BMI Survey ||

02323 Introduction to statistics

Mohamad Malaz Mohamed ALZarrad
s180424

November 10, 2020

Small Introduction about our project

Let's us remember What Body mass index (BMI) is.
(BMI) is a value derived from the mass (weight) and height of a person.
As we have already explained in the first section of the project that overweight is a common health problem between different groups, it can have an impact on a person and their general health.
Here in this part of the assignment, I will continue working on analyzing the data from BMI .
Here also I will check the results of using p-value and t-tests and I will also use Multipe linear regression.
In this assignment, I have a statistically describe BMI from a sample of 847 persons I will work with.

Question a: 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.

Short Description of Data Set and Variables:

BMI (Body Mass Index) appears to be as strongly correlated with various metabolic and dis-ease, gender, education, fast food, age outcome
this dataset is collected which includes features like:

- **Gender**: gender of respondant man or woman.
- **height**: height of respondant in Centimeter.
- **weight**: weight of respondant in kilogram.
- **Fast Food level**: number of days respondant eats fast food in a year

scatter plots :

By using the inbuilt R script functions here is the plots generated

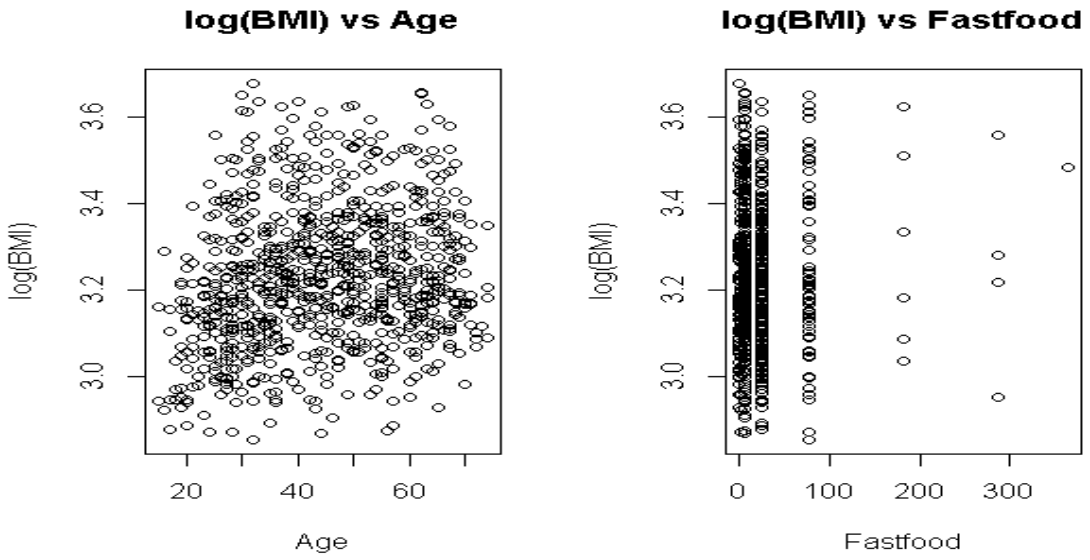

Figure1: Scatter Plots of log_transformed BMI data versus variables

We can see clearly from the plots shown above that the distribution of observations are spread i.e. “randomly scattered” and there are no clear correlation between age and the log-transformed BMI. Here too, there does not appear to be a clear link between the log-transformed BMI and Eating fast food.

