02402 - Introduction to Statistics

Project 2: BMI survey analysis

Author: Kornel Kowalczyk (s202401)

Overview:

The following report presents a descriptive and statistical analysis of BMI survey data collected among Danish population. First part describes results of a survey paper, giving explanations of used attributes and providing basic summary statistics of the dataset. Second part of the report focuses on statistical analysis including the proposal of a linear model for BMI score prediction as well as model validation and reduction.

Table of contents

1. Descriptive analysis

The content of the dataset is a result of a BMI survey carried out on a sample size of 847 people. From the following survey 3 attributes *(fig. 1)* were extracted which are used for statistical analysis and model development for BMI score prediction.

Variable	Variable type	Description
BMI	Quantitative	The respondent's BMI score.
Age	Quantitative	The respondent's age in years.
Fast food	Quantitative	The frequency of respondent eating fast food: **0** - Never **1.0** - Less than 1 time per year **6.0** - 1-11 times per year **24.0** - 1-3 times per month **78.2** - 1-2 times per week **182** - 3-4 times per week **286.7** - 5-6 times per week **365** - Every day

Fig. 1. Attributes overview.

BMI values are log-transformed for the purpose of further analysis and a logBMI variable is added to the dataset. Data points are distributed as shown on the histograms (*fig. 2).* LogBMI is normally distributed while age follows uniform distribution. Fastfood variable is most likely distributed exponentially, while most of the data points are clumped up on the left side of the histogram, very little information is included about consumers who eat fast food more frequently.

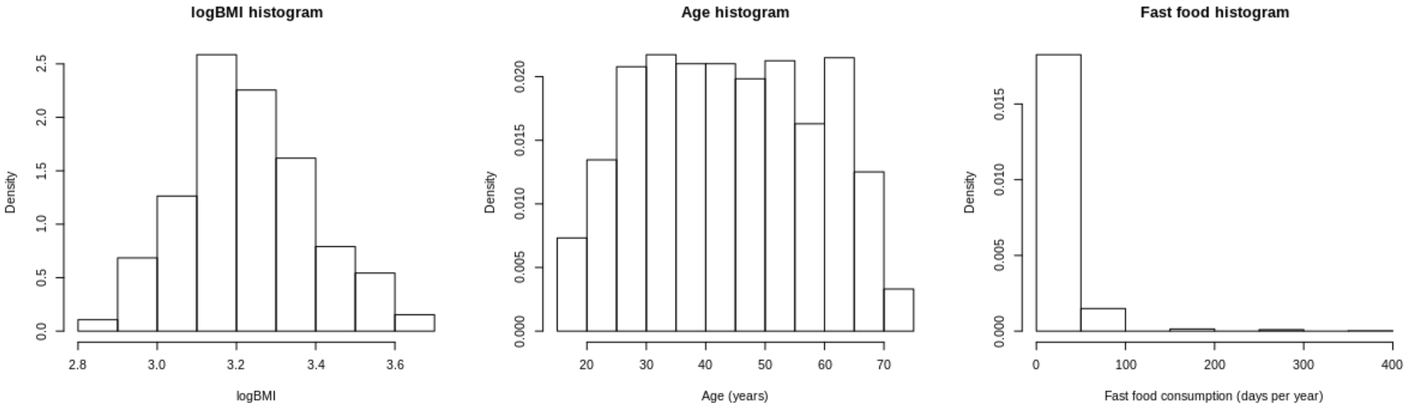

Fig. 2. Histograms of dataset variables.

Scatter plots (*fig. 3)* are representing logarithmically-transformed BMI scores plotted against age and fastfood variables. As can be seen on the left plot, values are covering the area more thoroughly while on the right plot, data points are not covering the middle and the right side of the scatter plot sufficiently.

