an intercept equal to 535.2752743067.

| | Variable | Coefficient |
|---|---|---|
| 1 | VOORS_MND | -0.03968640330607055 |
| 2 | LEERJAAR | 0.04766189098672846 |
| 3 | VBJ_BO | 0.010902828486668015 |
| 4 | VBJ_SBO | -0.004535958858701172 |
| 5 | VBJ_SO | 0.002413867769439193 |
| 6 | VBJ_VSO | -0.00576283563725527 |
| 7 | VBJ_INS | -1.73928402094992 |
| 8 | LJR_INS_1E | -1.6825181109090939 |
| 9 | GROEPSGR | 0.05848789201770743 |
| 10 | VROEG_MND | -0.015465876491399605 |
| 11 | NVS | -0.22176669547823946 |
| 12 | GENERATIE | 0.26067473988247203 |
| 13 | GEBJAAR | 0.4515339627943983 |
| 14 | AFSTAND | -0.002969164817831091 |
| 15 | PC4_NOORDHOLLAND | -0.023679506722256716 |
| 16 | PC4_ZUIDHOLLAND | -0.1718766310362769 |
| 17 | PC4_ZEELAND | 0.02200049399546955 |
| 18 | PC4_NOORDBRABANT | 0.14983009657526164 |
| 19 | PC4_LIMBURG | 0.1892248632617686 |
| 20 | PC4_UTRECHT | 0.08619522874084992 |
| 21 | PC4_FLEVOLAND | -0.11405444721477898 |
| 22 | PC4_OVERIJSSEL | -0.01697483699401308 |
| 23 | PC4_GELDERLAND | 0.02261570788608605 |
| 24 | PC4_FRIESLAND | -0.010736229273216907 |

| | | |
|---|---|---|
| 25 | PC4_GRONINGEN | 0.03006089263911428 |
| 26 | PC4_DRENTHE | 0.01842481388849133 |
| 27 | GESLACHT_M | 1.8462336431130693 |
| 28 | GESLACHT_V | 1.6395703705497602 |
| 29 | ACHTERGR_1 | -0.08618854049460623 |
| 30 | ACHTERGR_2 | -0.17952518931608316 |
| 31 | TYPE_PO_SBO | -0.21515754341966012 |
| 32 | GEWICHT_0.25 | -0.04077406742445694 |
| 33 | GEWICHT_0.3 | -0.4384386728082275 |
| 34 | GEWICHT_0.4 | -0.015240394172047911 |
| 35 | GEWICHT_0.7 | -0.011409664573044237 |
| 36 | GEWICHT_0.9 | -0.046418739907701345 |
| 37 | GEWICHT_1.2 | -0.3466328469875535 |
| 38 | LFT_TOETS | -0.35673436051161 |
| 39 | NATIO1_NL | -0.011191554630694933 |
| 40 | LAND_GEB_NL | -0.05296782197244629 |
| 41 | LAND_OUDER1_NL | 0.27416927346598524 |
| 42 | LAND_OUDER2_NL | 0.4168291570976393 |
| 43 | NATIO2_NVT | 0.04405434534135677 |
| 44 | MND_TOT_VEST_NL | -0.026310258896004635 |
| 45 | ADVIES_VO_PRO | -0.6687650805149613 |
| 46 | ADVIES_VO_VMBOB | -1.3057121751366811 |
| 47 | ADVIES_VO_VMBOB_K | -0.34719517250290816 |
| 48 | ADVIES_VO_VMBOK | -0.03178217311269825 |
| 49 | ADVIES_VO_VMBOK_T | 0.39268645489774273 |
| 50 | ADVIES_VO_VMBOGT | 3.6087823042354463 |
| 51 | ADVIES_VO_VMBOGT_HAVO | 2.5469235605145797 |
| 52 | ADVIES_VO_HAVO | 4.468423340016744 |
| 53 | ADVIES_VO_HAVOVWO | 4.740397001936828 |
| 54 | ADVIES_VO_VWO | 6.846659283014065 |
| 55 | DENOMINATIE_RK | 0.1612836186146127 |
| 56 | DENOMINATIE_OPB_ABZ | 0.027427657802128677 |
| 57 | DENOMINATIE_PC | 0.11501802571852032 |
| 58 | DENOMINATIE_REF_GEV | 0.14983507916345076 |
| 59 | DENOMINATIE_ISL | 0.05927793166807496 |
| 60 | SCHOOLGROOTTE | 0.11488574236997615 |

Table 1: Coefficients of linear regression.

To see which variables have the most influence on or are most correlated with the attainment test scores and thus the largest coefficient in table 1, a plot is made of the absolute value of the variables:

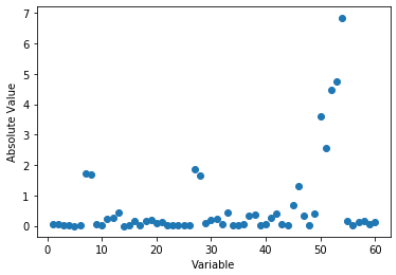



Figure 7: Absolute coefficients from linear regression (see table 1).

It seems like the secondary education recommendation (variable 45-54) is the most correlated with the attainment test scores. This is reasonable as the secondary education recommendation is made by the pupil's teacher. His or her recommendation is based on the child's learning process, achievements and progress during primary school. Moreover, starting from 2015 a new system has been introduced in which the teacher's recommendation is more important than the attainment test (Centrale Eindtoets) [15].

Looking at the variables with the smallest absolute coefficients it is remarkable that the variables representing the years of residence at the different primary school types (variable 3-6) are located in the 6 smallest coefficients. Furthermore, the second smallest absolute coefficient belongs to the variable **AFSTAND** (variable 14). This is not unlikely as the traveling distance should not affect a pupil's school performance.

After finding the coefficients and making the model, it is now possible to make predictions on the test set. In the figure below the predictions of 25 pupils are plotted next to the actual attainment test scores.

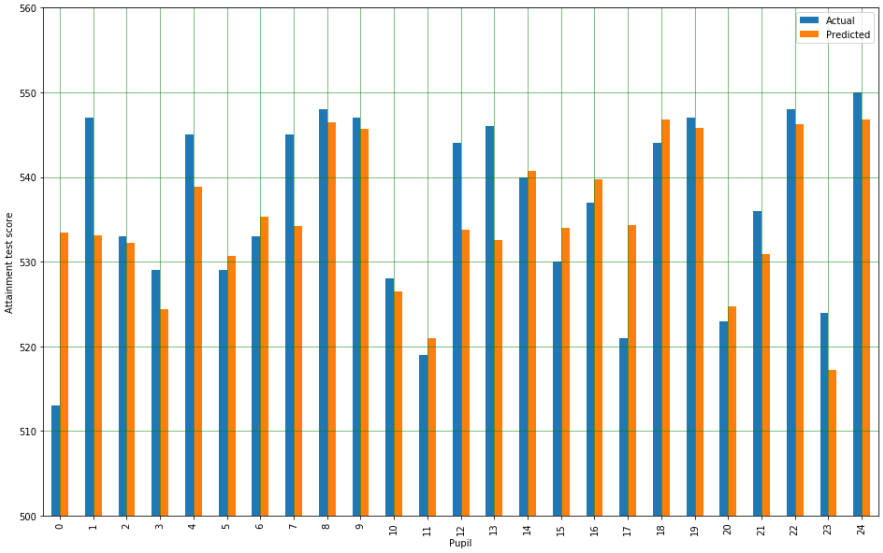



Figure 8: Predictions for 25 pupils of the test set.

In order to get some insight at the errors per test score, the average difference is plotted in a bubble plot. The average difference is calculated in the following way: Suppose $T$ students got an attainment test score of value $s$ and their predicted test score is denoted in a vector $\hat{\mathbf{y}}$ with length $T$. Then the average difference is $\frac{\sum_{j=1}^{T}(\hat{\mathbf{y}}_j - s)}{T}$. The sizes of bubbles are in proportion with the number of pupils achieving that test score.

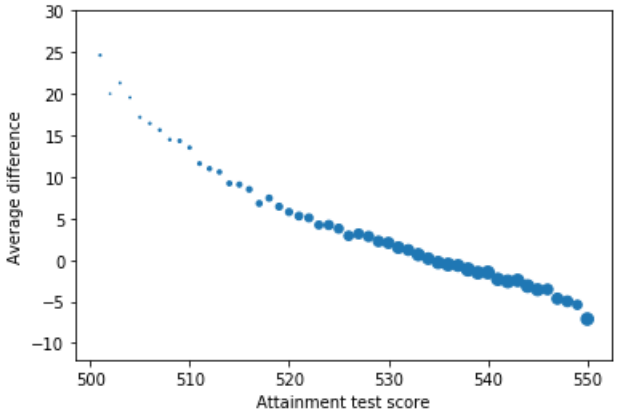



Figure 9: Average errors per attainment test score.

The bubble plot in figure 9 shows that around the intercept ($\approx 535$) the model makes pretty good predictions. However, the further the test score is from the intercept the larger the errors generally become.

Finally the performance of the model must be evaluated. These values are particularly useful later for comparing models made with PCA.

| $R^2$ | Adjusted $R^2$ | MAE | MSE | RMSE |
|---|---|---|---|---|
| 0.5838798 | 0.5837122 | 4.5614668 | 38.4256585 | 6.1988433 |

$R^2$ and Adjusted $R^2$ have a range from 0 to 1. The higher the value, the better fit a model provides. Knowing this, the values in the table above do not indicate high accuracy. However, since the research is a social study, $R^2$ and adjusted $R^2$ value higher than 0.5 are (relatively) adequate.

# 6 Multiple linear regression and PCA

Last chapter a linear model was created for the educational data set. In this chapter a linear model is created but after applying Principal Component Analysis and keeping different amounts of principal components.

## 6.1 Applying PCA

PCA is applied to the attribute matrix $\mathbf{X}$ from section 5.1. The steps of section 3.1 are followed. After preprocessing, the data step 1 and partly step 2 are already completed so this section starts with computation of the singular values and principal components. In Appendix A the largest and smallest 5 principal components can be found. As principal components consist of linear combinations of the original variables, it is hard to explain and describe what the principal components represent. This is one of the disadvantages of PCA.

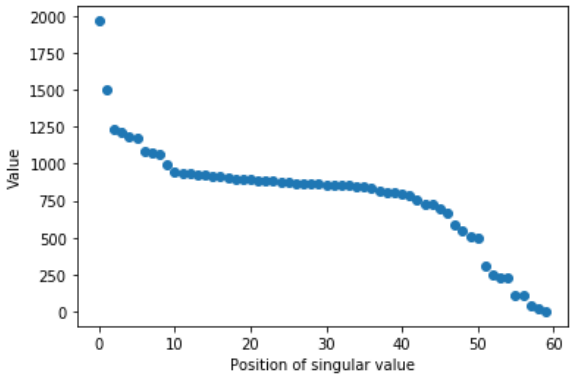



Figure 10: Scatterplot of the singular values.

As discussed in subsection 3.2 one way of choosing the number of components is by the elbow test. In the figure above it is possible to see an elbow around 10 principal components. However, keeping 10 principal components only preserves 36% of the total variance (see figure 11). This percentage feels a bit skimpy, so the decision is made to look at the procedures when keeping 30, 35, 40, 45, 50, 55 and 60 principal components.

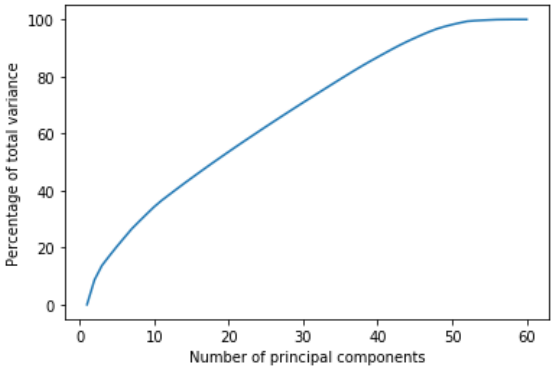



Figure 11: Percentage of variance per number of considered principal components.

Moreover, in figure 10 can be seen that the smallest few singular values are relatively close to 0. This usually points to dependency in the principal components. Suppose the Singular Value Decomposition of a matrix $\mathbf{X}$ is equal to $\mathbf{P\Delta Q^T}$ (see step 3 of section 3.1). As $\mathbf{Q}$ is orthonormal it holds that $\mathbf{XQ} = \mathbf{P\Delta}$. If the j'th singular value is 0, $\mathbf{Xq}_j = 0$. As not all entries of $\mathbf{q}_j$ are equal to 0, the columns of $\mathbf{X}$ are dependent and using the values of $\mathbf{q}_j$ it is possible to deduce which ones are dependent.