Histograms :

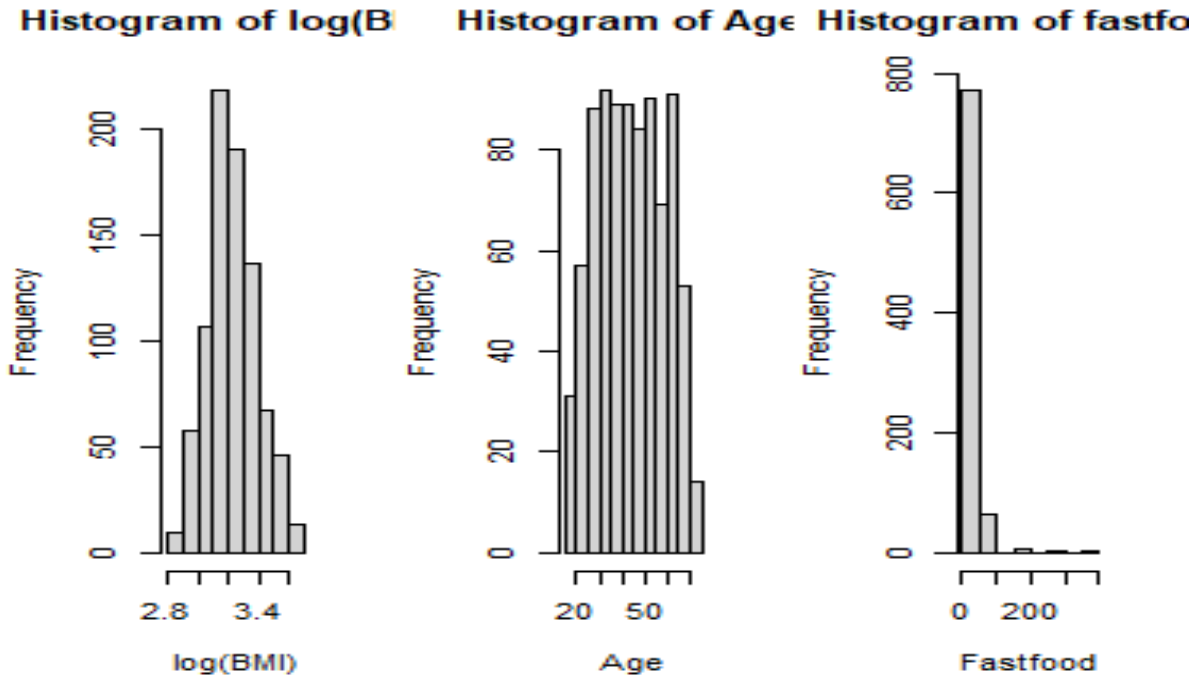

Figure 2: Histograms for the three variables

We can see from the figure shown above that the log transformed BMI is the only one that has a normal distribution. As for the age distribution has no inherent distribution which is unbiased survey and fair. As for the fast food distribution is randomly scattered with a top close to zero and high values.

Box plots :

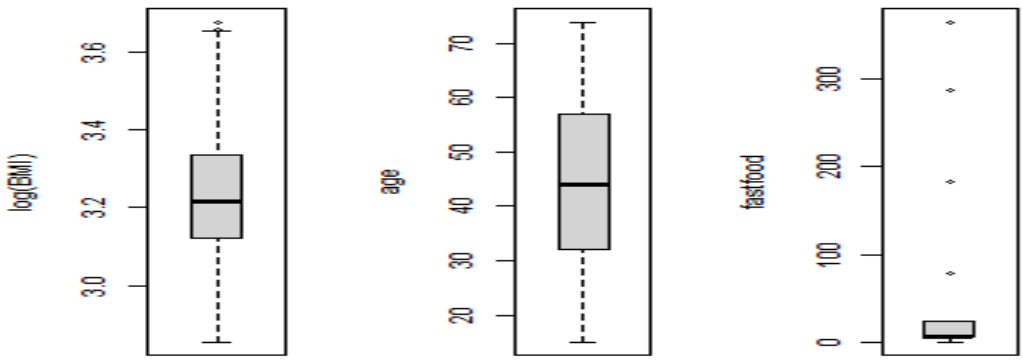

Figure 3: Box plots of the three variables

We can see from the figure shown above that there are two outliers and mean in center in the log transformed on the other hand there are NO outliers in the age distribution.
But we can see in the fast food that mean is located in the lower quartile which referenced the survey is collected from people who eat so much fast food which are referenced in outliers.