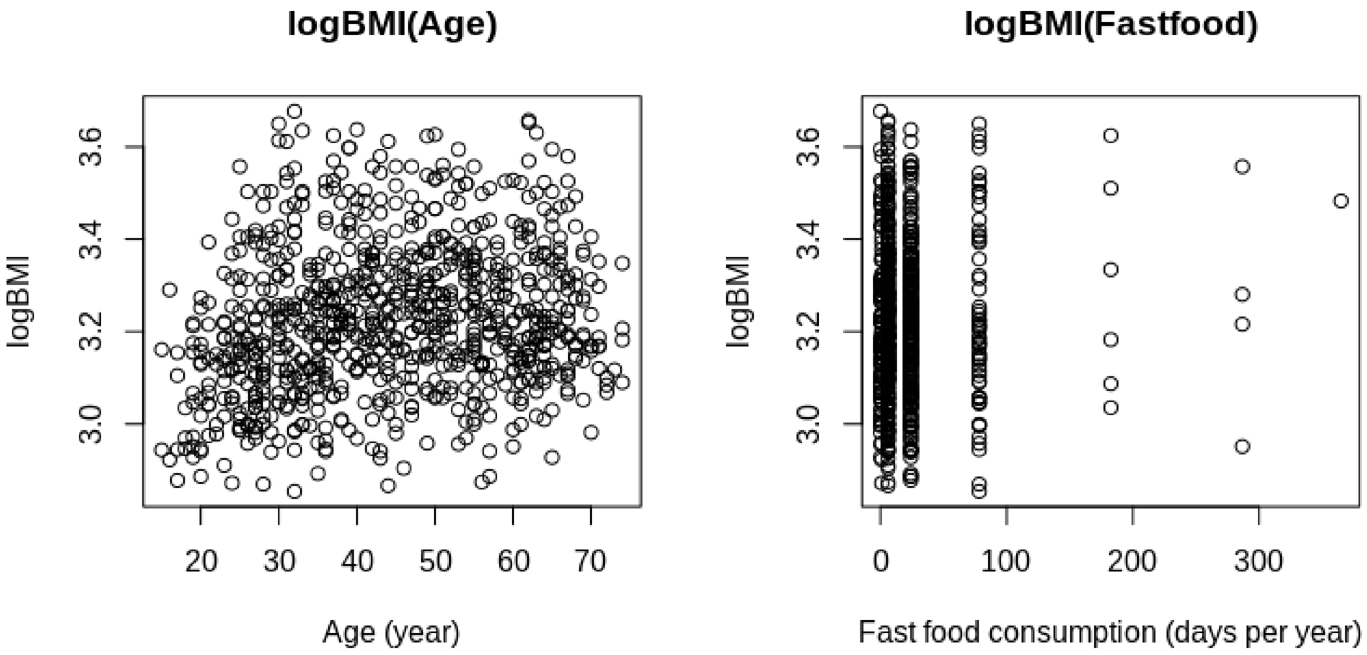

Fig. 3. Scatter plots of logBMi against other variables.

There is only information for certain discrete values which is dictated by the structure of a survey as can be seen on attributes overview (*fig. 1)*. As fast food consumption values are not continuous but rather wide intervals recoded into discrete values. This leads to information loss which may corrupt further model accuracy. Boxplots below (*fig. 4)* are further visualizing the problem of data distribution within fast food consumption variable, as the median is equal to 6 and upper quartile is equal to 24 which are extremely low values considering the whole range for this variable.