No singular value is exactly equal to 0. In fact, the smallest eigenvalue is equal to 1.2. Consequently because during the preprocessing of the data it is attempted to make the input variables independent. However, looking at the values of the smallest principal components can still say something about which variables are likely to be dependent of each other. This is deduced based on their absolute value. The variables with the largest absolute value in a principal component with a singular value near 0 are expected to be dependent. These variables are quite probable to be dependent based on the smallest 5 principal components, starting with the smallest principal component (see Appendix A)):

1. **GESLACHT_M** and **GESLACHT_V**: whether a pupil is male or female
2. **ADVIES_VO**-dummies, starting from VMBO B till VWO: most frequent secondary school recommendations
3. **PC4**-dummies, all provinces variables
4. **VBJ_INS** and **LJR_INS_1E**: years of residence at the school and first grade entering the school
5. Denomination dummies

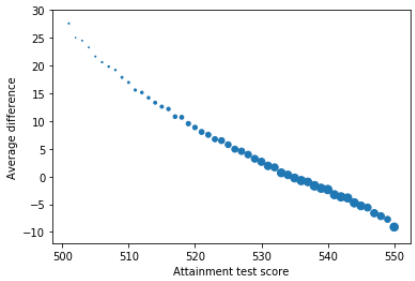
This is not a surprising result. Most of them are dummy variables. Per category one dummy variable has been left out to avoid dependency. When the least frequent value of a category is left out, the dummies are dependent for most of their entries. **VBJ_INS** and **LJR_INS_1E** can also be explained. The majority of the pupils do not skip a grade or do the same grade twice. Hence **LJR_INS_1E** is usually equal to 8-**VBJ_INS**.

Looking at the largest few principal components, a few things stand out. For the largest principal component the following variables have the largest absolute entries: **LAND_OUDER1_NL** (-0.363209), **LAND_OUDER2_NL** (-0.360827), **GENERATIE** (0.359564), **ACHTERGR_2** (0.354333), **ACHTERGR_1** (-0.338584) and **NATIO2_NVT** (-0.281772). All these variables are related to the background of a child and in combination with their values point to a pupil with a non-Dutch background.

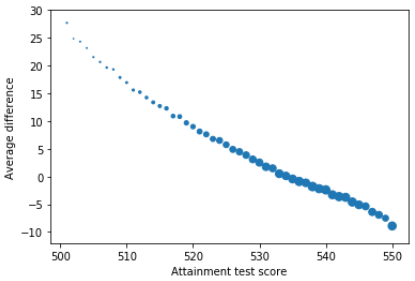
For the second principal component the following variables have the largest absolute entries: **VBJ_INS** (-0.496388), **LJR_INS_1E** (0.487234),**VBJ_BO** (-0.370645) and **NVS** (0.355818). These variables are all related to a pupil's primary education history.

## 6.2 Linear Regression after PCA

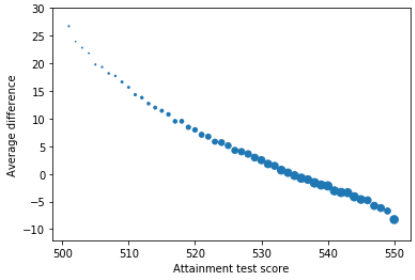
After deciding how many principal components are kept, the data set matrix $\mathbf{X}$ is transformed as described in step 5 of subsection 3.1. To see if PCA leads to better models, bubble plots are made and the criteria for the model selection are calculated. The coefficients of linear regressions can be found in appendix B.



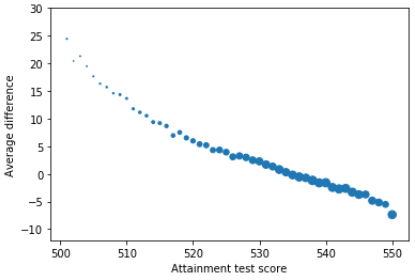

(a) 30 principal components



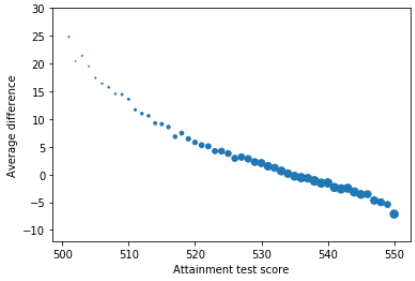

(b) 35 principal components



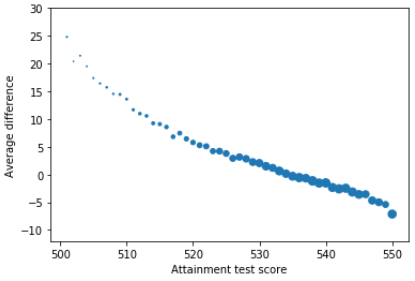

(c) 40 principal components



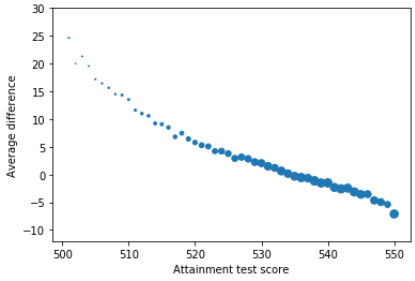

(d) 45 principal components



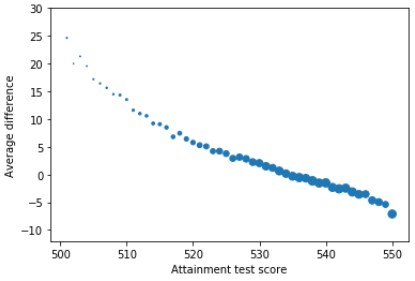

(e) 50 principal components




(f) 55 principal components




(g) 60 principal components




(h) Bubble plot of model without PCA

Figure 12: Average errors per attainment test scores when considering different numbers of principal components.

Looking carefully at figure 12 it is noticeable that the more principal components are kept, the less straight the plot becomes. The plots seems to 'flatten' towards 0. Also when keeping all 60 principal components the plot seems identical to the one without PCA. Furthermore, the range also becomes smaller.

| Model | total variance | $R^2$ | Adjusted $R^2$ | MAE | MSE | RMSE |
|---|---|---|---|---|---|---|
| 30 components | 72.6% | 0.4271701 | 0.4269393 | 5.7226669 | 54.5147375 | 7.3834096 |
| 35 components | 80.7% | 0.4165774 | 0.4193719 | 5.6666914 | 53.8748208 | 7.3399469 |
| 40 components | 88.2% | 0.4733848 | 0.4734239 | 5.3222746 | 48.629075 | 6.9734550 |
| 45 components | 94.7% | 0.5712632 | 0.5710904 | 4.6712722 | 39.5907197 | 6.2921156 |
| 50 components | 98.8% | 0.5825229 | 0.5823546 | 4.5646249 | 38.5509662 | 6.2089424 |
| 55 components | 99.9% | 0.5827078 | 0.5825396 | 4.5640610 | 38.5338893 | 6.2075671 |
| 60 components | 100% | 0.5838799 | 0.5837122 | 4.5614668 | 38.4256585 | 6.1988433 |
| Without PCA | 100% | 0.5838799 | 0.5837122 | 4.5614668 | 38.4256585 | 6.1988433 |

Table 2: Model selection criteria values for different models.

It seems like PCA does not improve linear regression. Looking at the model that keeps all principal components, all the model selection criteria have the same values. The same model is likely made when all principal components are considered.

The fact that PCA with fewer components does not improve the models can be explained due to the withdrawal of information and deleting dependent variables during the preprocessing of the data. To make a better comparison between PCA and no PCA the next chapter will look at what happens when 50 principal components, 50 random variables and 50 variables with the highest absolute coefficients in section 5.2 are kept.

# 7 Comparison model with and without PCA

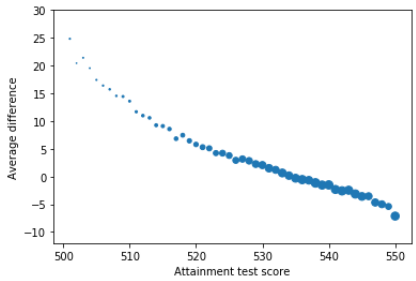
To make a better comparison between models with and without PCA this chapter will look at what happens when 50 principal components are kept and 50 variables are kept. For the latter 6 different sets with 50 variables will be used: 1 with the variables with the largest absolute coefficients of section 5.2, and the other sets are constructed by choosing 50 variables after shuffling all variables in different random ways.

The model with 50 principal component ('PCA50') has already been calculated in the chapter before. For the other models 10 variables are left out as described above. The model with the largest absolute coefficients is called 'LargeCoeff', the other models with randomly selected variables are called 'Random1', 'Random2', 'Random3', 'Random4' and 'Random5'. Their linear regression coefficients can be found in appendix C. In the following table the deleted variables can be found:

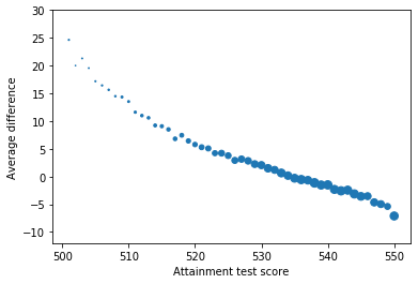
| LargeCoeff | Random1 | Random2 |
|---|---|---|
| **VBJ_SO** | **LAND_GEB_NL** | **GEWICHT_0.25** |
| **AFSTAND** | **LAND_OUDER2_NL** | **PC4_FRIESLAND** |
| **VBJ_SBO** | **VBJ_BO** | **SCHOOLGROOTTE** |
| **VBJ_VSO** | **ADVIES_VO_VMBOK_T** | **PC4_NOORDBRABANT** |
| **PC4_FRIESLAND** | **ADVIES_VO_VMBOGT_HAVO** | **NVS** |
| **VBJ_BO** | **ADVIES_VO_PRO** | **ADVIES_VO_PRO** |
| **NATIO1_NL** | **GEWICHT_0.4** | **GEWICHT_0.7** |
| **GEWICHT_0.7** | **GEWICHT_0.9** | **PC4_ZUIDHOLLAND** |
| **GEWICHT_0.4** | **LAND_OUDER1_NL** | **PC4_OVERIJSSEL** |
| **VROEG_MND** | **GESLACHT_V** | **TYPE_PO_SBO** |

| Random3 | Random4 | Random5 |
|---|---|---|
| **ADVIES_VO_PRO** | **PC4_FRIESLAND** | **VBJ_SO** |
| **ADVIES_VO_VMBOB_K** | **ADVIES_VO_HAVO** | **ACHTERGR_2** |
| **VBJ_BO** | **PC4_NOORDBRABANT** | **PC4_LIMBURG** |
| **VBJ_INS** | **PC4_GELDERLAND** | **LJR_INS_1E** |
| **ADVIES_VO_VMBOGT_HAVO** | **VBJ_BO** | **TYPE_PO_SBO** |
| **ADVIES_VO_VMBOK** | **PC4_UTRECHT** | **ADVIES_VO_PRO** |
| **ACHTERGR_2** | **GEWICHT_0.25** | **GEWICHT_1.2** |
| **PC4_UTRECHT** | **LJR_INS_1E** | **ACHTERGR_1** |
| **ADVIES_VO_VMBOK_T** | **ACHTERGR_1** | **SCHOOLGROOTTE** |
| **GEWICHT_0.9** | **ADVIES_VO_VMBOGT_HAVO** | **ADVIES_VO_HAVO** |

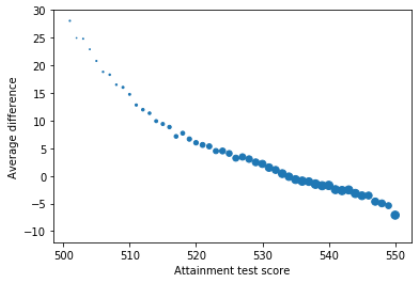
Table 3: The 10 deleted variables for the different models with 50 variables.



