BMI Survey 2

02402 Introduction to statistics

Author:

M V A Suhas Kumar s191382

November 12, 2019

DTU

Technical
University of
Denmark

Contents

List of Figures

List of Tables

1 Introduction

Overweight is a very common health-issue, and can have great impac on a person's well bei n gand general health. Body mass index (BMI) is one way of relatin g a person's weight to thei rheight and it is a common measurement of human body weight.

In this assignment we will continue to analyse the data from BMI survey. Unlike the last time this time we use Multipe linear regression and check the results using p and t tests.This report will seek to statistically describ e BMI from a sample of 847 persons.

2 Question A

2.1 Question

Present a short descriptive analysis and summary of the data for the variables logbmi, age, and fastfood. Include scatter plots of the log-transformed BMI scores against the two other variables, as well as histograms and box plots of all three variables. Present a table containing summary statistics, which includes the number of observations, and the sample mean, standard deviation, median, and 0.25 and 0.75 quantiles for each variable.

2.2 Solution

Description of Data Set and Variables

A person's BMI (Body Mass Index) score is a measure of the person's overweight which depends on various factors like gender, age ,education . In order to analyse BMI this dataset is collected which includes features like:

- **Gender** : which corresponds to gender of respondant
- **height** : which corresponds to height of respondant in cm
- **weight** : which corresponds to weight of respondant in kgs
- **Fast Food level** : which corresponds to no of days respondant eats fast food in a year

scatter plots

These are the plots genarated using the inbuilt R script function

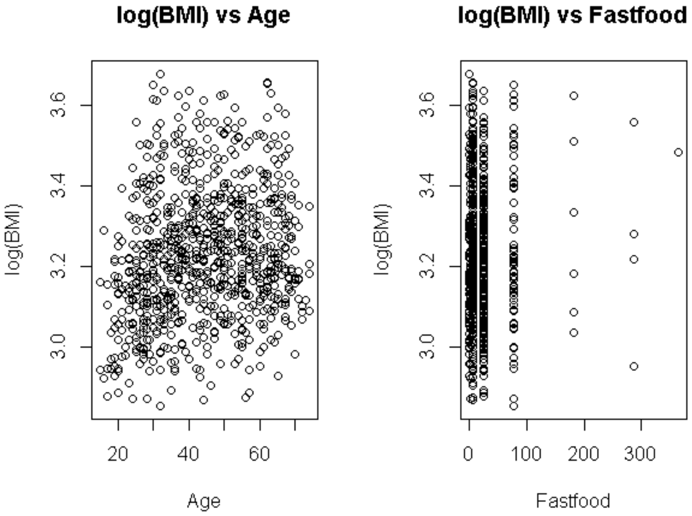

Figure 1: Scatter Plots of log transformed BMI data vs other variables

From above Figure it is seen that the observations are very spread and there seems to be no clear correlation between the log-transfored BMI and age. here also does not seem to be a clear link between the log-transformed BMI and fastfood intake.

Histograms

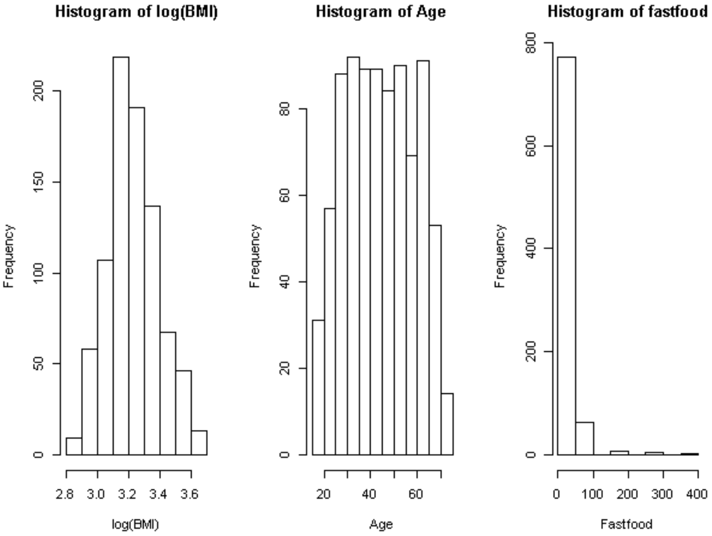

Figure 2: Histograms of the three variables

From above figure we could see that log transformed BMI is normally distributed. Age distribution has no inherent distribution which is truly fair and un-biased survey. Fast food distribution is scattered with peak near zero and very high values.