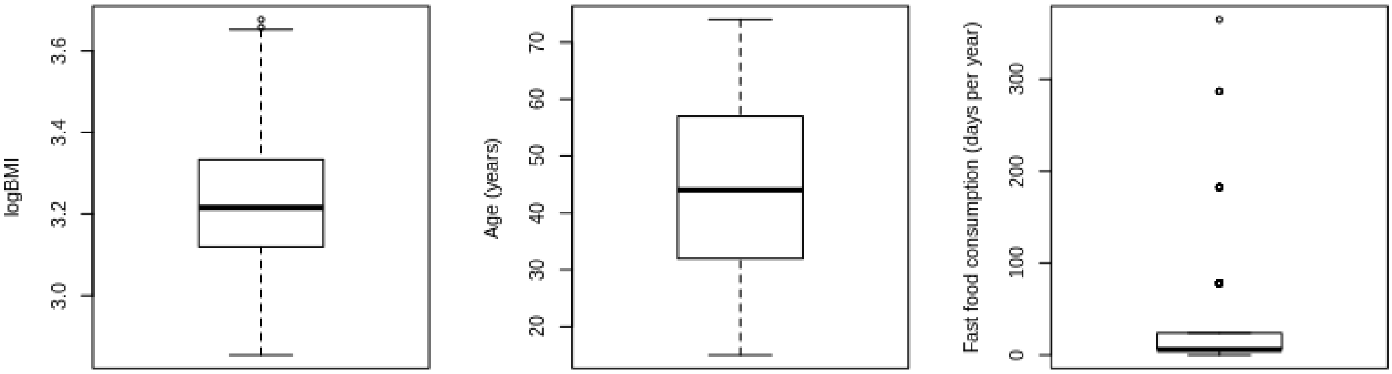

Fig. 4. Boxplots of dataset variables, from left logBMI, Age, Fastfood.

The following table (*fig. 5)* represents summary statistics for all variables in the dataset used in the further modelling.

Variable:	***Number of obs.***	***Sample mean***	***Sample variance***	***Sample std. dev.***	***Lower quartile***	***Median***	***Upper quartile***
	n	$(\bar{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
logBMI	847	3.23	0.03	0.16	3.12	3.22	3.33
Age	847	44.62	211.20	14.53	32.00	44.00	57.00
Fastfood	847	19.04	1066.10	32.65	6.00	6.00	24.00

Fig. 5. Summary statistics.

2. Statistical analysis

Designating logBMI as outcome variable while age and fast-food consumption as the explanatory variables the following linear model is proposed:

$$Y_{\log BMI} = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

it is assumed that the residuals are independent and identically distributed normal random variables with zero mean and unknown constant variance.

Parameters estimation

Using the R script, following model parameters are computed and whose values are presented in the table below (*fig. 6*).

Parameter	***Parameter value***	***Standard deviation***
$\widehat{\beta}_0$	3.1124	0.0193
$\widehat{\beta}_1$	0.0024	0.0004
$\widehat{\beta}_2$	0.0005	0.0002
$\widehat{\sigma}^2$	0.0250	-
R^2	0.0450	-
Degrees of freedom	837	-

Fig. 6. Linear model parameters.

From the model regression coefficients it can be concluded that with increasing age by 1 year the logBMI value is growing by 0.0024, and by changing consumption of fast food by 1 day per year the logBMI is proportionally changing by 0.0005. As in model creation 840 observations were used as well as 2 explanatory variables, degrees of freedom are equal to 837.

Model validation

To provide model validation the following plots (*fig. 7*) are presented. It can be observed that residuals are not having any systematic dependence.