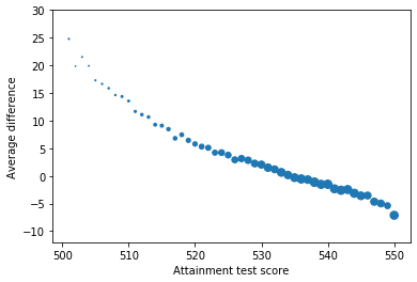

(a) PCA50



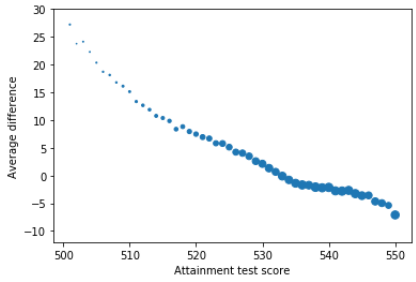

(b) LargeCoeff



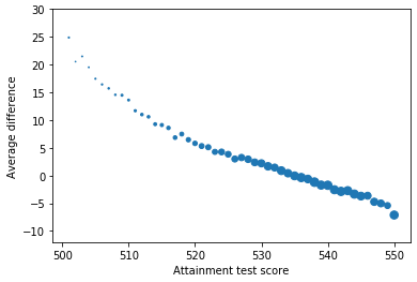

(c) Random1



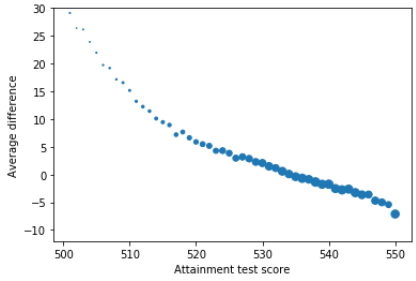

(d) Random2




(e) Random3




(f) Random4




(g) Random5

Figure 13: Average errors per attainment test scores when considering different models with 50 variables or principal components.

| Model | $R^2$ | Adjusted $R^2$ | MAE | MSE | RMSE |
|---|---|---|---|---|---|
| PCA50 | 0.5825229 | 0.5823546 | 4.5646249 | 38.5509662 | 6.2089424 |
| LargeCoeff | 0.5838902 | 0.5837505 | 4.5613890 | 38.4246994 | 6.1987660 |
| Random1 | 0.5615769 | 0.5614297 | 4.6896923 | 40.4851702 | 6.3627958 |
| Random2 | 0.5823417 | 0.5822014 | 4.5685596 | 38.5676972 | 6.2102896 |
| Random3 | 0.5250400 | 0.5248805 | 4.9680023 | 43.8590888 | 6.6226195 |
| Random4 | 0.5751122 | 0.5749696 | 4.6328139 | 39.2352852 | 6.2638076 |
| Random5 | 0.5599131 | 0.5597653 | 4.6436433 | 40.6388142 | 6.3748580 |

Table 4: Model selection criteria values for different models with 50 variables or principal components.

When comparing table 4 and figure 13 it is remarkable that the models which perform best with the model selection criteria also have the flattest curve in the plots at the most common attainment test scores. Furthermore, the following things can be deduced from table 4 and figure 13:

Comparing the plots in figure 13 of the random models and PCA50 it is noticeable that the range of the average error is bigger for most of the random models. In addition, the curve around the most frequent attainment test is flatter for PCA50. The values in table 4 are also in favor of PCA50: the errors are smaller and $R^2$ & adjusted $R^2$ are greater. Hence:
**The PCA model is a better fit than a random variable model.** This is as expected since PCA keeps most of the information by keeping the linear combinations of variables with the largest variances.

In figure 13 no clear difference can be seen between the plot of PCA50 and LargeCoeff. However, the model selection criteria are more in favor of LargeCoeff. Thus the following is concluded:
**The PCA model is a worse fit than the model which kept the variables with large absolute coefficients.** This is a surprising result. One can argue that as PCA takes linear combinations of the variables, each variable is taken (partly) into account and less important variables are less strongly included. However, this applies mainly to data sets with dependent variables. Figure 10 showed that all singular values were greater than 0. This indicates that the data did not contain dependent variables after preprocessing.
Another explanation is that the model selection criteria used in this research are not the right measures for PCA. It could be that PCA gives better results in other norms.

# 8 Hypothesis testing

In the last few chapters, all variables have been used in linear regression to describe a model. However, it is also interesting to know which variables significantly matter. It could be the case that the relationship that is found in the data set is the result of a random error and the variable does not matter for the response in general. In this chapter this will be done with statistical tests. There are various ways to do this: for example using confidence intervals, t-tests or with p-values. The latter has been chosen for this chapter.

## 8.1 Hypothesis testing and p-values

A brief description of the procedure [11]:

**Step 1: null hypothesis and alternative hypothesis:**

For linear regression it has been assumed that

$$\hat{\mathbf{y}} = a_0\mathbf{1} + a_1\mathbf{x}_1 + ... + a_J\mathbf{x}_J$$

If a coefficient $a_j = 0$, the model would not contain the variable $\mathbf{x}_j$ anymore and it is unlikely to be related to the response $\mathbf{y}$. Therefore the null hypothesis is '$a_j$ equals 0' and the alternative hypothesis is '$a_j$ is not equal to 0'. This is equivalent to the following null hypothesis: 'There is no relation between variable $\mathbf{x}_j$ and response $\mathbf{y}$'. The corresponding alternative hypothesis becomes: 'There is a relation between variable $\mathbf{x}_j$ and response $\mathbf{y}$'.

**Step 2: rejecting or failing to reject the null hypothesis**

Usually a significance level $\alpha$ is chosen equal to 0.05. This choice is also made for this chapter. The p-value represents the probability of observing results equal to, or more extreme than those actually observed, under the assumption that the null hypothesis is correct. If the p-value is smaller than the prespecified significance level, the null hypothesis is rejected and it is believed that there is a relation between variable $\mathbf{x}_j$ and response $\mathbf{y}$. Otherwise, the null hypothesis is not rejected.

## 8.2 Application on data set

For every variable in the data set the p-values for the t-statistics are calculated. As it might be interesting to compare the p-values with the absolute coefficients of linear regression, the latter has been included in the table and sorted from smallest to largest.

| Variables | p-values | absolute coefficients |
|---|---|---|
| VBJ_SO | 0.7693321081631866 | 0.002413867773525702 |
| AFSTAND | 0.7189685437453237 | 0.0029691648177805063 |
| VBJ_SBO | 0.7835313408794146 | 0.0045359588585279775 |
| VBJ_VSO | 0.42718465922871607 | 0.005762835634355312 |
| PC4_FRIESLAND | 0.7576570408628722 | 0.010736229273300424 |
| VBJ_BO | 0.5789582634114921 | 0.01090282848601945 |
| NATIO1_NL | 0.26383317511538434 | 0.011191554630754774 |
| GEWICHT_0.7 | 0.14375821521961332 | 0.011409664572998524 |
| GEWICHT_0.4 | 0.060712274527473395 | 0.015240394172043481 |
| VROEG_MND | 0.05880972549787158 | 0.015465876491487479 |

| | | |
|---|---|---|
| PC4_OVERIJSSEL | 0.7182941098578746 | 0.01697483699406699 |
| PC4_DRENTHE | 0.5459967629827291 | 0.018424813888544667 |
| PC4_ZEELAND | 0.4189314758966848 | 0.022000493995408932 |
| PC4_GELDERLAND | 0.7231195930806844 | 0.022615707886095127 |
| PC4_NOORDHOLLAND | 0.682199959116474 | 0.023679506722203758 |
| MND_TOT_VEST_NL | 0.01173240528763668 | 0.02631025889599349 |
| DENOMINATIE_OPB_ABZ | 0.4368657840505408 | 0.02742765780216186 |
| PC4_GRONINGEN | 0.3778568655269695 | 0.030060892639145464 |
| ADVIES_VO_VMBOK | 0.7186426102691765 | 0.031782173112711476 |
| VOORS_MND | 1.236376968469539e-06 | 0.03968640330462103 |
| GEWICHT_0.25 | 7.10714288404855e-07 | 0.04077406742442846 |
| NATIO2_NVT | 9.517685028986277e-05 | 0.044054345341316345 |
| GEWICHT_0.9 | 2.8098573250218464e-08 | 0.04641873990771189 |
| LEERJAAR | 5.139164566439764e-09 | 0.04766189098751333 |
| LAND_GEB_NL | 7.600281865436683e-07 | 0.05296782197245953 |
| GROEPSGR | 7.0990203517494175e-12 | 0.05848789201774798 |
| DENOMINATIE_ISL | 1.4933092275974355e-09 | 0.05927793166809292 |
| ACHTERGR_1 | 6.327020411296391e-05 | 0.08618854049447677 |
| PC4_UTRECHT | 0.09119952076586321 | 0.0861952287408636 |
| PC4_FLEVOLAND | 0.000114290721605114 | 0.1140544472147752 |
| SCHOOLGROOTTE | 9.49110628734506e-40 | 0.11488574236995008 |
| DENOMINATIE_PC | 0.0003276728696752873 | 0.11501802571856512 |
| PC4_NOORDBRABANT | 0.028517537102756146 | 0.14983009657521862 |
| DENOMINATIE_REF_GEV | 9.267701147228158e-24 | 0.1498350791634106 |
| DENOMINATIE_RK | 7.362429984292299e-06 | 0.1612836186146288 |
| PC4_ZUIDHOLLAND | 0.021294249737984605 | 0.17187663103630513 |
| ACHTERGR_2 | 1.3301380929911343e-14 | 0.1795251893159669 |
| PC4_LIMBURG | 5.075109370208938e-05 | 0.18922486326170007 |
| TYPE_PO_SBO | 2.918760233296622e-40 | 0.21515754341970483 |
| NVS | 9.916009982161483e-31 | 0.22176669547764993 |
| GENERATIE | 2.4959607444513504e-33 | 0.26067473988225676 |
| LAND_OUDER1_NL | 4.727132196874826e-56 | 0.274169273465952 |
| GEWICHT_1.2 | 2.5106183215622746e-281 | 0.3466328469875189 |
| ADVIES_VO_VMBOB_K | 2.526959892502644e-07 | 0.347195172502919 |
| LFT_TOETS | 8.469189757310759e-285 | 0.3567343605115614 |
| ADVIES_VO_VMBOK_T | 2.083626003260744e-12 | 0.3926864548977058 |
| LAND_OUDER2_NL | 3.127988976455579e-150 | 0.41682915709762874 |
| GEWICHT_0.3 | 0.0 | 0.4384386728082623 |
| GEBJAAR | 0.0 | 0.45153396279447267 |
| ADVIES_VO_PRO | 6.189330068789756e-198 | 0.6687650805149956 |
| ADVIES_VO_VMBOB | 9.740229963430656e-65 | 1.3057121751366312 |
| GESLACHT_V | 0.4557240040357742 | 1.6395703706083822 |
| LJR_INS_1E | 0.0 | 1.682518110934088 |
| VBJ_INS | 1.1228418228099499e-259 | 1.7392840209812872 |
| GESLACHT_M | 0.4009504045484015 | 1.8462336431713453 |
| ADVIES_VO_VMBOGT_HAVO | 4.1804591718015335e-152 | 2.5469235605145926 |
| ADVIES_VO_VMBOGT | 1.549546086797715e-91 | 3.6087823042356826 |
| ADVIES_VO_HAVO | 4.902715892576029e-283 | 4.468423340016759 |
| ADVIES_VO_HAVOVWO | 0.0 | 4.740397001936729 |

| | | |
|---|---|---|
| ADVIES_VO_VWO | 0.0 | 6.846659283014043 |

Table 5: p-values and absolute linear regression coefficients for model with all 60 variables.

The following variables have a p-value greater than 0.05:

- **VBJ_SBO**
- **VBJ_SO**
- **PC4_FRIESLAND**
- **PC4_GELDERLAND**
- **AFSTAND**
- **ADVIES_VO_VMBOK**
- **PC4_OVERIJSSEL**
- **PC4_NOORDHOLLAND**
- **VBJ_BO**
- **PC4_DRENTHE**
- **GESLACHT_V**
- **DENOMINATIE_OPB_ABZ**
- **VBJ_VSO**
- **PC4_ZEELAND**
- **GESLACHT_M**
- **PC4_GRONINGEN**
- **NATIO1_NL**
- **GEWICHT_0.7**
- **PC4_UTRECHT**
- **GEWICHT_0.4**
- **VROEG_MND**

Most variables for which the null hypothesis is accepted have indeed small absolute coefficients. However, the gender variables (**GESLACHT_M** and **GESLACHT_V**) are outliers in this case. In chapter 6 the singular values revealed that no variables were linearly dependent since they are all unequal to 0. However, the smallest principal component showed that even though the variables for male or female gender are not linearly dependent, they are dependent in the sense that they influence each other. This is not surprising since the third value 'O' for the variable **GESLACHT** only applied to 0.0003% of the pupils. Hence, the dummy variables **GESLACHT_M** and **GESLACHT_V** are opposite indicators in most of their components. Furthermore, the third largest principal component has both dummy variables as variables with the largest absolute coefficient. This should indicate that the gender of a pupil influences the test score.

