02323- Statistik

Projekt 2 BMI 2

Muhammad Ali Khan Bangash

Brugernavn	Email	Studienummer	Uddannelse
s092512	s092512@student.dtu.dk	s092512	diploming. Softwaretek.

Indholdsfortegnelse

Introduktion

Over hel verden er fedme et hastigt voksende problem. På verdensplan er fedme tredoblet siden 1975, hvor i 2016 var 38% af voksne over 18 år overvægtige, hvor 13 % var fede[1]. En simpelt og nøgleindikator for overvægt og fedme er Body Mass Index (BMI). BMI er defineret som:

$$BMI = \frac{vægt}{højde^2}$$

Hvor vægten persons vægt målt i kg og højde er persons højde målt i meter.

For voksende, kategorier er defineret i følgende tabel:

BMI	*Kategori*
BMI < 15	Meget alvorlig undervægtig
15< BMI <16	Alvorlig undervægtig
16< BMI <18.5	Undervægtig
18.5< BMI <25	Normal
25< BMI <30	Overvægtig
30< BMI <35	Moderat overvægtig
35< BMI <40	Alvorlig overvægtig
40<B MI	Meget alvorlig overvægtig

a) Statistik analyse

Projektet omfatter et datasæt af fire variabler som består af ID, BMI, alder og fastfood forbrug af 847 respondanter.

Variabel	Betydning	Definition
ID	Respondents id nummer	Kvantitativ værdi
BMI	Respondents BMI	Kvantitativ værdi
Alder	Respondents alder	Kvantitativ værdi

[1] World Health Organization

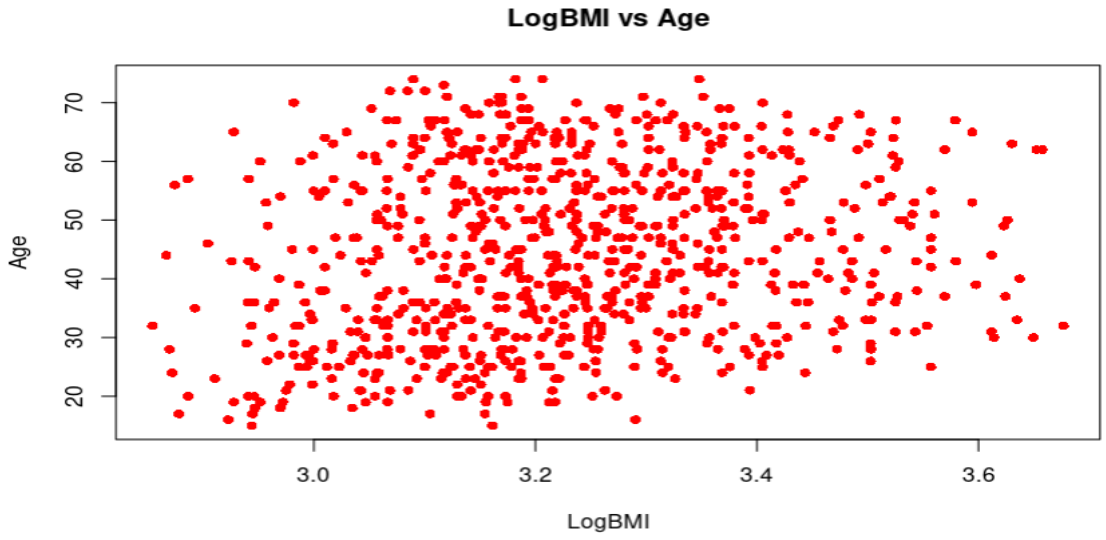
Fastfood forbrug	Antal af dage om året respondent spiser fast food.	Kvantitativ værdi.

Derudover, a Kvantitativ variabel LogBMI er blevet tilføjet til datasættet. Dette er defineret som det naturlige logartime af BMI af respondant.