Box plots

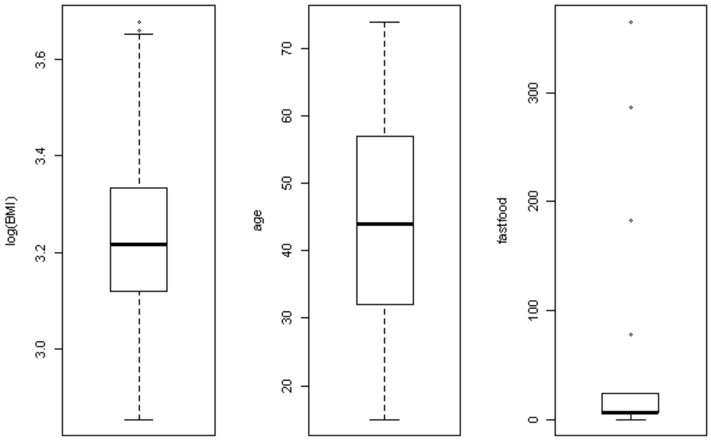

Figure 3: Box plots of the three variables

Log transformed BMI has only 2 outliers and centered mean, No outliers in age distribution.For the fast food we could see that mean is lower quartile which indicates the survey is collected from people who eat very fast food or very high fast food which are indicated in ouliers.

Summary Statistics

The Below results are obtained from the R-script.

	age	**fastfood**	**logbmi**
Mean	44.62	19.04	3.228
Standard Deviation	14.5328	32.65124	0.1603723
Median	44.00	6.00	3.216
Q1	32.00	6.00	3.120
Q3	57.00	24.00	3.334

Provides summary statistics more precisely than the gures, seeing as the exact values

3 Question B

3.1 Question

Formulate a multiple linear regression model with the log-transformed BMI scores as the dependent/outcome variable (Y_i), and age and fast-food consumption as the independent/explanatory variables ($x_{1,i}$ and $x_{2,i}$, respectively). Remember to state the model assumptions. (See Equation (6-1) and Example 6.1).

3.2 Solution

Here we take the target variable or output variable as the logarithmic of BMI(Y_i) and input variables are age and fast food consumption as ($x_{1,i}$ and $x_{2,i}$ respectively. Also we take β_0, β_1, β_2 as the weights of the model/parameters where β_0 as constant term ,β_1 as coefficient of $x_{1,i}$ and β_2 as coefficient of $x_{2,i}$.

Formulating the Multiple linear regression

$$Y_i = \beta_0 + \beta_1 * x_{1,i} + \beta_2 * x_{2,i} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2) \tag{1}$$

Here we assume that (ϵ_i) are independent normal random variables with the mean as zero and fixed variance (σ).

4 Question C

4.1 Question

Estimate the parameters of the model. These consist of the regression coefficients, which we denote byβ_0, β_1, β_2, and the variance of the residuals, σ^2 . Give an interpretation of the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, explaining what they tell us about the relation between the log-transformed BMI scores and the model's expla natory variables. (See Remark 6.14). Furthermore, present the estimated standard deviations of$\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$,2 , the degrees of freedom used for the estimated residual variance σ^2, and the explained variation, R^2.

4.2 Solution

The inbuilt R script code $fit < -lm(logbmi\ age + fastfood, data = D_model)$ is used to estimate the parameters of the model. summaray(fit) yields the parameters of the model.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
$\hat{\beta_0}$ (Constant)	3.1124	0.0194	160.84	0.0000
$\hat{\beta_1}$ (Age coefficient)	0.0024	0.0004	6.10	0.0000
$\hat{\beta_2}$ (Fastfood coefficient)	0.0005	0.0002	3.12	0.0019

Table 1: Estimated Parameters for Multiple linear Regression

We can find the residual variance using the R command $var(fit\$residuals)$. We get the value of the **variance** to be **0.02469646**. Also from the above table last column we could get result of the t statistics with the hypos thesis being $H_{0,i} : \beta_0 = 0$ performed on each of the estimated coefficient .The last column also gives the p values from which we could see that Age has a very relation with log(BMI) (since p <0.001) and also fast food has a decent relation (since 0.01>p>0.0001).

The summary yields the result of β_0 **= 3.1124** with **variance of 0.01942** ,β_1**= 0.0024** with a **variance of 0.00042**, and β_2**= 0.0005** with a **variance of 0.00022**.It should seen that standard deviation is fairly low for all the variables which implies we got a good values.
DF $= n - (p + 1)$ where n is the amount of observations and p is the amount of variables not including the intercept.
DF $= 840 - (2 + 1) =$ **837**
With the obtained DF value we could get the residual variance as 0.1573 and explained variance to be 0.0449.