It is likely that if the unknown gender value 'O' is substituted for the most frequent gender value 'V', the p-value for the gender dummy becomes small. This substitution is equivalent to deleting the dummy variable **GESLACHT_V** from the data set. The coefficients and p-values of linear regression can be found in appendix D. The corresponding p-value for **GESLACHT_M** turns out to be 1.64042e-144. This is small as expected and it can be concluded that there is a relation between the attainment test score and the gender of a pupil. As for the remaining variables, the variables with p-value greater than 0.05 stay the same.

The above is a nice example of how PCA can give insight into the data set and could help to improve a model.

# 9 Conclusion

The aim of this thesis was to find the variables which influence or correlated with the performance of primary school pupils, in particular their leavers attainment test score. The research's data is provided by DUO, the Dutch Executive Agency for Education. The data contains information about 744771 pupils who completed the attainment test in the years 2008-2013. Principal Component Analysis is applied due to a substantial amount of data. Linear regression models with and without PCA are made and compared to each other to evaluate if Principal Component Analysis indeed helps when handling a big data set.

Not all data was of interest and not all variable sets were complete. After a selection and preprocessing, only 60 variables were chosen to be considered besides the attainment test score. During this process, the variables were chosen carefully such that no variable was linearly dependent on the other variables.

After preprocessing linear regression is applied to the data set. In chapter 5 a linear model was made, taking all 60 variables into account. It turns out that teachers' secondary school recommendation correlated the most with the attainment test score. The years of residence at different primary school types and the traveling distance to school had the least influence on the attainment test score.

When applying PCA as well (in chapter 6), the linear model, however, did not give better results than the model without PCA. To conclude this, a model selection criteria was made, consisting of $R^2$, adjusted $R^2$, Mean Absolute Error, Mean Squared Error and Root Mean Squared Error. The worse result for the PCA model could be because there are no linearly dependent variables in the data set after preprocessing. Furthermore, when not all principal components are kept, some information will not be taken into account. However, the principal components did give insight into the data set. The smallest principal components gave an indication which variables are likely to influence each other, although they are not linearly dependent. The largest principal components also give information. Their components are constructed by the variables which provide the direction with the most variance.

To see if PCA keeps the most important information when not considering all principal components, 6 models with 50 variables and 1 model with 50 principal components were compared in chapter 7. For the 6 models with 50 variables, 1 model was made by keeping the variables with the largest absolute coefficient during linear regression with all 60 variables. For the other 5 models, 50 variables were randomly picked. As expected, the model with PCA was better than the randomly picked models. However, the model with the largest coefficients was a better fit according to the chosen model selection criteria. It could be that PCA does not optimize the model in the measures of the chosen criteria. Moreover, PCA is of great value when variables are not linearly independent of each other. This is not the case with the data after preprocessing.

Linear regression has been a great way to describe the relationship between the variables and the attainment test score. However, it is also interesting to know whether the relationship between a variable and the test score is significant. It could be that a variable has a significant high coefficient caused by a random error whilst the variable does not matter for the attainment test in general. To check this, hypothesis testing with p-values is used in chapter 8. In general, variables with a large p-value have a small absolute coefficient in the linear model as expected. However, this was not the case for the gender variables. They had p-values greater than the significant

level but a notable larger absolute coefficient. In chapter 6 the smallest principal component indicated that the gender variables were related although they were not linearly dependent. As well as that, the third largest principal component indicates that the gender variables influence the test score. Deleting the female gender variable resulted in a small p-value. Hence, there is indeed a relation between the attainment test score and the gender of a pupil.

In a nutshell, when one should take fewer variables into account it is in general better to use PCA than choose the variables randomly. However, when the variables are linearly independent, PCA might give a worse fitted model. Nevertheless, PCA is helpful in giving more insight into the data set.

# Appendices

## A Principal components

| Variable | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| VOORS_MND | 0.0216917 | -0.0177017 | -0.0095011 | 0.0050157 | -0.0024347 |
| LEERJAAR | -0.0023500 | -0.0145364 | -0.0052291 | 0.0213154 | -0.0216036 |
| VBJ_BO | -0.1394512 | -0.3706450 | -0.0334787 | -0.0834222 | -0.0320349 |
| VBJ_SBO | 0.02651354 | 0.1232668 | 0.2357719 | 0.0783706 | 0.5969530 |
| VBJ_SO | 0.0157769 | 0.0670214 | 0.0239234 | -0.0001008 | 0.0171421 |
| VBJ_VSO | 0.0018366 | 0.0096798 | 0.0061824 | 0.0006873 | 0.0123372 |
| VBJ_INS | -0.1663627 | -0.4963880 | 0.03690599 | -0.0009921 | 0.0977358 |
| LJR_INS_1E | 0.1724627 | 0.4872343 | -0.0234135 | 0.0075117 | -0.0924538 |
| GROEPSGR | -0.0543273 | -0.0297083 | -0.03043203 | 0.0837087 | -0.1187964 |
| VROEG_MND | 0.0240019 | -0.0189437 | -0.0134805 | -0.0294600 | 0.0030057 |
| NVS | 0.1110177 | 0.3558177 | -0.0514385 | -0.0642997 | -0.0889723 |
| GENERATIE | 0.3595643 | -0.1414819 | -0.0111209 | -0.0257950 | 0.0106879 |
| GEBJAAR | -0.0279272 | 0.0211530 | -0.1547634 | -0.1869646 | -0.0753717 |
| AFSTAND | 0.0252266 | 0.1063728 | -0.0009549 | -0.0483157 | 0.0191814 |
| PC4_NOORDHOLLAND | 0.0684281 | -0.0102842 | -0.0011622 | -0.0218246 | 0.0171185 |
| PC4_ZUIDHOLLAND | 0.0729314 | -0.0197191 | -0.0265583 | -0.1399667 | 0.0054775 |
| PC4_ZEELAND | -0.0129955 | 0.0173040 | 0.0006259 | -0.0403071 | 0.0287781 |
| PC4_NOORDBRABANT | -0.0360475 | -0.0267829 | 0.0622129 | 0.3342073 | -0.0683555 |
| PC4_LIMBURG | -0.0214381 | -0.0037959 | 0.0375566 | 0.2423702 | -0.0574547 |
| PC4_UTRECHT | -0.0032683 | -0.0129599 | -0.0136839 | -0.0509648 | -0.0049328 |
| PC4_FLEVOLAND | 0.0249711 | 0.0310881 | -0.0196380 | -0.0848052 | 0.0109627 |
| PC4_OVERIJSSEL | -0.0307644 | -0.0034473 | -0.0038122 | -0.0075356 | -0.0019532 |
| PC4_GELDERLAND | -0.0429333 | 0.0134577 | -0.0082918 | -0.0556164 | 0.0297190 |
| PC4_FRIESLAND | -0.0300627 | 0.0225601 | -0.0270880 | -0.1491470 | 0.0460584 |
| PC4_GRONINGEN | -0.0159585 | 0.0313694 | -0.0185466 | -0.1328490 | 0.0336007 |
| PC4_DRENTHE | -0.0214101 | 0.0249250 | -0.0158826 | -0.1047116 | 0.0267651 |
| GESLACHT_M | -0.0017894 | -0.0102981 | 0.6296803 | -0.1934756 | -0.2402279 |
| GESLACHT_V | 0.0017828 | 0.0102898 | -0.6296793 | 0.1934726 | 0.2402348 |
| ACHTERGR_1 | -0.3385842 | 0.1621075 | 0.0025049 | 0.0140088 | -0.0305370 |
| ACHTERGR_2 | 0.3543329 | -0.1687154 | -0.0013459 | -0.0109508 | 0.0300350 |
| TYPE_PO_SBO | 0.0268609 | 0.1249214 | 0.2351318 | 0.0778039 | 0.5967947 |
| GEWICHT_0.25 | -0.0100573 | 0.0009728 | 0.0249532 | 0.0379473 | 0.0236659 |
| GEWICHT_0.3 | 0.0118377 | -0.0060616 | 0.0012178 | 0.0033643 | 0.0246156 |
| GEWICHT_0.4 | -0.0012840 | 0.0043955 | 0.0015162 | 0.0025917 | 0.0020454 |
| GEWICHT_0.7 | -0.0010803 | 0.0005046 | 0.0042906 | 0.0059499 | 0.0043187 |
| GEWICHT_0.9 | 0.0776917 | -0.0389816 | 0.0281753 | 0.0475621 | 0.0136197 |
| GEWICHT_1.2 | 0.2377271 | -0.1149496 | -0.0050107 | -0.0057500 | 0.0074695 |
| LFT_TOETS | 0.0751711 | -0.0259350 | 0.1375115 | 0.1003495 | 0.0458696 |
| NATIO1_NL | -0.1843562 | -0.0494182 | -0.0113889 | -0.1118223 | 0.0919894 |
| LAND_GEB_NL | -0.1518885 | -0.1110549 | -0.0085296 | -0.1399568 | 0.1225206 |
| LAND_OUDER1_NL | -0.3632085 | 0.1138449 | 0.0046887 | -0.0051338 | 0.0111315 |
| LAND_OUDER2_NL | -0.3608266 | 0.1030547 | 0.0053216 | 0.0008373 | 0.0027035 |
| NATIO2_NVT | -0.2817724 | 0.1531543 | 0.0082769 | 0.0227080 | -0.0290040 |

| | | | | | |
|---|---|---|---|---|---|
| MND_TOT_VEST_NL | 0.1528813 | 0.1876935 | 0.0109276 | 0.1376723 | -0.1352156 |
| ADVIES_VO_PRO | 0.0269692 | 0.0033858 | 0.0368126 | 0.0191402 | 0.0825822 |
| ADVIES_VO_VMBOB | 0.0569041 | -0.0188877 | 0.0483718 | 0.0418341 | 0.0875512 |
| ADVIES_VO_VMBOB_K | 0.0219299 | -0.0087235 | 0.0301712 | 0.0191244 | 0.0397469 |
| ADVIES_VO_VMBOK | 0.0263475 | -0.0178801 | 0.0335788 | 0.0418155 | 0.0354471 |
| ADVIES_VO_VMBOK_T | 0.0001764 | -0.00834 | 0.0224363 | 0.0549141 | 0.0107097 |
| ADVIES_VO_VMBOGT | 0.0105715 | 0.0041407 | -0.1114954 | -0.1405877 | -0.0660173 |
| ADVIES_VO_VMBOGT_HAVO | -0.0021068 | -0.0067663 | 0.0179635 | 0.0482343 | 0.0142757 |
| ADVIES_VO_HAVO | -0.0208737 | 0.0023871 | 0.0184557 | 0.0396940 | -0.0048189 |
| ADVIES_VO_HAVOVWO | -0.0254899 | 0.0027013 | 0.0267302 | 0.0314820 | 0.0014830 |
| ADVIES_VO_VWO | -0.0403540 | 0.0260150 | 0.01666913 | 0.0032693 | -0.0332668 |
| DENOMINATIE_RK | -0.0537236 | -0.0590061 | 0.1005263 | 0.5656643 | -0.1179117 |
| DENOMINATIE_OPB_ABZ | 0.0686812 | 0.0539556 | -0.0508232 | -0.2906471 | 0.0384134 |
| DENOMINATIE_PC | -0.02150498 | 0.0063922 | -0.0483125 | -0.2786571 | 0.0816982 |
| DENOMINATIE_REF_GEV | -0.0410493 | 0.0134282 | -0.0123249 | -0.0765458 | 0.0209905 |
| DENOMINATIE_ISL | 0.0954370 | -0.0423667 | -0.0118549 | -0.0108495 | 0.0075098 |
| SCHOOLGROOTTE | 0.0064329 | -0.0293878 | -0.0103776 | 0.1174497 | -0.1261503 |

Table 6: First 5 principal components.