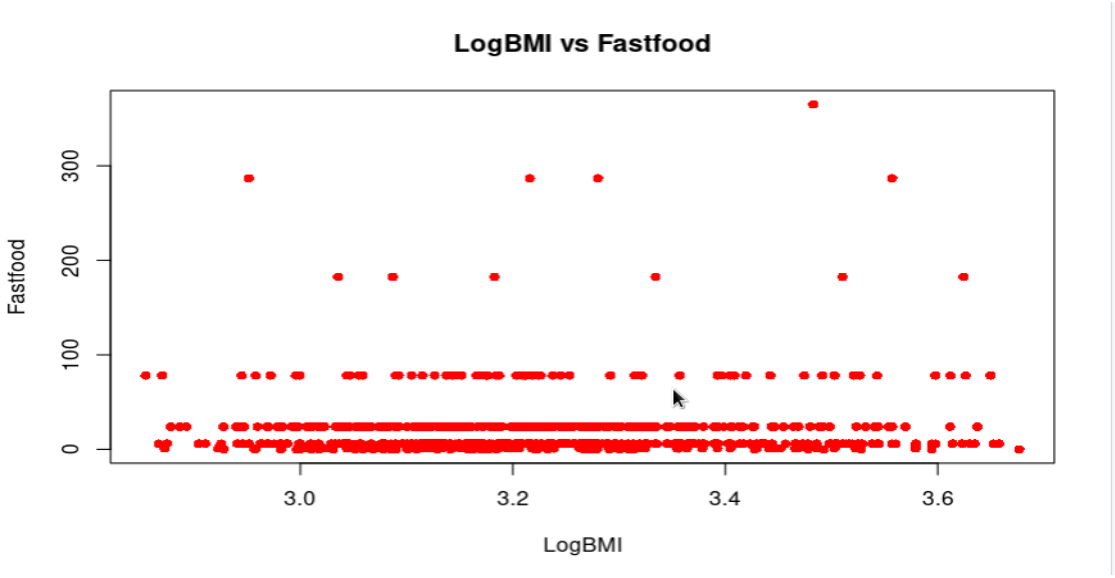
LogBMI = log (BMI)

Datasættet er komplet og består ikke af nogen manglende værdier. Man kan tage summary af datasættet for at undersøge kvartiler, median, minimum og maksimum.

```
> summary (D)
       id              bmi             age            fastfood
 Min.   :  1.0   Min.   :17.36   Min.   :15.00   Min.   :  0.00
 1st Qu.:212.5   1st Qu.:22.64   1st Qu.:32.00   1st Qu.:  6.00
 Median :424.0   Median :24.93   Median :44.00   Median :  6.00
 Mean   :424.0   Mean   :25.57   Mean   :44.62   Mean   : 19.04
 3rd Qu.:635.5   3rd Qu.:28.04   3rd Qu.:57.00   3rd Qu.: 24.00
 Max.   :847.0   Max.   :39.52   Max.   :74.00   Max.   :365.00
>
```



Figurer 1a: Scatter Plot of LogBMI vs Age



Figurer 1b :Scatter Plot of LogBMI vs Age

I både figur 1a og figur 1b, kan det ses ar der ikke ser ud til at være en sammenhæng mellem LogBMI og Alder og BMI og Fastfood. Dataene fra begge grafer ser ud til at være spredt ud uden noget klar mønster.