5 Question D

5.1 Question

Perform model validation with the purpose of assessing whether the model assumptions hold. Use the plots, which can be made using the R code below, as a starting point for your assessment. (See section 6.4 on residual analysis).

5.2 solution

The R commands are used to check the assumption wether residuals are normally distributed or not.The concept is to check with q-q plot.

$$qqnorm(fitresiduals, ylab = "Residuals", xlab = "Z - scores", vmain = "")$$

$$qqline(fitresiduals)$$

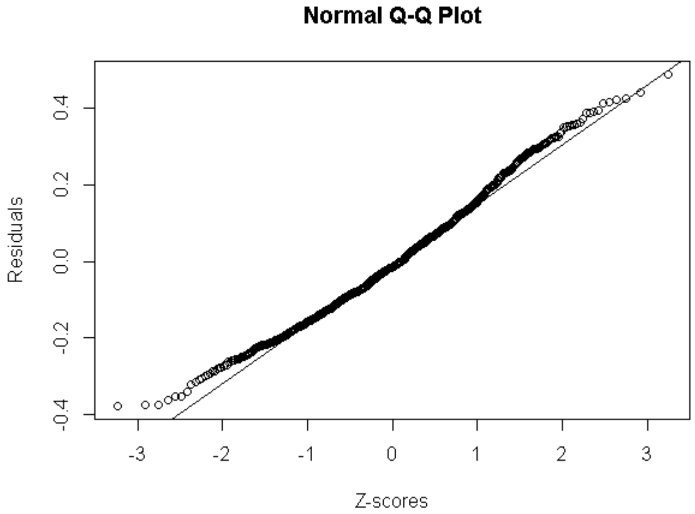

Figure 4: Normal Q-Q Plot

Clearly from the above plot we could see that our assumption is true.

System Behaviour plots

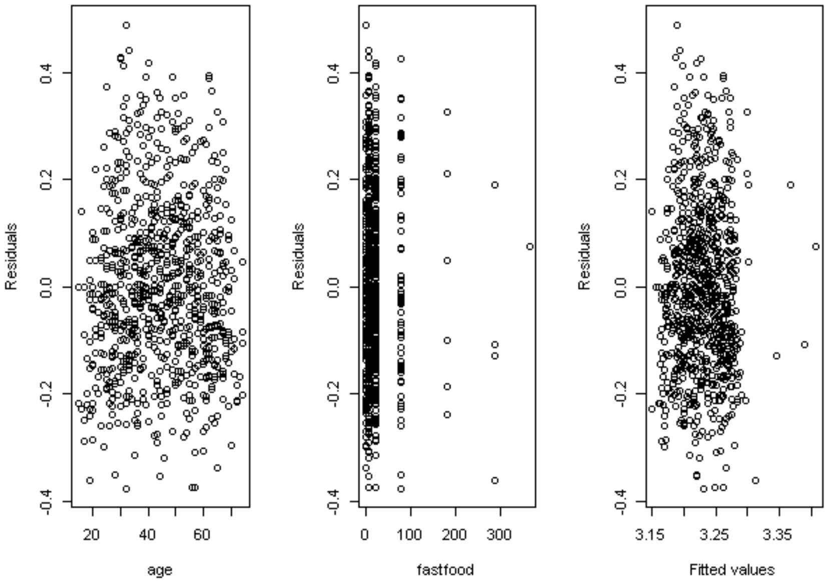

Figure 5: System Behaviour Plots

Considering the first figure we see that there is apparent distribution between residual values and Age. Again even from the middle figure we see that there is no apparent relation between residuals and Fast food,probably that is the reason why beta coefficient for the fast food is nearly zero and we could drop it as there is no real effect by removing it.From the residual vs fitted values plot, It shows that a large portion of the data set is below 3.3 on the scale of the tted values, with a few outliers. The residuals are spread out, but there seems to be no correlation between the residuals and the tted values.

6 Question E

6.1 Question

State the formula for a 95% confidence interval for the age coefficient, here deno- ted by β_1.(See Method 6.5). Insert numbers into the formula, and compute the confidence interval. Use the R code below to check your result, and to determine confidence intervals for the two other regression coefficients.

6.2 Solution

The formula for the 95% CI for the $(1 - \alpha)$ quartile would be

$$CI = \hat{\beta}_i + t_{1-\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}$$

where $t_{1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of a i-distribution with $n - (p + 1)$ degrees of freedom.

As found in Exercise C, the model estimate for the age coefficient is 0.0024 with a variance of 0.00042. The needed quantile (with DF 837) is 1.9628.