| Variable | 56 | 57 | 58 | 59 | 60 |
|---|---|---|---|---|---|
| VOORS_MND | 0.0003814 | -0.0003261 | 1.9542787e-05 | 3.6721730e-05 | -5.1215340e-08 |
| LEERJAAR | 0.0016410 | 0.0164890 | 9.5315789e-06 | -5.1995518e-06 | 2.8696892e-07 |
| VBJ_BO | 0.0020025 | 0.1275860 | 0.0017418 | 7.2740703e-05 | -4.8942961e-06 |
| VBJ_SBO | 4.8488287e-05 | 0.0213433 | 0.0003743 | -0.0003409 | -1.2206521e-06 |
| VBJ_SO | 0.0002584 | 0.0108897 | 0.0002608 | 5.8881733e-05 | -9.7371319e-07 |
| VBJ_VSO | 0.0001549 | 0.0012053 | 3.6444587e-05 | 1.3749264e-05 | -1.3457353e-07 |
| VBJ_INS | -0.0085687 | -0.7683452 | -0.0009673 | -0.0001292 | -1.0483289e-05 |
| LJR_INS_1E | -0.0079927 | -0.61642679 | -0.0009232 | -0.0002090 | -2.4902938e-05 |
| GROEPSGR | 0.0028069 | 0.0003567 | 0.0001207 | 0.0001235 | -3.1175190e-06 |
| VROEG_MND | 0.0003023 | 0.0002116 | 3.0488603e-05 | -6.7091573e-05 | 6.8242050e-08 |
| NVS | -0.0016605 | -0.0990775 | -0.0005982 | 9.1259481e-05 | 1.4520225e-05 |
| GENERATIE | 0.0039690 | 0.0014381 | -0.0140454 | -6.3401309e-05 | 1.3346772e-05 |
| GEBJAAR | -0.0022498 | -0.0039379 | -0.0002196 | 0.0001568 | -4.1915789e-06 |
| AFSTAND | -0.0027746 | -0.0001497 | 0.0001925 | -0.0001949 | -5.2716419e-07 |
| PC4_NOORDHOLLAND | -0.0098825 | 0.0014836 | -0.3373403 | -0.0002632 | -1.9792392e-05 |
| PC4_ZUIDHOLLAND | -0.0010074 | 0.0015910 | -0.4371065 | -0.0004146 | -2.6576106e-05 |
| PC4_ZEELAND | -0.0019599 | -0.0001379 | -0.1530437 | -0.0003523 | -1.0354017e-05 |
| PC4_NOORDBRABANT | 0.0040194 | 0.0004757 | -0.4000921 | -0.0004478 | -2.5555146e-05 |
| PC4_LIMBURG | 0.0018208 | 2.1274114e-05 | -0.2707500 | -0.0004213 | -1.7589118e-05 |
| PC4_UTRECHT | 0.0033909 | 8.0417754e-05 | -0.2969115 | -0.0003269 | -1.9166824e-05 |
| PC4_FLEVOLAND | -0.0032634 | 0.0002286 | -0.1675293 | -0.0002291 | -1.0052033e-05 |
| PC4_OVERIJSSEL | 0.0010792 | -0.0018984 | -0.2731884 | -0.0003753 | -1.8319263e-05 |
| PC4_GELDERLAND | -0.0011678 | 0.0006048 | -0.3731271 | -0.0005179 | -2.4883349e-05 |
| PC4_FRIESLAND | 0.0021389 | 0.0003949 | -0.1993017 | -0.0001786 | -1.3767521e-05 |
| PC4_GRONINGEN | 0.0026492 | 0.0018399 | -0.1950394 | -0.0001738 | -1.3067738e-05 |
| PC4_DRENTHE | 0.0012870 | 0.0002232 | -0.1733487 | -0.0002001 | -3.4679753e-05 |
| GESLACHT_M | -2.4645268e-05 | 0.0008137 | 4.2539133e-06 | -0.0001444 | -0.7071070 |
| GESLACHT_V | 2.9610813e-05 | -0.0007834 | 9.1365316e-05 | 0.0001505 | -0.7071066 |

| | | | | | |
|---|---|---|---|---|---|
| ACHTERGR_1 | 0.0019941 | -0.0020402 | -0.0002969 | -4.5343919e-05 | -1.6723482e-06 |
| ACHTERGR_2 | 0.0090001 | -0.0011811 | 0.0029387 | 5.9486773e-05 | -1.9725599e-06 |
| TYPE_PO_SBO | 0.0017337 | -0.0028001 | -7.7581921e-05 | -0.0006747 | 8.1469271e-08 |
| GEWICHT_0.25 | 7.6530401e-05 | 0.0001146 | -3.9854365e-05 | -5.1886962e-06 | -4.2132398e-07 |
| GEWICHT_0.3 | -0.0004189 | 0.0015403 | -0.0001103 | -9.9779187e-05 | 1.4587845e-06 |
| GEWICHT_0.4 | 0.0001131 | -0.0001136 | -1.3279057e-06 | 6.5116160e-06 | -6.7183282e-08 |
| GEWICHT_0.7 | -0.0010456 | -6.6894864e-05 | -1.7077242e-05 | 7.7127189e-06 | -7.7919449e-08 |
| GEWICHT_0.9 | 0.0009655 | 3.1392847e-05 | 0.0005526 | -1.4170328e-06 | 2.6429438e-06 |
| GEWICHT_1.2 | 0.0011229 | 0.0015345 | 0.0007397 | -5.9573576e-05 | 8.2030413e-06 |
| LFT_TOETS | 0.0006787 | 0.0484826 | -0.0002498 | -0.0001162 | 6.0024208e-06 |
| NATIO1_NL | -0.0047584 | -0.0007257 | 0.0072974 | 5.1984146e-05 | 1.9266335e-05 |
| LAND_GEB_NL | -0.0002120 | 0.0013221 | 0.0084788 | -7.0205841e-06 | 2.2918908e-06 |
| LAND_OUDER1_NL | 0.0124384 | 0.0004799 | -0.0063080 | -6.2243048e-05 | 9.4469878e-06 |
| LAND_OUDER2_NL | 0.0117876 | -0.0004583 | -0.0047806 | -2.6376181e-05 | 8.5925431e-06 |
| NATIO2_NVT | -0.0070148 | 0.0002737 | -0.0012172 | 3.5028526e-05 | 4.9901762e-07 |
| MND_TOT_VEST_NL | 0.00058380 | 0.0057345 | 0.0066098 | 1.3374063e-05 | -1.5120060e-05 |
| ADVIES_VO_PRO | -0.0003579 | 0.0017684 | 9.0341000e-05 | -0.0640583 | 3.7305632e-07 |
| ADVIES_VO_VMBOB | -0.0004374 | 0.0057579 | 0.0003419 | -0.2357852 | -2.2728077e-07 |
| ADVIES_VO_VMBOB_K | 0.0006385 | 0.0037416 | 0.0001615 | -0.2063197 | -4.8421743e-07 |
| ADVIES_VO_VMBOK | -0.0001985 | 0.0043292 | 0.0002681 | -0.2711035 | -2.8407641e-07 |
| ADVIES_VO_VMBOK_T | 0.0003066 | 0.00211577 | 0.0001755 | -0.1706029 | -4.3677350e-07 |
| ADVIES_VO_VMBOGT | -6.0084115e-05 | 0.0020427 | 0.0005597 | -0.5481596 | -2.3096864e-06 |
| ADVIES_VO_VMBOGT_HAVO | -0.0003392 | -0.0003628 | 0.0003060 | -0.2980887 | -4.7006584e-07 |
| ADVIES_VO_HAVO | 0.0002035 | -0.0027311 | 0.0003688 | -0.3825770 | -5.4692852e-06 |
| ADVIES_VO_HAVOVWO | 0.0006479 | -0.0036192 | 0.0004579 | -0.3313769 | -9.9686918e-07 |
| ADVIES_VO_VWO | -0.0001010 | -0.0060251 | 0.00062450 | -0.3868046 | -2.9112712e-07 |
| DENOMINATIE_RK | -0.5872077 | 0.0077252 | -0.0004608 | -5.1705364e-05 | -1.5376233e-06 |
| DENOMINATIE_OPB_ABZ | -0.5764217 | 0.0068064 | 0.0004471 | -0.0001089 | -3.2383383e-06 |
| DENOMINATIE_PC | -0.5197525 | 0.0047711 | 0.0006491 | -0.0001005 | -5.3868692e-07 |
| DENOMINATIE_REF_GEV | -0.2094590 | 0.0045976 | 0.0001940 | 9.5924378e-05 | -3.7549982e-07 |
| DENOMINATIE_ISL | -0.0900368 | 0.0008905 | -0.0004949 | 2.4890596e-05 | 1.2242951e-07 |
| SCHOOLGROOTTE | -0.0001780 | 0.0002819 | 0.0005328 | 5.2701305e-05 | -3.8945637e-06 |

Table 7: Last 5 principal components.

# B Coefficients from linear regression with PCA (chapter 6)

Table with :

| Model | intercept |
|---|---|
| 30 components | 535.2775224931693 |
| 35 components | 535.2772404246008 |
| 40 components | 535.2772904798455 |
| 45 components | 535.2753179090021 |
| 50 components | 535.2751694680306 |
| 55 components | 535.2753185545196 |
| 60 components | 535.2752743066999 |

Table 8: Intercepts of linear regression with PCA considering different number of principal components.

| Principal component | Coefficient |
|---|---|
| 1 | -0.9612193079438958 |
| 2 | 0.2620313394637823 |
| 3 | -0.1428714963331799 |
| 4 | -0.09962318397730188 |
| 5 | -0.9382381884117886 |
| 6 | 0.32241311698388575 |
| 7 | -2.2067621979065204 |
| 8 | 1.7469978163218034 |
| 9 | 1.4330427411245161 |
| 10 | -2.988949646571449 |
| 11 | 0.8961693795164289 |
| 12 | -0.3433437958779084 |
| 13 | 0.6571279278614697 |
| 14 | -0.036839043646693825 |
| 15 | -0.3646337192711625 |
| 16 | 0.42065102221682205 |
| 17 | -0.549335020722833 |
| 18 | -0.01618064634377303 |
| 19 | 0.1243827185129832 |
| 20 | 0.9217821740260445 |
| 21 | 0.25258142801772576 |
| 22 | 0.11671901890453296 |
| 23 | -0.05557105697897097 |
| 24 | -0.030053176636535728 |
| 25 | -0.06813615814234927 |
| 26 | -0.04055793586815898 |
| 27 | -0.05314952579687138 |
| 29 | -0.04076844307112783 |
| 30 | -0.3223512902559822 |

Table 9: Linear regression coefficients for model with 30 principal components.

| Principal component | Coefficient |
|---|---|
| 1 | -0.9613872587583775 |
| 2 | 0.262125496642333 |
| 3 | -0.14953844932260527 |
| 4 | -0.0885694721723683 |
| 5 | -0.9067092037146367 |
| 6 | 0.3160803977364943 |
| 7 | -2.2208991330054277 |
| 8 | 1.7336935427938547 |
| 9 | 1.4285601440964455 |
| 10 | 3.0527142618854404 |
| 11 | 0.9147373143554578 |
| 12 | -0.7141503123854125 |
| 13 | 0.3991561787698055 |
| 14 | 0.11273418581721131 |
| 15 | 0.31133029367654264 |
| 16 | -0.2525040679532017 |
| 17 | 0.6511331144244731 |
| 18 | 0.06275934184071195 |
| 19 | -0.2879966901013076 |
| 20 | 0.7123535122170872 |
| 21 | 0.19057061501269473 |
| 22 | 0.14390817903360395 |
| 23 | 0.11988908260617966 |
| 24 | -0.2456064650676204 |
| 25 | -0.12355002846337869 |
| 26 | 0.19952809016335435 |
| 27 | -0.022810111705303532 |
| 28 | -0.12982953862849383 |
| 29 | 0.6055184807599273 |
| 30 | -0.009609682809515394 |
| 31 | -0.648857107373407 |
| 32 | -0.5905350414044285 |
| 33 | -0.4298906085964583 |
| 34 | 0.4496522798190996 |
| 35 | 0.46355508498569287 |

Table 10: Linear regression coefficients for model with 35 principal components.