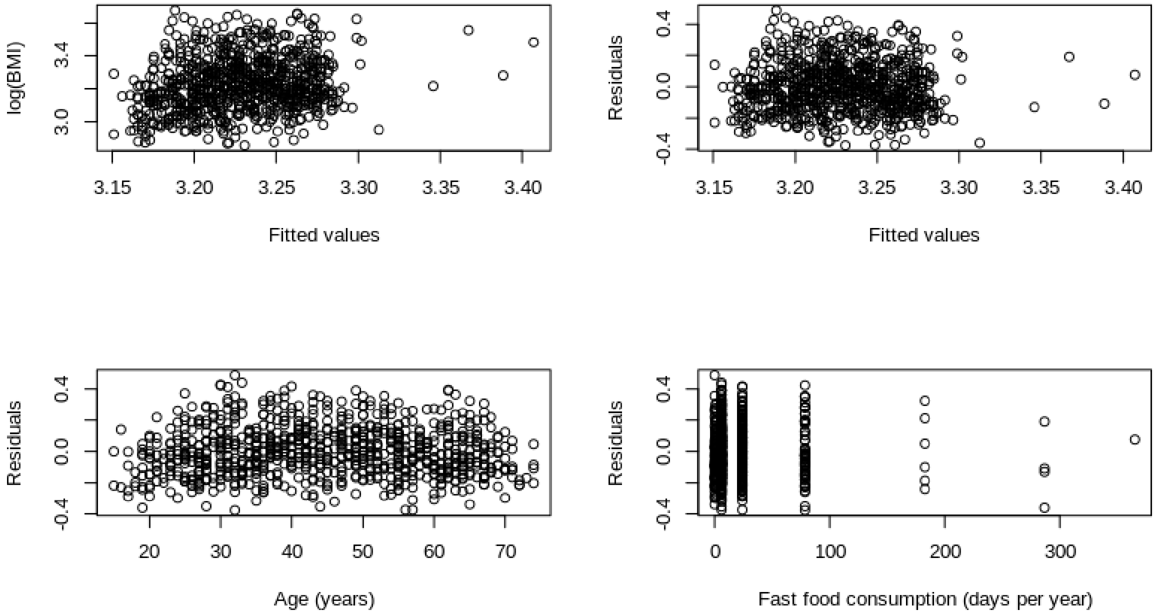

Fig. 8. Model validation plots

To check normality assumption QQ-plot is shown (*fig. 9*).

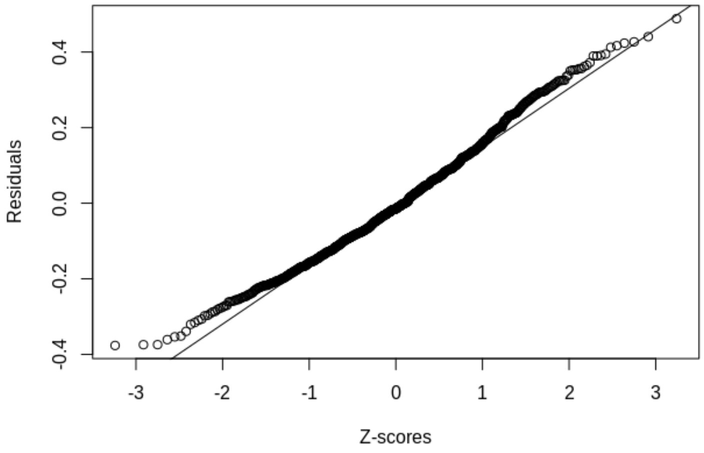

Fig. 9. QQ-plot of the residuals

As residuals are lying close to the qq-line it leads to the conclusion that residuals are identically distributed normal random variables as was assumed at the beginning.

Confidence Interval

To estimate range of plausible values of our model parameters, 95 % confidence interval for $\widehat{\beta}_1$ is calculated using following formula:

$$\widehat{\beta}_1 \pm t_{1-\alpha/2} \widehat{\sigma}_{\beta_1} = 0.0024 \pm 1.96 \cdot 0.0004$$

$$0.0024 \pm 0.0076 = [0.0016,\ 0.0031]$$

Confidence intervals for all of the regression coefficients are presented in the table below (*fig. 10*).