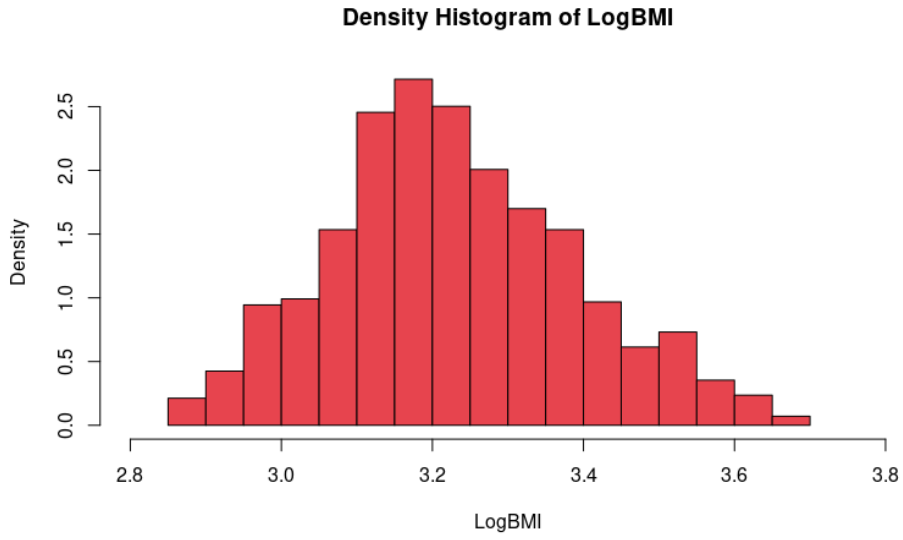


Fig 2a: Density Histrogram of LogBMI

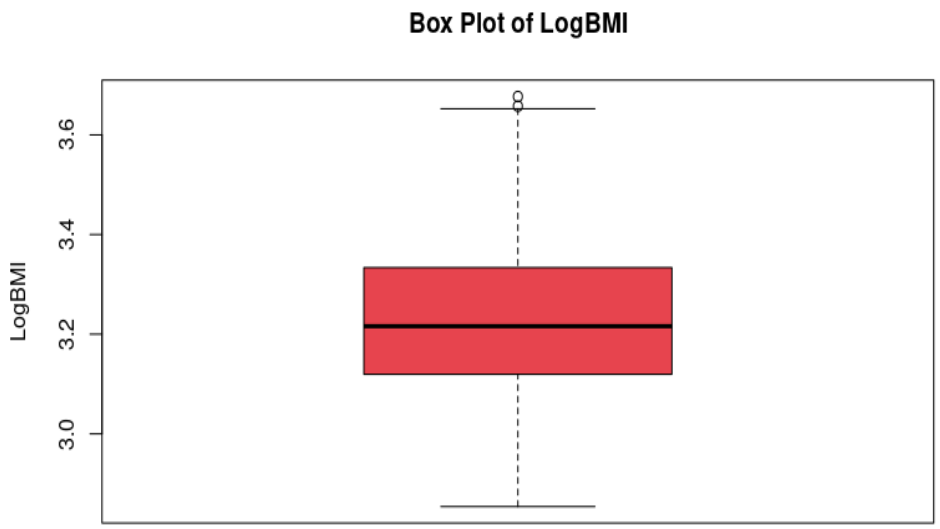


Fig 2b: Box Plot og LogBMI

På Figurer 2a kan det ses LogBMI ser ud til at være næsten normalt fordelt. Tilsvarende viser figur 2b at LogBMI ser ud til at være mere symmetrisk med two outlier over den øverste whisker.