Summary Statistics :

By using the R-script I obtain the results in the table shown below:

	Fast-food	**age**	**Log-bmi**
Median	6.00	44.00	3.216
Standard Deviation	32.65124	14.5328	0.1603723
Mean	19.04	44.62	3.228
Q1	6.00	32.00	3.120
Q3	24.00	57.00	3.334

Table 1: Summary Statistics

Question b: 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).

Here we have to focus on output variable and take it as the logarithmic of BMI (Y_i) and our input variables are the consuming of fast food as (x_1,*i*) and age as (x_2,*i)*.
We have also to take β_0, β_1, β_2 as the weights of model
where →
β_0 is constant term
β_1 is coefficient of (x_1,*i*)
β_2 is coefficient of (x_2,*i)*

the Multiple linear regression :

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

(*εi*) are normal random variables with the mean equal zero and **(*σ)*** which is fixed variance.

Question c: Estimate the parameters of the model. These consist of the regression coefficients, which we denote by β_0, β_1, β_0, and the variance of the residuals σ^2,
The code in R:

fit < −lm(logbmi ~ age + fastfood, data = D_model)

I estimate the parameters of the model by using this code. The output of the summaray (fit) is parameters of the model.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
β_0 **(Constant)**	3.1124	0.0194	160.84	0.0000
β_1 **(Agecoefficient)**	0.0024	0.0004	6.10	0.0000
β_2 **(Fastfood coefficient)**	0.0005	0.0002	3.12	0.0019

Table 2: Estimated Parameters for Multiple linear Regression

By using the R command *var* (*fit*$*residuals*) we can find the residual variance where **variance** is **0.02469646**.

From the table we could get the result of the t statistics with the hypothesis $\mathbf{H_{0,i}: \beta_0 = 0}$ Conducted on each of the estimated coefficient. We can also see from column get the p_values and from these values we can that Age has a link with log (BMI)→($p<0.001$) and Fast food has also a relation → (since **0.01>p>0.0001**).

If we take a general look at our summaray We can see the results that it gave is

β_0 **=3.1124** with a **variance = 0.0194**^{2}

β_1 **= 0.0024** with a **variance = 0.0004**^{2}

β_2 **= 0.0005** with a **variance = 0.0002**^{2}

It implies we got a good value because standard deviation is fairly low for all variables.

$$\mathbf{DF = n - (p + 1)}$$

here we have **n** as the amount of observations and **p** is the amount of variables without the intercept. That means by using DF value: **DF** = 840 − (2 + 1) = **837**, I could get the residual variance equal to **0.157** and explained variance equal to **0.0449**.

Question d: 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).

By using R commands, I will check the assumption whether residuals are normally distributed or isn't.

Here we have the concept to check with q-q plot.

```
qqnorm(fitresiduals, ylab = "Residuals", xlab = "Z − scores", vmain = "")

qqline(fitresiduals)
```