	2.5%	97.5%
$\widehat{\beta}_0$	3.0744	3.1504
$\widehat{\beta}_1$	0.0016	0.0031
$\widehat{\beta}_2$	0.0002	0.0009

Fig. 10. Confidence intervals for model parameters

Hypothesis testing

To investigate if β_1 is equal to *0.001* the following hypothesis test is provided:

$$H_0: \beta_1 = 0.001$$

$$H_1: \beta_1 \neq 0.001$$

Significance level $\alpha = 0.05$ is chosen and test statistic is calculated using following formula:

$$t_{obs,\beta_1} = \frac{\widehat{\beta}_1 - \beta_{0,1}}{\widehat{\sigma}_{\beta_1}} = \frac{0.0024 - 0.001}{0.0004} = 3.53$$

next, using t_{obs} the *p-value* is calculated with degrees of freedom equal to n-3 which is 837:

$$p - value_1 = 2 \cdot P\left(T > \left| t_{obs,\beta_1} \right| \right) = 0.0004$$

As calculated *p-value* is lower than significance level $\alpha = 0.05$ The null hypothesis is rejected. According to this it is statistically not plausible to say that the regression coefficient for age is equal to 0.001.

Backward selection

To investigate if the model can be reduced, p-values of regression coefficients need to be compared (*fig. 11*) in order to designate the first variable to remove. As the highest p-value is for fast food consumption, this variable is chosen to be removed.

Parameter	Test statistic	p-value
$\widehat{\beta}_0$	160.835	$2\cdot 10^{-16}$
$\widehat{\beta}_1$	6.104	$1.58\cdot 10^{-9}$
$\widehat{\beta}_2$	3.119	$1.88\cdot 10^{-3}$

Fig. 11. Validation table for current model.

Therefore test statistic and p-values are calculated for the reduced model (*fig. 12*). As the p-value for age regression coefficient had increased after model reduction it can be concluded that the model presented at the beginning couldn’t be reduced.

Parameter	Test statistic	p-value
$\widehat{\beta}_0$	178.327	$2\cdot 10^{-16}$
$\widehat{\beta}_1$	5.412	$8.15\cdot 10^{-8}$

Fig. 12. Validation table for reduced model

According to these results the final model is as follows:

$$Y_{\log BMI} = 3.1124 + 0.0024x_{1,i} + 0.0005x_{2,i} + \varepsilon_i$$
$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

Prediction interval

The final model prediction intervals are presented in following table (*fig. 13*):

Id	Measured value	Predicted value	Lower bound of PI	Upper bound of PI
841	3.143436	3.236993	2.927972	3.546015
842	3.269232	3.210875	2.901802	3.519949
843	3.269438	3.232245	2.923231	3.541258
844	3.324205	3.232245	2.923231	3.541258
845	3.106536	3.229870	2.920857	3.538883
846	3.263822	3.229641	2.920601	3.538681
847	3.058533	3.211670	2.901898	3.521443

Fig. 13. Prediction intervals for test data.

LogBMI scores for all of the test dataset elements are within 95 % prediction interval of the final model. What needs to be pointed is that this interval is covering wide range of BMI score, for example for observation with ID 842 the prediction interval is from 18.2 to 33.8 therefore according to WHO classification of BMI scores (*fig. 14*) such prediction interval gives us no useful information as person’s BMI score falls between underweight as well as class I obesity.

BMI score	**Assessment**
Less than 18.5	The person is underweight
Between 18.5 and 25	The person’s weight is normal
Between 25 and 30	The person is moderately overweight
Between 30 and 35	The person is severely overweight (Obesity Class I)
Between 35 and 40	The person is severely overweight (Severe Obesity Class II)
Above 40	The person is severely overweight (Extremely severe obesity Class III)

Fig. 14. The WHO classification of BMI scores

Predicted BMI scores are somewhat accurate therefore models may be used for coarse assessment of persons weight, however it is not recommended to rely completely on this model. Such outcome may be a result of not the best selection of explanatory variables or caused by small sample size. Also as was mentioned before, fast food consumption observations are distributed very unevenly within possible range while also being discretized with high loss of information. Therefore it may have a strong impact on model effectiveness. To overcome these problems it is recommended to increase sample size, reconsider selection of explanatory variables - maybe by adding some referring to physical activity - as well as make sure that all observations are well distributed within possible range with little information loss during data acquisition process.