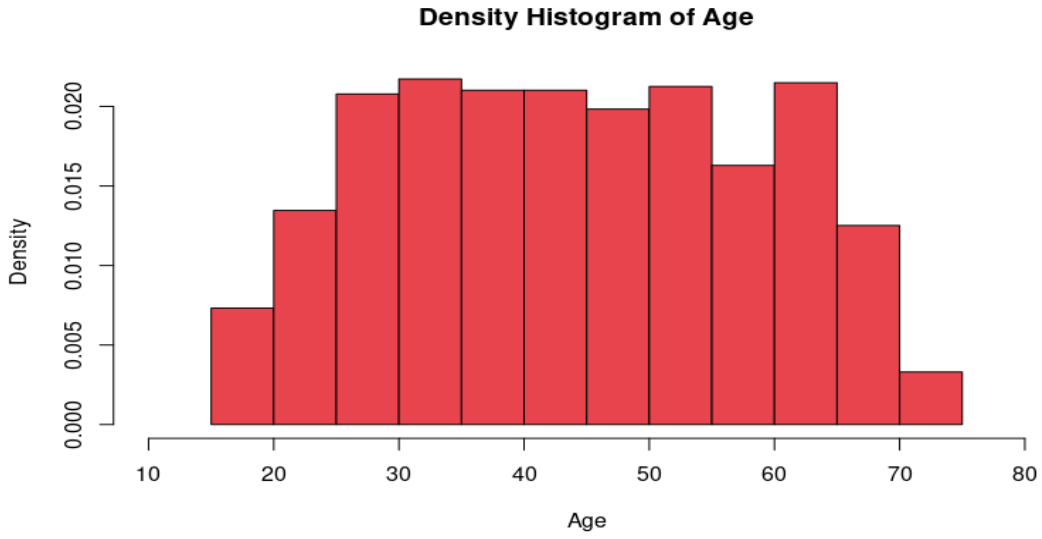


Fig3a : Density Histogram og age

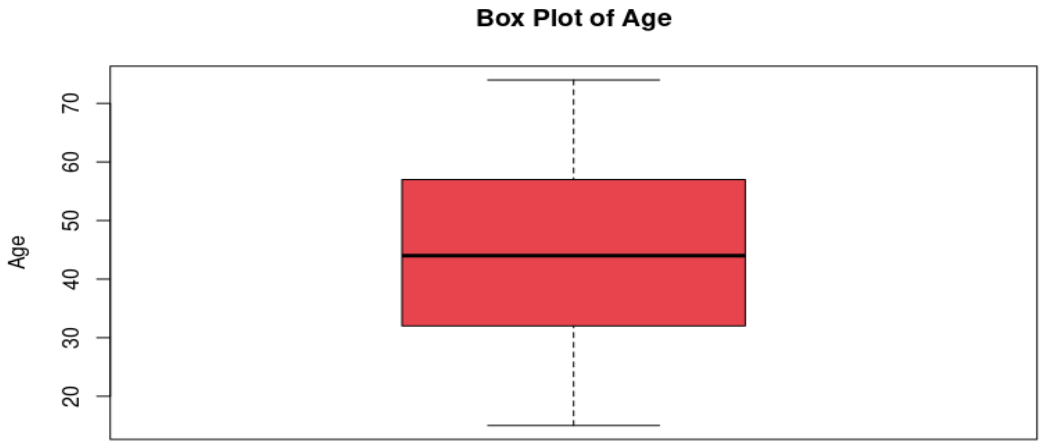
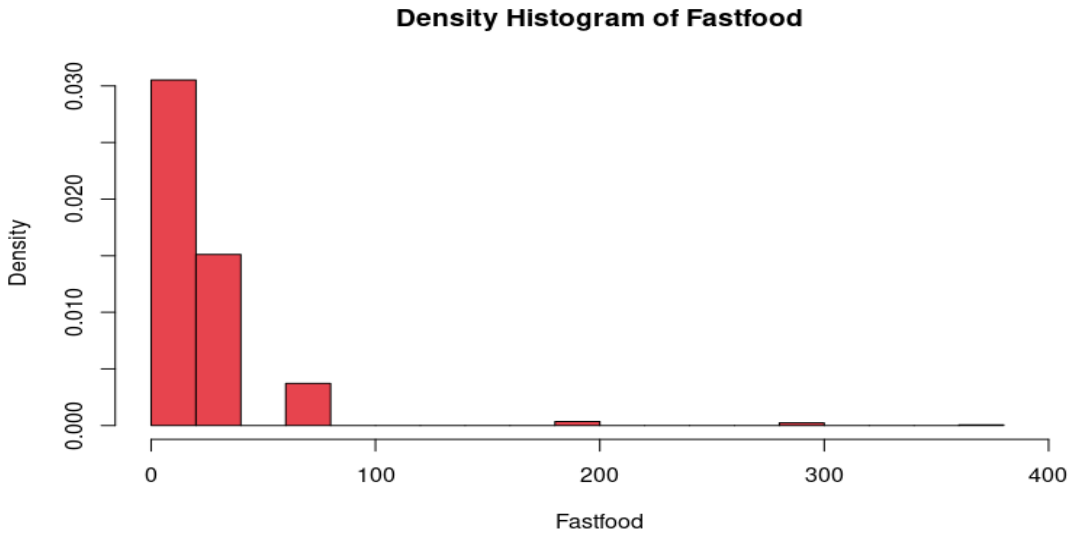
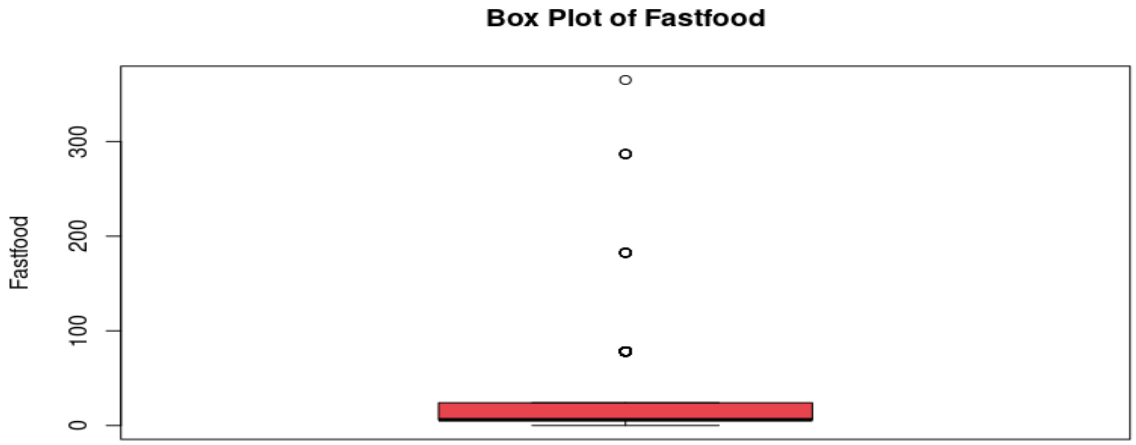