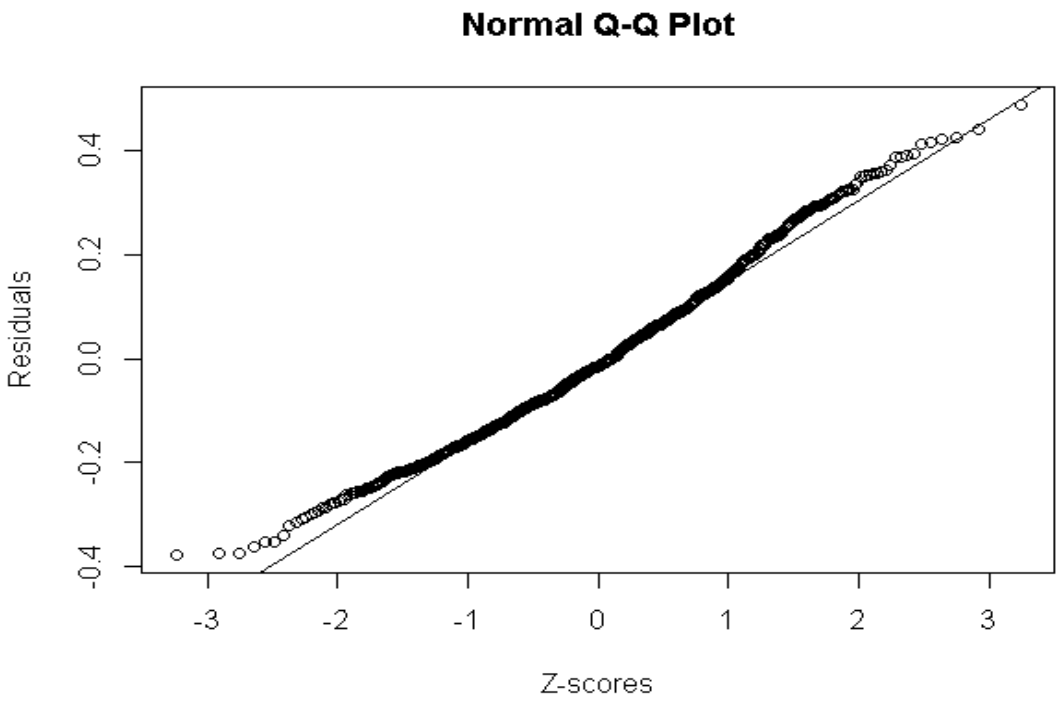

Figure 4: Normal Q-Q Plot

Form the above plot we can clearly see that **assumption is true.**

System Behaviour plots :

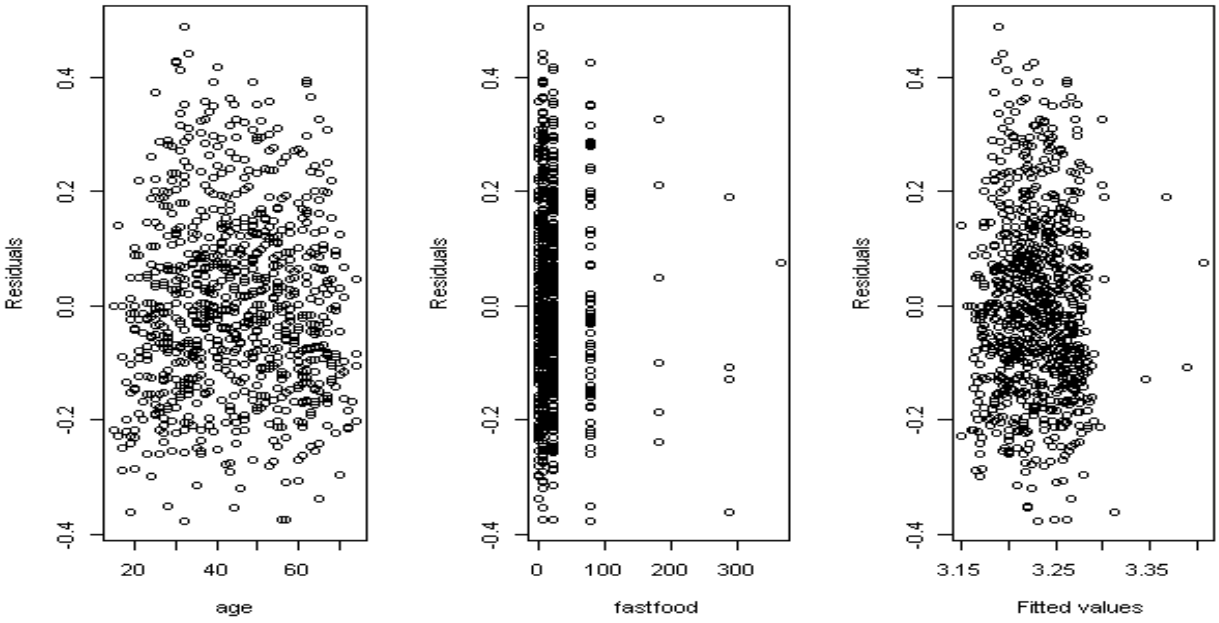

Figure 5: System Behavior Plots

We will take each figure separately; from the first figure we can see that there is apparent distribution between Age and residual values.