Calculating 95% CI around median where $(\alpha = 0.05)$ and also obtaining median from R script

$$CI = 0.0024 + 1.9728 * \frac{0.00042}{\sqrt{837}} \sim (0.0016, 0.0031)$$

The **95% confidence interval of mean** of the is **0.0016-0.0031**,meaning true mean will lie in this class with 95% probability.

By running R scripts for the remaining 2 parameters we get the following.

	2.5 %	**97.5 %**
$\hat{\beta}_0$ **(Intercept)**	3.0744463234	3.1504132672
$\hat{\beta}_1$ **(age)**	0.0016108861	0.0031378342
$\hat{\beta}_2$ **(fastfood)**	0.0002003159	0.0008803957

Table 2: Confidence Interval Regression Coefficients

We could clearly see that our manually calculation checks out with R script.

7 Question F

7.1 Question

It is of interest whether β_1 might be 0.001. Formulate the corresponding hypothesis. Use the significance level $\alpha = 0.05$. State the formula for the relevant test statistic (see Method 6.4), insert numbers, and compute the test statistic. State the distribution of the test statistic (including the degrees of freedom), compute the p-value, and write a conclusion.

7.2 Solution

A null hypothesis set up :

$$H_0 : \beta_1 = \beta_{0,1}$$

, where $\beta_{0,1} = 0.001$

$$H_1 : \beta_1 \neq \beta_{0,1}$$

. Firstly using t-statistics is used to calculate

$$t_{obs} = \frac{x - \mu_0}{\frac{s}{\sqrt{n}}}$$

$$t_{obs} = \frac{0.0023743602 - 0.001}{\frac{0.00042}{\sqrt{837}}} = 3.53306$$

Now calculating the P-value with knowning 837 degrees of freedom

$$p - value = 2 * pt(-abs(t_{obs}), df = n - 1) = 0.0004332699$$

Also significance level is set to $\alpha = 0.05$, clearly the P-value is less than α , therefore our initial assumption of null hypothesis is false and therefore we reject the null hypothesis. Hence $H_1 : \beta_1 \neq 0.001$

8 Question G

8.1 Question

Use backward selection to investigate whether the model can be reduced. (See Example 6.13). Remember to estimate the model again, if it can be reduced. State the final model, including estimates of its parameters.

8.2 Solution

From the p values for the coefficients obtained in the section C. We could see that P value for fast food coefficient is higher than for that of age coefficient.So we will drop the fast food parameter and check the results wether model can be reduced.
By running the R-script following are the results:

	Estimate	Std. Error	t value	Pr(>|t|)
$\hat{\beta_0}$ (Intercept)	3.1382	0.0176	178.33	0.0000
$\hat{\beta_1}$ (age)	0.0020	0.0004	5.41	0.0000

Table 3: Estimated Parameters for Reduced Model

We could see from the above table that the p values for the estimated coefficients still significant. Also we could also see that new estimated β_1 still lie in the range of standard deviation of the previous model. Hence ,this reduced model is good enough for prediction.

9 Question H

9.1 Question

Use your final model from the previous question as a starting point. Determine predictions and 95% prediction intervals for the log-transformed BMI scores, for each of the seven observations in the validation set (D test). See Example 6.8, Method 6.9 and the R code below. Compare the predictions to the observed log-BMI scores for the seven observations in the validation set and make an assessment of the prediction capabilities of the final model.

9.2 Solution

The 95% prediction intervals and predicted values for the last 7 observation calculated using the R-script are as follows

	id	logbmi	fit	lower	upper
841	841	3.143436	3.233456	2.922838	3.544073
842	842	3.269232	3.211150	2.900472	3.521828
843	843	3.269438	3.229400	2.918787	3.540013
844	844	3.324205	3.229400	2.918787	3.540013
845	845	3.106536	3.227372	2.916759	3.537986
846	846	3.263822	3.235483	2.924860	3.546106
847	847	3.058533	3.186817	2.875833	3.497801

Table 4: Prediction values and 95% prediction intervals

From the above table we could see that the predicted log(BMI) values are in the 95% range of true values.Hence our model performs well in 95% range.

The above statement could be observed quantitatively from below.

Observation No	841	842	843	844	845	846	847
% error	2.8637261	1.7766118	1.2246269	2.8519647	3.8897358	0.8682536	4.1942840

Table 5: Errors in predicted values.

We could see that maximum error in the observations is 4.1942840 which is less that 5%. Hence the model works well in 95% confidence range. Also we could see that the **mean error is 2.524712** which is practically acceptable. So,our predictions are very good and acceptable.