Fig3b: Boxplot of age

Af figur 3a fremgår det af der er en høj graf af varians i respondenternes alder, mens figuren 3b viser at boksplotten af respondenternes alder er fuldstændig symmetrisk uden afvigelser.



Figurer 4a: Density Histogram of fastfood



Figurer 4b: Boxplot of fastfood

Figuren 4a viser at værdierne af fastfood-indtag hos respondenterne er meget spredt ud med markante toppe og bunde mens de fleste værdier ligger under 100. I figuren 4b er det tydeligt at fastfood indtag har mange outlier med medianen, som er meget tæt på de laveste kvantil.

Tabel 3: Summary Statistik af datasæt

	Antal af observationer	*Sample Mean*	*Standard afvigelse*	*0.25 Kvantil*	*Median*	*0.75 Kvantil*
	n	$\bar{x}$	s	Q1	Q2	Q3
LogBMI	847	3.23	0.160	3.120	3.21	3.33
Age	847	44..62	14.53	32.00	44.00	57.00
Fastfood	847	19.04	32.65	6.00	6.00	24.00
BMI	847	25.57	4.22	22.64	24.93	28.04

Tabel 3 giver en mere kvantitiv og præcis beskrivelse af dataene, der viser antallet af observationer, gennemsnit, standardafvigelse og kvantiler og median af datasæt variablerne.

b) multipel lineær regressionsmodel

I denne opgave, skal der opstilles en multipel reggressionsmodel for datasættet. LogBMI vil være den afhængig variabel Y_i mens alder vil være den uafhængig variebel x1i og fastfood variebel vil også være den uafhængig variebel x2i. Modellen kan formuleres som:

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

Hvor $\beta_0, \beta_{1,}\beta_2$ er regressions parametre mens ε_i er uafhængig og ens fordelte tilfældige variabler med middelværdi 0 og en ukendt varians.

c) Modellens parametre

Parametrene som $\beta_0, \beta_{1,}\beta_2$ og residual varians σ^2 vil blive beregnet bed hjælp af R studio funktion "summary". De første 840 observationer bruges til at estimere modelparametrene mens de sidste 7 bruges til at validere. Output ser således ud:

```
> summary(fit)

Call:
lm(formula = logbmi ~ age + fastfood, data = D_model)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37643 -0.11304 -0.01488  0.09736  0.48839

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.1124298  0.0193517 160.835  < 2e-16 ***
age         0.0023744  0.0003890   6.104 1.58e-09 ***
fastfood    0.0005404  0.0001732   3.119  0.00188 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1573 on 837 degrees of freedom
Multiple R-squared:  0.04487,   Adjusted R-squared:  0.04259
F-statistic: 19.66 on 2 and 837 DF,  p-value: 4.53e-09
```