| Principal component | Coefficient |
|---|---|
| 1 | -0.9601226626854137 |
| 2 | 0.26220054631100753 |
| 3 | -0.14911776268508575 |
| 4 | -0.08960769997501258 |
| 5 | -0.9131161169403658 |
| 6 | 0.31727303315306554 |
| 7 | -2.215122634170209 |
| 8 | 1.7445519406817007 |
| 9 | 1.424206929510197 |

| | |
|---|---|
| 10 | 3.052080066848769 |
| 11 | 0.7976996764595469 |
| 12 | -0.8019973850482218 |
| 13 | 0.4071174835845186 |
| 14 | 0.0786460140980898 |
| 15 | -0.28581284456242734 |
| 16 | -0.2681561611324681 |
| 17 | 0.6080981246687529 |
| 18 | 0.03350666305696127 |
| 19 | -0.2745146806473837 |
| 20 | 0.6588667284885038 |
| 21 | 0.003473237568493659 |
| 22 | 0.12635041010832182 |
| 23 | 0.07744104416563272 |
| 24 | 0.25937826623379917 |
| 25 | -0.13145161786231743 |
| 26 | 0.12104327805163936 |
| 27 | -0.0621072158634311 |
| 28 | -0.04323235676556664 |
| 29 | 0.6098297573493682 |
| 30 | 0.0988164746209288 |
| 31 | -0.6777159558230479 |
| 32 | 0.18186430343208565 |
| 33 | -0.2576270335170783 |
| 34 | -0.2656876781871654 |
| 35 | -0.025532609616330637 |
| 36 | 0.3149470740909925 |
| 37 | 0.09791191454894699 |
| 38 | 1.2383950962105348 |
| 39 | 0.498980462865391 |
| 40 | -2.0119054881868506 |

Table 11: Linear regression coefficients for model with 40 principal components.

| Principal component | Coefficient |
|---|---|
| 1 | -0.9599252941455159 |
| 2 | 0.26143141397244674 |
| 3 | -0.1522420927401214 |
| 4 | -0.08933409053928246 |
| 5 | -0.9147551725457616 |
| 6 | 0.31676574777051597 |
| 7 | -2.214674009242326 |
| 8 | 1.7468978465611014 |
| 9 | 1.4223355712084733 |
| 10 | 3.0526527392768115 |
| 11 | 0.7982921777455165 |
| 12 | -0.8022060431164075 |
| 13 | 0.4095234183475527 |
| 14 | 0.08054559046977236 |

| | |
|---|---|
| 15 | -0.28661373760831754 |
| 16 | -0.2710024356404884 |
| 17 | 0.6047654455262049 |
| 18 | 0.03293755823013611 |
| 19 | -0.273090034008286 |
| 20 | 0.661604846547468 |
| 21 | 0.005844630152040953 |
| 22 | 0.1260228888827805 |
| 23 | 0.08083611025656151 |
| 24 | 0.25916728548688894 |
| 25 | -0.1294463977031617 |
| 26 | 0.1269954598012767 |
| 27 | -0.05186657651519827 |
| 28 | -0.04776062819280033 |
| 29 | 0.6078699738589275 |
| 30 | 0.09585582993182777 |
| 31 | -0.664600875560921 |
| 32 | 0.20044757027842086 |
| 33 | -0.2599626951635347 |
| 34 | -0.26085480981427767 |
| 35 | -0.03095714975116609 |
| 36 | 0.3139070435979446 |
| 37 | 0.09694675838383326 |
| 38 | 1.2314613008610131 |
| 39 | 0.46296237051148226 |
| 40 | -2.0222928515187926 |
| 41 | -0.5240823753329973 |
| 42 | -2.9544530830728553 |
| 43 | 1.4058459647579435 |
| 44 | -0.3456332845687043 |
| 45 | -0.0671279987692115 |

Table 12: Linear regression coefficients for model with 45 principal components.

| Principal component | Coefficient |
|---|---|
| 1 | -0.9603058999691908 |
| 2 | 0.2615025720523081 |
| 3 | -0.1519040547871885 |
| 4 | -0.09005968725711488 |
| 5 | -0.9150927609844761 |
| 6 | 0.31629115639974004 |
| 7 | -2.2153877867869083 |
| 8 | 1.746513298889934 |
| 9 | 1.4218519020779334 |
| 10 | 3.052748471003039 |
| 11 | 0.7979394626963247 |
| 12 | -0.8014948601822556 |
| 13 | 0.4093287131074125 |
| 14 | 0.07952084342353745 |

| | |
|---|---|
| 15 | -0.28717618828573555 |
| 16 | -0.2715278254174632 |
| 17 | 0.6049798858819508 |
| 18 | 0.032676148648517624 |
| 19 | -0.2733456967025728 |
| 20 | 0.6618055083793194 |
| 21 | 0.0049741360729064765 |
| 22 | 0.12579083239948008 |
| 23 | 0.0801906052104236 |
| 24 | 0.2605214694182617 |
| 25 | -0.12902412996265614 |
| 26 | 0.12703135885014005 |
| 27 | -0.05167520086707239 |
| 28 | -0.04885131744987936 |
| 29 | 0.607562793375104 |
| 30 | 0.09653242092982717 |
| 31 | -0.6632952522757627 |
| 32 | 0.20056144504860773 |
| 33 | -0.26023321711898495 |
| 34 | -0.2607019345934719 |
| 35 | -0.03044382687429692 |
| 36 | 0.31303121409536416 |
| 37 | 0.09782319208002987 |
| 38 | 1.2325656548981763 |
| 39 | 0.4632884817525757 |
| 40 | -2.0226929851268847 |
| 41 | -0.5236647438316834 |
| 42 | -2.9547786171961326 |
| 43 | 1.4050434942251666 |
| 44 | -0.3454315479601584 |
| 45 | -0.06652267301219439 |
| 46 | 0.2250097353264413 |
| 47 | 0.8381745527844957 |
| 48 | 0.7121835408668333 |
| 49 | -0.8594984631583907 |
| 50 | -0.07681464366538515 |

Table 13: Linear regression coefficients for model with 50 principal components.

| Principal component | Coefficient |
|---|---|
| 1 | -0.9602516206826853 |
| 2 | 0.26160648387689345 |
| 3 | -0.1519054482151374 |
| 4 | -0.09009435171440083 |
| 5 | -0.9149561811604161 |
| 6 | 0.3163934483821258 |
| 7 | -2.215371927131956 |
| 8 | 1.7463999301853868 |
| 9 | 1.4218086006596955 |

| | |
|---|---|
| 10 | 3.052819539552167 |
| 11 | 0.7979158803503887 |
| 12 | -0.8014616014691922 |
| 13 | 0.40932177516678003 |
| 14 | 0.07954200767360578 |
| 15 | -0.28711732293331094 |
| 16 | -0.27146194152003433 |
| 17 | 0.605132340708331 |
| 18 | 0.032602134359136066 |
| 19 | -0.273235648861082 |
| 20 | 0.6617417969500685 |
| 21 | 0.004887828562330804 |
| 22 | 0.125764259248825 |
| 23 | 0.0801927784129699 |
| 24 | 0.2604457745992694 |
| 25 | -0.12897897274527642 |
| 26 | 0.12713699685513422 |
| 27 | -0.05178655057547754 |
| 28 | -0.04889369406456742 |
| 29 | 0.6076134007987619 |
| 30 | 0.09659460906724962 |
| 31 | -0.6634268941725989 |
| 32 | 0.20062349469188956 |
| 33 | -0.2600983867689184 |
| 34 | -0.2608337511895559 |
| 35 | -0.03044784398750383 |
| 36 | 0.3131183251571302 |
| 37 | 0.09791723706793765 |
| 38 | 1.232600701369724 |
| 39 | 0.46331891831556987 |
| 40 | -2.022513595227837 |
| 41 | -0.5235490130664546 |
| 42 | -2.954814444455777 |
| 43 | 1.4050114673981784 |
| 44 | -0.34550638297961533 |
| 45 | -0.06670222958238192 |
| 46 | 0.22495247786557643 |
| 47 | 0.8381888641225257 |
| 48 | 0.712126728786478 |
| 49 | -0.8593962962113338 |
| 50 | -0.07674417363691344 |
| 51 | -0.043053180028127525 |
| 52 | 0.13601845622055536 |
| 53 | -0.42321876418714033 |
| 54 | 0.23186205801047205 |
| 55 | 0.06350008599570844 |

Table 14: Linear regression coefficients for model with 55 principal components.

| Principal component | Coefficient |
|---|---|
| 1 | -0.9603716787513387 |
| 2 | 0.2612403735183215 |
| 3 | -0.15242360519919118 |
| 4 | -0.09018072303610702 |
| 5 | -0.9159811504330808 |
| 6 | 0.3162940702252721 |
| 7 | -2.215602870766765 |
| 8 | 1.7458126424178744 |
| 9 | 1.4214715047171624 |
| 10 | 3.0526941191303765 |
| 11 | 0.7978846605984065 |
| 12 | -0.8017318160803575 |
| 13 | 0.4093187190606035 |
| 14 | 0.07962133924050747 |
| 15 | -0.2869758595789268 |
| 16 | -0.27157257127137113 |
| 17 | 0.6049548103386534 |
| 18 | 0.0329717147559038 |
| 19 | -0.27296455042994183 |
| 20 | 0.6619703249821518 |
| 21 | 0.005246029079573167 |
| 22 | 0.12572389710586807 |
| 23 | 0.0798472727286002 |
| 24 | 0.2602448791893589 |
| 25 | -0.12895519306363618 |
| 26 | 0.12720030958403855 |
| 27 | -0.05147294484359738 |
| 28 | -0.048730967955499804 |
| 29 | 0.607332018004547 |
| 30 | 0.09648891929486705 |
| 31 | -0.6626836711765303 |
| 32 | 0.20039799401964825 |
| 33 | -0.26064620977574066 |
| 34 | -0.2612591424181 |
| 35 | -0.0309988212096296 |
| 36 | 0.31360151732127617 |
| 37 | 0.09710713811680705 |
| 38 | 1.232221253037026 |
| 39 | 0.46368215206605556 |
| 40 | -2.0224318007088304 |
| 41 | -0.5238304670598505 |
| 42 | -2.9547685201672413 |
| 43 | 1.4049275813258464 |
| 44 | -0.3450814088226246 |
| 45 | -0.06653221946055687 |
| 46 | 0.22458628387339188 |
| 47 | 0.8379072351271322 |
| 48 | 0.7119977762178702 |

| | |
|---|---|
| 49 | -0.8600488365106758 |
| 50 | -0.07695295440411165 |
| 51 | -0.0426172748515905 |
| 52 | 0.1360180231218683 |
| 53 | -0.42237526922395696 |
| 54 | 0.23309933529069948 |
| 55 | 0.06315353515566346 |
| 56 | -0.167694680216951 |
| 57 | 2.307790701593801 |
| 58 | -0.04348714372761553 |
| 59 | -8.301402217193736 |
| 60 | -2.464821920105003 |

Table 15: Linear regression coefficients for model with 60 principal components.

# C Coefficients for the model comparison (chapter 7)

| Variable | Coefficient |
|---|---|
| VOORS_MND | -0.041593955035025894 |
| LEERJAAR | 0.04769255080008561 |
| VBJ_INS | -1.7251057108611239 |
| LJR_INS_1E | -1.6808010880440272 |
| GROEPSGR | 0.058685871645344 |
| NVS | -0.21373889287293885 |
| GENERATIE | 0.26083974258027554 |
| GEBJAAR | 0.45214172575865474 |
| PC4_NOORDHOLLAND | -0.00701117203145174 |
| PC4_ZUIDHOLLAND | -0.15075694649727212 |
| PC4_ZEELAND | 0.029649949997763658 |
| PC4_NOORDBRABANT | 0.16985185528550018 |
| PC4_LIMBURG | 0.2027781503025471 |
| PC4_UTRECHT | 0.10108243954365448 |
| PC4_FLEVOLAND | -0.10590830483927077 |
| PC4_OVERIJSSEL | -0.003325745362127236 |
| PC4_GELDERLAND | 0.041314336575261346 |
| PC4_GRONINGEN | 0.03948538505608981 |
| PC4_DRENTHE | 0.027089921206455536 |
| GESLACHT_M | 1.827953573340229 |
| GESLACHT_V | 1.6212181976678508 |
| ACHTERGR_1 | -0.08619426992935969 |
| ACHTERGR_2 | -0.17835003248089476 |
| TYPE_PO_SBO | -0.22055599652709004 |
| GEWICHT_0.25 | -0.040669780887159224 |
| GEWICHT_0.3 | -0.4384585060049618 |
| GEWICHT_0.9 | -0.04599088492824521 |
| GEWICHT_1.2 | -0.3455716825084313 |
| LFT_TOETS | -0.3562512319995586 |
| LAND_GEB_NL | -0.05663186555260946 |
| LAND_OUDER1_NL | 0.2718641474127452 |
| LAND_OUDER2_NL | 0.4153075150611082 |
| NATIO2_NVT | 0.0483793699026028 |
| MND_TOT_VEST_NL | -0.026724824019203397 |
| ADVIES_VO_PRO | -0.6687094728029384 |
| ADVIES_VO_VMBOB | -1.3051186761725129 |
| ADVIES_VO_VMBOB_K | -0.3466882960654396 |
| ADVIES_VO_VMBOK | -0.031026864478035945 |
| ADVIES_VO_VMBOK_T | 0.3931955620676232 |
| ADVIES_VO_VMBOGT | 3.6098601467460765 |
| ADVIES_VO_VMBOGT_HAVO | 2.5477749074821263 |
| ADVIES_VO_HAVO | 4.469521416877054 |
| ADVIES_VO_HAVOVWO | 4.74128951661166 |
| ADVIES_VO_VWO | 6.847708716460304 |
| DENOMINATIE_RK | 0.1630427413715851 |

| | |
|---|---|
| DENOMINATIE_OPB_ABZ | 0.02883825102631822 |
| DENOMINATIE_PC | 0.11622651107417215 |
| DENOMINATIE_REF_GEV | 0.14981208104007754 |
| DENOMINATIE_ISL | 0.05948192234748795 |
| SCHOOLGROOTTE | 0.11486345472484982 |