We still cannot see from the second figure shown above any apparent relation between residuals and Fast food, and that is the cause why coefficient of beta for the fast food is very close to zero and we could remove it as there is no effect by dropping it.

And from the residual versus plot fitted values, it shows that a large part of the data is below 3.3 on the scale of the fitted values, and with a few outliers.
Residuals are spread, but there looks to be no connexion amidst the residuals and the fitted values.

Question e: State the formula for a 95% confidence interval for the age coefficient, here deno- ted by β_1.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.

We have formula for the 95% CI for the (1 − α) quartile

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

Where→

$t_{1-\alpha/2}$ means the **(1− α /2)-quantile** from a ***i*-distribution** and with ***n*−(*p*+1)** degrees of freedom.

I have to come back to my answer "**c**" to take a look at the results, I can find that the model estimate for the age coecient equal to **0.0024** and variance of **0.0004^2**. The necessary quantile with DF **837** is equal to **1.9628**.

Calculating **95%** CI around median where (**α = 0.05**) and also obtaining median from R Script

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

And the **95%** confidence interval of mean of the is **0.0016-0.0031**, This means that the true mean lies in this type with a probability of **95%**, and I use R scripts I get the following

	2.5 %	**97.5** %
β_0 **(Intercept)**	3.0744463234	3.1504132672
β_1 **(age)**	0.0016108861	0.0031378342
β_2 **(fastfood)**	0.0002003159	0.0008803957

Table 3: Confidence Interval Regression Coefficients

Our manually calculation checks out with R script **is seen clearly**.

Question f: It is of interest whether β_1 might be 0.001. Formulate the corresponding hypothesis. Use the
significance level α = 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.

First will A null hypothesis set up:

Where→ $\beta_{0,1}$ = **0.001**

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

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

second t-statistics is used to calculate

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

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

Calculating the P-value where **837** degrees of freedom

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

We set significance level to α**= 0.05** and we can see that the P-value is less than α, when we compare between them we can say that initial assumption of null hypothesis is not true and therefore we reject the null hypothesis.Hence $H_1 : \beta_1 \neq$ **0.001**

Question g: 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.

Back to our answer ”C” ,by using our p_values for the coefficients. We can clearly see that P_value for fast food coefficient is bigger than for that of age coefficient.

So we have to drop parameter of the fast food and check if we can reduce the results of the model.

I will run the R-script to get following results:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
β_0 **(Intercept)**	3.1382	0.0176	178.33	0.000
β_1 **(age)**	0.0020	0.0004	5.41	0.000

Table 4: Estimated Parameters for Reduced Model

From the previous table, we can conclude several things:

1- The p_values for the estimated coefficients still significant.
2- New estimated β_1 still lie in the range of standard deviation

Therefore, the reduced model is good enough and useful for prediction.

Question h: 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.

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

ID	LOG_BMI	FIT	LOWER	UPPER
841	3.143436	3.233456	2.922838	3.544073
842	3.269232	3.211150	2.900472	3.521828
843	3.269438	3.229400	2.918787	3.540013
844	3.324205	3.229400	2.918787	3.540011
845	3.106536	3.227372	2.916759	3.537986
846	3.263822	3.235483	2.924860	3.546106
847	3.058533	3.186817	2.875833	3.497801

Table 5: Prediction values and 95% prediction intervals

By using the table 5 we can clearly see that the predicted log (BMI) values are in the 95% range of true values. Hence our model performs in 95% range.

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 6: Errors in predicted values

By Searching for the greatest error in the observations we will find this **4.1942840** as the biggest value and it is less that **5%**, that means our model works in **95%** confidence range.
And our mean error is equal to **2.524712** and it is acceptable. That means, our predictions are fine and also practically acceptable.