Figur 5: **Output fra Summary Function**

$$\beta_0 = 3.1124298 \approx 3.11$$
$$\beta_1 = 0.0023744 \approx 0.0024$$
$$\beta_2 = 0.0005404 \approx 0.0054$$
$$\sigma^2 = 0.1573^2 \approx 0.025$$

β_0 viser skæringspunkt med y -aksen og β_1 og β_2 viser hældning. Det kan ses at både β_1 og β_2 er positiv og meget lav værdi, hvilket betyder at regression funktion er positiv, men langsomt voksende funktion.

I den nedenstående tabel vises værdier for varians af hver variebel og frihedsgrader.

$$sd_0 = 0.0193517 \approx 0.019$$
$$sd_1 = 0.0003890 \approx 0.00039$$
$$sd_2 = 0.0001732 \approx 0.00017$$

Antal af frihedsgrader er givet ved :

$$DF = n - (p + 1)$$
$$hvor\ n = antal\ af\ observationer$$
$$p = antal\ af\ variabler$$
$$DF = 840 - (2 + 1) = 837$$

Antallet af frihedsgrader bruges til at finde restvariansen på 0.01573 og $R^2 = 0.049$

d) Model validering

I denne opgave , vil multipel lineær regression model bruges til at validere ved hjælp af residual analyse

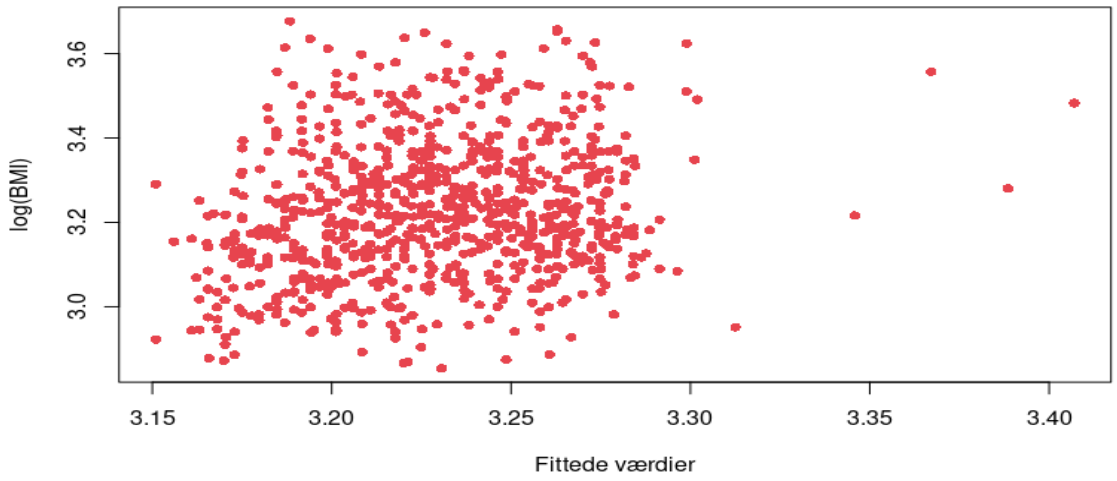


Fig 6 : Fitted Values VS logBMI

I ovenstående figur, er det tilsyneladende at de fleste værdier ligger mellem 3.15 og 3.30 med få outliers.