Table 16: Linear regression coefficients for model LargeCoeff.

| Variable | Coefficient |
|---|---|
| PC4_GRONINGEN | -0.04037178614056482 |
| ADVIES_VO_VWO | 4.322425466999827 |
| ADVIES_VO_VMBOB_K | -1.6763147163228305 |
| VBJ_SBO | -0.030426439669972716 |
| GESLACHT_M | 0.20905288306336123 |
| GEWICHT_1.2 | -0.4412310069874585 |
| PC4_OVERIJSSEL | -0.12416565698853987 |
| NVS | -0.2114996578593376 |
| ADVIES_VO_HAVOVWO | 2.581018659228672 |
| PC4_UTRECHT | 0.0034648512927671904 |
| NATIO2_NVT | 0.06493497613610283 |
| DENOMINATIE_PC | 0.1652066419095145 |
| PC4_NOORDBRABANT | 0.05291143895550454 |
| GEWICHT_0.25 | -0.06170359037080303 |
| ADVIES_VO_VMBOGT | 0.05176751253381917 |
| GEWICHT_0.3 | -0.5086899004012335 |
| ADVIES_VO_VMBOK | -1.7802914877760228 |
| NATIO1_NL | 0.0590374898086897 |
| DENOMINATIE_OPB_ABZ | 0.07100759506032395 |
| AFSTAND | -0.0008646675789038549 |
| PC4_GELDERLAND | -0.0854156052559063 |
| ADVIES_VO_HAVO | 1.9773923912403806 |
| TYPE_PO_SBO | -0.26611952856874455 |
| GEWICHT_0.7 | -0.025490398518671353 |
| PC4_FRIESLAND | -0.0740904038010311 |
| VBJ_SO | 0.0043485123484954605 |
| PC4_NOORDHOLLAND | -0.12057829946826004 |
| ACHTERGR_2 | -0.2746558537871088 |
| ACHTERGR_1 | -0.08543915676160241 |
| DENOMINATIE_RK | 0.20509790929518146 |
| DENOMINATIE_ISL | 0.06619427036829251 |
| PC4_LIMBURG | 0.09315914825003846 |
| SCHOOLGROOTTE | 0.12666403356004574 |
| PC4_FLEVOLAND | -0.16901742445801632 |
| PC4_DRENTHE | -0.02558085544535893 |
| VBJ_INS | -1.8799113824184703 |
| ADVIES_VO_VMBOB | -2.8164826073428673 |
| LJR_INS_1E | -1.8489582524882089 |
| DENOMINATIE_REF_GEV | 0.1646641881864773 |
| LEERJAAR | 0.06931341790457246 |

| | |
|---|---|
| PC4_ZEELAND | -0.0013146586823694906 |
| VOORS_MND | -0.038826371151623795 |
| PC4_ZUIDHOLLAND | -0.2966397954456983 |
| VBJ_VSO | -0.005630807970177024 |
| GENERATIE | -0.21506179009665516 |
| VROEG_MND | -0.014950071716723912 |
| GROEPSGR | 0.07072867565340957 |
| GEBJAAR | 0.4512662615584938 |
| MND_TOT_VEST_NL | -0.06384700996885742 |
| LFT_TOETS | -0.42508344153710725 |

Table 17: Linear regression coefficients for model Random1.

| Variable | Coefficient |
|---|---|
| LFT_TOETS | -0.36044259102317955 |
| ADVIES_VO_VMBOB | 0.8476822378769873 |
| GEBJAAR | 0.44760300886829407 |
| AFSTAND | -0.007409782658190731 |
| GESLACHT_V | 1.4796616253192083 |
| PC4_LIMBURG | 0.18233764178899503 |
| GEWICHT_0.3 | -0.4392636884506125 |
| LEERJAAR | 0.050028516450033145 |
| PC4_ZEELAND | 0.02955787219524386 |
| GEWICHT_0.9 | -0.04716780456550973 |
| MND_TOT_VEST_NL | -0.033651570412176035 |
| PC4_GELDERLAND | 0.04770724904294721 |
| DENOMINATIE_OPB_ABZ | 0.05280145207470066 |
| ADVIES_VO_VMBOGT | 8.620553578838985 |
| ACHTERGR_2 | -0.17204400766336114 |
| ADVIES_VO_HAVO | 7.971475494917828 |
| LAND_OUDER1_NL | 0.29576597756657297 |
| GEWICHT_1.2 | -0.34854865156715037 |
| ADVIES_VO_VMBOGT_HAVO | 5.276041964670859 |
| VROEG_MND | -0.02017016214517009 |
| GEWICHT_0.4 | -0.013654107095573853 |
| VBJ_BO | -0.16414166946811048 |
| LAND_OUDER2_NL | 0.4424863738941686 |
| PC4_GRONINGEN | 0.04738186867775529 |
| VBJ_VSO | -0.008482364242264673 |
| DENOMINATIE_PC | 0.0995950666851966 |
| VBJ_INS | -1.4761089077204068 |
| GENERATIE | 0.2696410210288698 |
| VOORS_MND | -0.03429950689906079 |
| LJR_INS_1E | -1.6965825033725377 |
| DENOMINATIE_ISL | 0.06736463094588191 |
| NATIO1_NL | -0.016153196377816648 |
| ADVIES_VO_HAVOVWO | 7.771782699594434 |
| ADVIES_VO_VMBOK_T | 1.95476537732275 |
| GESLACHT_M | 1.6890268493184175 |

| | |
|---|---|
| NATIO2_NVT | 0.03272440335179544 |
| ADVIES_VO_VMBOB_K | 1.5349676819732279 |
| DENOMINATIE_REF_GEV | 0.14306429342762006 |
| GROEPSGR | 0.08307218605433508 |
| ADVIES_VO_VWO | 10.386952900977118 |
| PC4_NOORDHOLLAND | 0.012953125963742929 |
| ACHTERGR_1 | -0.06807848374677683 |
| LAND_GEB_NL | -0.05386201321320094 |
| PC4_FRIESLAND | 0.0019396118519894034 |
| PC4_UTRECHT | 0.11905960137126256 |
| VBJ_SBO | -0.21557539466714148 |
| DENOMINATIE_RK | 0.26040769360499 |
| PC4_DRENTHE | 0.030562205478065374 |
| VBJ_SO | -0.02190443001694231 |
| ADVIES_VO_VMBOK | 2.4501203802724283 |

Table 18: Linear regression coefficients for model Random2.

| Variable | Coefficient |
|---|---|
| DENOMINATIE_REF_GEV | 0.09133460101989843 |
| ADVIES_VO_HAVOVWO | 3.701891837697102 |
| GEWICHT_1.2 | -0.4698561328982632 |
| GEWICHT_0.7 | -0.029149930602179186 |
| PC4_NOORDHOLLAND | -0.1125820225357339 |
| GEWICHT_0.4 | -0.020629903032365338 |
| TYPE_PO_SBO | -0.279850287669399 |
| LAND_GEB_NL | -0.07164236898032617 |
| GROEPSGR | 0.06338284238809155 |
| DENOMINATIE_OPB_ABZ | 0.014827484372980315 |
| LEERJAAR | 0.02396317266811554 |
| PC4_LIMBURG | 0.1255722570199585 |
| GENERATIE | 0.31978733914909285 |
| PC4_OVERIJSSEL | -0.0733344878752179 |
| VROEG_MND | -0.011785013852466303 |
| DENOMINATIE_RK | 0.1191254282489671 |
| LFT_TOETS | -0.7166694451087046 |
| PC4_NOORDBRABANT | 0.03705500853616406 |
| DENOMINATIE_ISL | 0.062281206003113626 |
| GESLACHT_V | 1.7375441061628676 |
| GEWICHT_0.25 | -0.10328874856164688 |
| LAND_OUDER2_NL | 0.490945362603194 |
| ADVIES_VO_VWO | 5.629768946943436 |
| PC4_ZUIDHOLLAND | -0.29705393568563637 |
| VBJ_SBO | -0.017625457903413133 |
| VBJ_VSO | -0.0038845931889318966 |
| PC4_FLEVOLAND | -0.20137081668542156 |
| DENOMINATIE_PC | 0.09750980927155875 |
| VOORS_MND | -0.03433697686128478 |
| ADVIES_VO_VMBOGT | 1.9554737882966482 |

| | |
|---|---|
| ADVIES_VO_HAVO | 3.284246235439582 |
| AFSTAND | 0.0014162922159738356 |
| VBJ_SO | 0.009982740894975056 |
| GEWICHT_0.3 | -0.61030692110737 |
| SCHOOLGROOTTE | 0.13321660186723072 |
| PC4_FRIESLAND | -0.10106894426575802 |
| PC4_DRENTHE | -0.03974395861015731 |
| GESLACHT_M | 1.9404550160128031 |
| ADVIES_VO_VMBOB | -1.984743379788458 |
| LAND_OUDER1_NL | 0.3387574804258756 |
| PC4_ZEELAND | -0.01787428311742345 |
| NVS | -0.17148571036223337 |
| ACHTERGR_1 | 0.06730318643151534 |
| NATIO2_NVT | 0.0697219635020756 |
| NATIO1_NL | 0.003379826184942647 |
| LJR_INS_1E | -0.040897797661913926 |
| PC4_GELDERLAND | -0.10749952849678446 |
| GEBJAAR | 0.37063054256499045 |
| PC4_GRONINGEN | -0.07108952848273098 |
| MND_TOT_VEST_NL | 0.011479978085094877 |

Table 19: Linear regression coefficients for model Random3.

| Variable | Coefficient |
|---|---|
| PC4_OVERIJSSEL | -0.09086904732907242 |
| PC4_GRONINGEN | -0.006758712030285996 |
| GESLACHT_M | 1.7396437311187287 |
| GEWICHT_0.3 | -0.45304813636848174 |
| PC4_NOORDHOLLAND | -0.10025012740978845 |
| TYPE_PO_SBO | -0.22534360671929754 |
| DENOMINATIE_REF_GEV | 0.1638400448997928 |
| VBJ_SBO | -0.003414908233946523 |
| GENERATIE | 0.26354348814820383 |
| PC4_DRENTHE | -0.019339414332652516 |
| ADVIES_VO_VWO | 2.78739088009457 |
| VOORS_MND | -0.04018888173620985 |
| AFSTAND | -0.005023317161229834 |
| ACHTERGR_2 | -0.09901324593853489 |
| SCHOOLGROOTTE | 0.14162799485992705 |
| NATIO2_NVT | 0.04707051161180503 |
| ADVIES_VO_VMBOB | -3.801194507198745 |
| DENOMINATIE_ISL | 0.06285614752788124 |
| LAND_GEB_NL | -0.053123744463435596 |
| GEWICHT_0.4 | -0.015421555622908395 |
| DENOMINATIE_RK | 0.19948352093805855 |
| LEERJAAR | 0.0029063905133042255 |
| GESLACHT_V | 1.534645973104479 |
| PC4_ZEELAND | -0.020666493207042447 |
| LAND_OUDER1_NL | 0.27660868588541887 |

| | |
|---|---|
| ADVIES_VO_HAVOVWO | 1.2615316190517114 |
| VBJ_INS | -0.11761287422283773 |
| PC4_LIMBURG | 0.11731527450751511 |
| LFT_TOETS | -0.4916504554697122 |
| ADVIES_VO_VMBOK | -2.8974532775860595 |
| VBJ_VSO | -0.004271478676402828 |
| GEBJAAR | 0.5077478879566946 |
| MND_TOT_VEST_NL | -0.01783221618615205 |
| PC4_FLEVOLAND | -0.135197933135308 |
| NATIO1_NL | -0.011875147145861131 |
| GEWICHT_1.2 | -0.35663498583747233 |
| LAND_OUDER2_NL | 0.42287613183002626 |
| ADVIES_VO_PRO | -1.3472002943509747 |
| ADVIES_VO_VMBOB_K | -2.5281377662807225 |
| GROEPSGR | 0.06877417244155448 |
| NVS | -0.2727876536079041 |
| ADVIES_VO_VMBOGT | -2.1815657620564792 |
| GEWICHT_0.7 | -0.011648192300141805 |
| GEWICHT_0.9 | -0.04459333954226721 |
| DENOMINATIE_OPB_ABZ | 0.03963531301321588 |
| ADVIES_VO_VMBOK_T | -1.4077315510132917 |
| VBJ_SO | 0.005038485860736797 |
| DENOMINATIE_PC | 0.12221826796798196 |
| PC4_ZUIDHOLLAND | -0.26947695464242905 |
| VROEG_MND | -0.018243624124389547 |