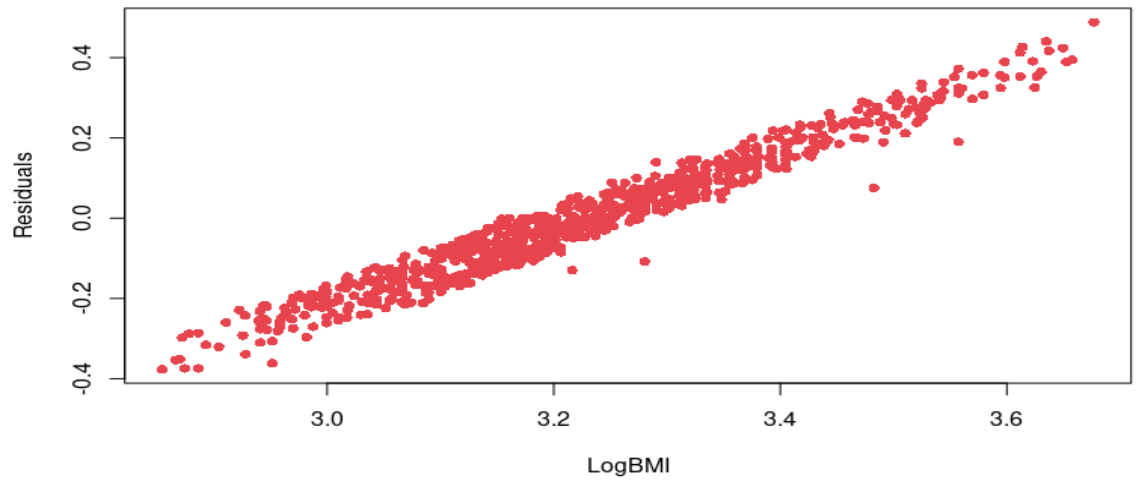


Fig 7a : LogBMI VS Residual

I figurer 7a kan det ses at er der er et klar mønster om lineær regression mellem LogBMI og Residual med få outliers.

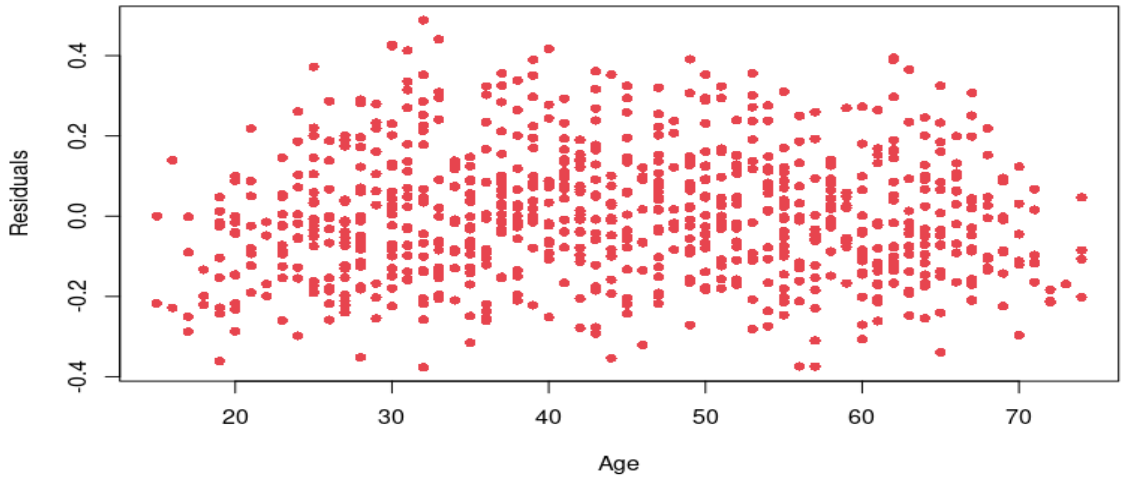


Fig 7a : AgeVS Residual

I figuren 7b , er værdierne meget spredt hvilket betyder at der ikke er noget sammenhæng mellem age og residual.