Table 20: Linear regression coefficients for model Random4.

| Variable | Coefficient |
|---|---|
| VOORS_MND | -0.042592079335691986 |
| GROEPSGR | 0.12133873551084562 |
| ADVIES_VO_VMBOB | -3.948026536569052 |
| GESLACHT_V | 1.1940333333673983 |
| ADVIES_VO_VWO | 2.556761575511462 |
| LEERJAAR | 0.021454649825315864 |
| ADVIES_VO_VMBOGT_HAVO | -0.7687988071617389 |
| GEWICHT_0.7 | -0.027008272284896773 |
| PC4_UTRECHT | -0.10421337213116866 |
| NVS | -0.21945664190306635 |
| ADVIES_VO_VMBOK_T | -1.5097255023178069 |
| AFSTAND | -0.0015945099793528983 |
| LAND_OUDER2_NL | 0.6185747025680697 |
| PC4_DRENTHE | -0.0995021868700087 |
| PC4_ZUIDHOLLAND | -0.4573984420201884 |
| GEWICHT_0.25 | -0.05733453389063459 |
| ADVIES_VO_VMBOB_K | -2.6525427387591187 |
| VBJ_VSO | -0.005339230807116145 |
| VBJ_INS | -0.04709021222603371 |
| PC4_NOORDBRABANT | -0.09913237609857617 |

| | |
|---|---|
| PC4_OVERIJSSEL | -0.1953349378423198 |
| PC4_NOORDHOLLAND | -0.24870001295841704 |
| ADVIES_VO_HAVOVWO | 1.0610151469889568 |
| PC4_FRIESLAND | -0.14858823038768923 |
| DENOMINATIE_PC | 0.10224587584259193 |
| DENOMINATIE_OPB_ABZ | 0.004837212882121467 |
| LAND_OUDER1_NL | 0.48766349566896094 |
| GEWICHT_0.9 | -0.01953145481818591 |
| VBJ_BO | -0.0658066256160908 |
| ADVIES_VO_VMBOK | -3.0602367314635437 |
| LAND_GEB_NL | -0.10008558999557376 |
| PC4_ZEELAND | -0.08415639008467045 |
| VBJ_SBO | -0.2850538204312446 |
| DENOMINATIE_ISL | 0.036629485361193814 |
| NATIO2_NVT | 0.12866736188519146 |
| LFT_TOETS | -0.578849847303466 |
| GEBJAAR | 0.45626945080414716 |
| PC4_GRONINGEN | -0.11656049984532688 |
| GEWICHT_0.3 | -0.4542891797125169 |
| NATIO1_NL | 0.07287331000533287 |
| GEWICHT_0.4 | -0.019180471383450852 |
| DENOMINATIE_REF_GEV | 0.14097764019840858 |
| PC4_GELDERLAND | -0.23439516936285817 |
| ADVIES_VO_VMBOGT | -2.491780688070671 |
| PC4_FLEVOLAND | -0.21370874482731322 |
| GESLACHT_M | 1.401289745258078 |
| MND_TOT_VEST_NL | -0.023253038496793887 |
| VROEG_MND | -0.01932649871647585 |
| GENERATIE | 0.43670214385686035 |
| DENOMINATIE_RK | 0.14999694660318874 |

Table 21: Linear regression coefficients for model Random5.

# D Coefficients and p-values for model without GESLACHT_V (chapter 8)

| Variable | pvalue | coefficient |
|---|---|---|
| ADVIES_VO_VWO | 0.0 | 6.8466611168245155 |
| ADVIES_VO_HAVOVWO | 0.0 | 4.740396605769574 |
| LJR_INS_1E | 0.0 | -1.6825919145582025 |
| GEWICHT_0.3 | 0.0 | -0.43843429373536896 |
| GEBJAAR | 0.0 | 0.4515218228686162 |
| LFT_TOETS | 8.970825752375126e-285 | -0.3567175069034075 |
| ADVIES_VO_HAVO | 4.918685153993545e-283 | 4.468410382187198 |
| GEWICHT_1.2 | 2.7232340852882658e-281 | -0.3466088561638398 |
| VBJ_INS | 1.0973266212053451e-259 | -1.7393164697139856 |
| ADVIES_VO_PRO | 6.202025858807428e-198 | -0.6687633067154511 |
| ADVIES_VO_VMBOGT_HAVO | 4.178187974728961e-152 | 2.546924614540628 |
| LAND_OUDER2_NL | 2.998100846750077e-150 | 0.4168540093316959 |
| GESLACHT_M | 1.6404199780183158e-144 | 0.20667311234412783 |
| ADVIES_VO_VMBOGT | 1.5497305560805932e-91 | 3.608779917729614 |
| ADVIES_VO_VMBOB | 9.74225388637001e-65 | -1.3057107515095012 |
| LAND_OUDER1_NL | 4.607593273980661e-56 | 0.27419667511674994 |
| TYPE_PO_SBO | 2.9188749760880193e-40 | -0.21515741559279433 |
| SCHOOLGROOTTE | 9.651819879527008e-40 | 0.11487450678072869 |
| GENERATIE | 2.4414797639985057e-33 | 0.26071338737070215 |
| NVS | 1.0163376381258756e-30 | -0.22172482196270926 |
| DENOMINATIE_REF_GEV | 9.273907187637389e-24 | 0.14983403894338387 |
| ACHTERGR_2 | 1.3275280575511887e-14 | -0.1795309601232118 |
| ADVIES_VO_VMBOK_T | 2.08352958914772e-12 | 0.39268666946082176 |
| GROEPSGR | 7.151272770290935e-12 | 0.05847886728883858 |
| DENOMINATIE_ISL | 1.4928994624313505e-09 | 0.05927834339345184 |
| LEERJAAR | 5.135554273993016e-09 | 0.047662827375077044 |
| GEWICHT_0.9 | 2.8244969748868948e-08 | -0.0464110962978162 |
| ADVIES_VO_VMBOB_K | 2.526961914555498e-07 | -0.3471950331307563 |
| GEWICHT_0.25 | 7.101961646235946e-07 | -0.04077521685238028 |
| LAND_GEB_NL | 7.624053057628753e-07 | -0.05296126870482054 |
| VOORS_MND | 1.2362786097704342e-06 | -0.039686517664626275 |
| DENOMINATIE_RK | 7.366456454702254e-06 | 0.1612793604392594 |
| PC4_LIMBURG | 5.098285435812925e-05 | 0.18917482559550425 |
| ACHTERGR_1 | 6.320985332283981e-05 | -0.08619337004175609 |
| NATIO2_NVT | 9.512576670197061e-05 | 0.044055794705095436 |
| PC4_FLEVOLAND | 0.00011383660631130245 | -0.1140830789166225 |
| DENOMINATIE_PC | 0.00032772560305848516 | 0.11501664047082516 |
| MND_TOT_VEST_NL | 0.011590513832469859 | -0.026354506550249998 |
| PC4_ZUIDHOLLAND | 0.021237142132564557 | -0.1719522090372676 |
| PC4_NOORDBRABANT | 0.028594433989139936 | 0.14975742014049542 |
| VROEG_MND | 0.05881305813816816 | -0.01546566692835874 |
| GEWICHT_0.4 | 0.06070967888340735 | -0.015240541975399813 |
| PC4_UTRECHT | 0.09140401057947614 | 0.08614069495049181 |
| GEWICHT_0.7 | 0.1437526182479075 5 | -0.011409819712233815 |

| | | |
|---|---|---|
| NATIO1_NL | 0.266212687839906 | -0.011135647082526068 |
| PC4_GRONINGEN | 0.3784473952384382 | 0.03002364722361006 |
| PC4_ZEELAND | 0.41955276642643324 | 0.021971092858333763 |
| VBJ_VSO | 0.427155567654096 | -0.00576319617405166 |
| DENOMINATIE_OPB_ABZ | 0.43701889146296125 | 0.027418488607287322 |
| PC4_DRENTHE | 0.5481670734981681 | 0.018325126228364685 |
| VBJ_BO | 0.579442375466369 | 0.010888916669109905 |
| PC4_NOORDHOLLAND | 0.6814872972060336 | -0.0237356594807685 |
| PC4_OVERIJSSEL | 0.7174633943573556 | -0.017027107867774616 |
| ADVIES_VO_VMBOK | 0.7186546927197464 | -0.03178073582880203 |
| AFSTAND | 0.7188066587322042 | -0.002970949778757337 |
| PC4_GELDERLAND | 0.7239504197621367 | 0.022544914164785546 |
| PC4_FRIESLAND | 0.7567999167741648 | -0.010775420296987082 |
| VBJ_SO | 0.7696027909067265 | 0.0024109515338955334 |
| VBJ_SBO | 0.7833612522597937 | -0.004539612205708199 |

Table 22: p-values and linear regression coefficients for model without variable **GESLACHT_V**.

# References


[1] Abdi, H., and Williams, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics 2*, 4 (2010), 433–459.

[2] Alkharusi, H. Categorical variables in regression analysis: A comparison of dummy and effect coding. *International Journal of Education 4*, 2 (2012), 202.

[3] Bramer, M. *Principles of data mining*, vol. 180. Springer-Verlag London, 2007.

[4] Brownlee, J. Why one-hot encode data in machine learning? https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/, 2017.

[5] Centraal Bureau voor de Statistiek. Basisscholen en denominatie. https://www.cbs.nl/nl-nl/achtergrond/2017/47/basisscholen-en-denominatie, 2017.

[6] de Pagter, B., Groenevelt, W., and Janssens, B. Lecture notes in analysis 2 (tw1070), 2019.

[7] Den Blanken, M., and Van der Vegt, A. L. *Voorschoolse educatie: waarom wel, waarom niet?* Sardes, Utrecht, 2007.

[8] Franco-Mariscal, A.-J., Oliva, J., and Gil, M. L. A. Students' perceptions about the use of educational games as a tool for teaching the periodic table of elements at the high school level. *Journal of Chemical Education 92* (02 2015), 278–285.

[9] Government of the Netherlands. Secondary education. https://www.government.nl/topics/secondary-education.

[10] Heumann, C., Schomaker, M., and Shalabh. Linear regression. In *Introduction to Statistics and Data Analysis : With Exercises, Solutions and Applications in R*. Springer International Publishing, Cham, 2016, pp. 249–295.

[11] Heumann, C., Schomaker, M., and Shalabh. Hypothesis testing. In *Introduction to Statistics and Data Analysis*. Springer International Publishing, Cham, 2017, pp. 209–247.

[12] Igual, L., and Seguí, S. Regression analysis. In *Introduction to Data Science*. Springer, Cham, 2017, pp. 97–114.

[13] Jolliffe, I. T. *Principal Component Analysis*, second ed. Springer Series in Statistics. Springer-Verlag New York, 2002.

[14] Onderwijs in Cijfers. Schooladvies en heroverweging schooladvies. https://www.onderwijsincijfers.nl/kengetallen/po/leerlingen-po/prestaties-schooladvies.

[15] Rijksoverheid. Toelating voortgezet onderwijs gebaseerd op definitief schooladvies. https://www.rijksoverheid.nl/onderwerpen/schooladvies-en-eindtoets-basisschool/toelating-voortgezet-onderwijs-gebaseerd-op-definitief-schooladvies.