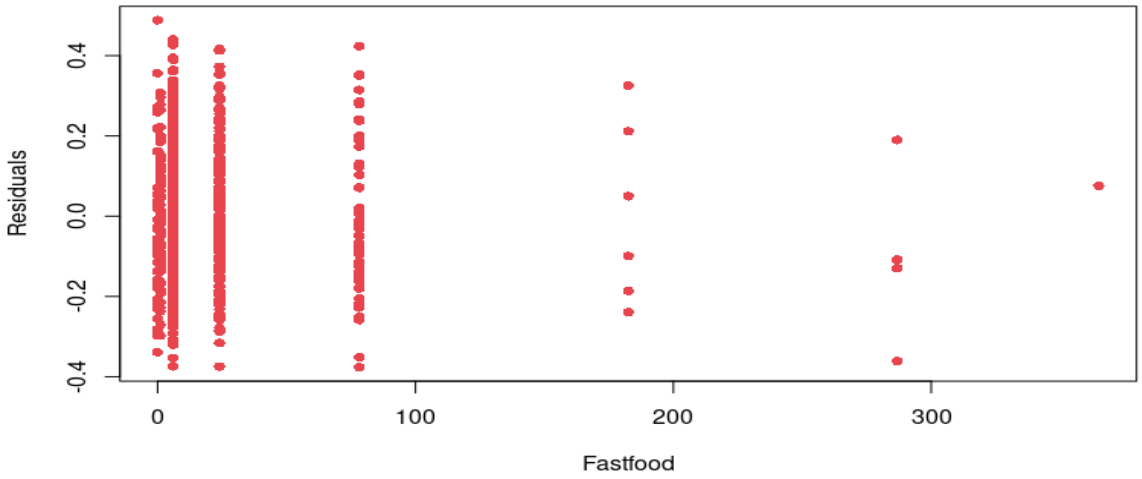


Fig 7a : FastfoodVS Residual

I figuren 7c kan det også ses at der ikke er en nogen sammenhæng mellem fastfood intaget og residualerne.

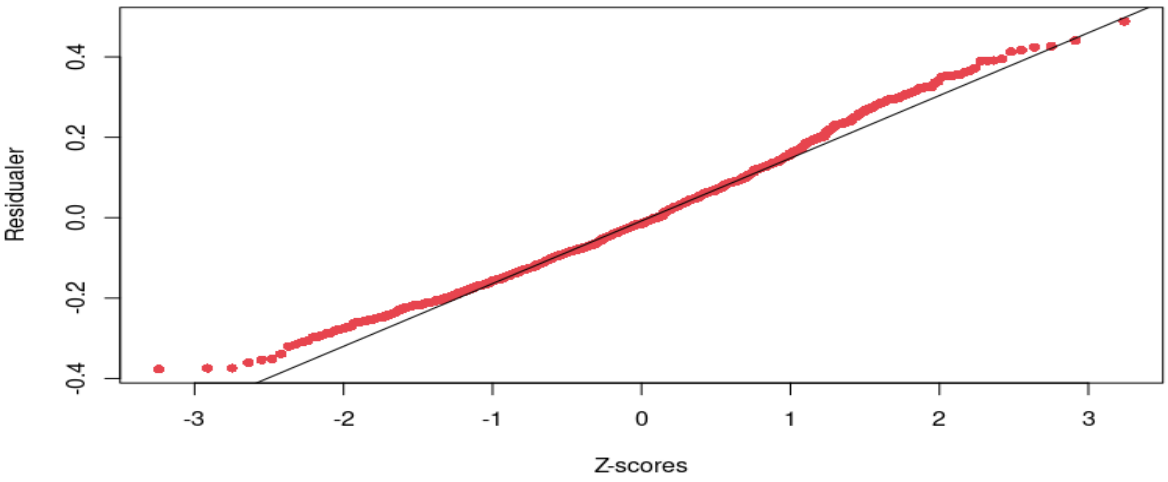


Figure 8 :QQ-plot of Residuals

Figuren 8 viser en QQ plot af residualerne. Det kan ses klart og tydeligt at værdierne ligger tæt på den lineær linje hvilket betyder at residulerne må være normalt fordelt.

e) 95% konfidensinterval for koefficienten for alder

For at finde $1-\alpha$ konfidensinterval for variabel af indeks **i** , vil der benyttes følgende formel :

$$CI_i = \hat{\beta}_i \pm t_1 - \frac{\alpha}{2} * \hat{\sigma}_{\beta_i}$$

Hvor $t_1 - \frac{\alpha}{2}$ er $1 - \frac{\alpha}{2}$ kvantil fra t-fordelingen med $n-(p+1)$ antal frihedsgeader. Vi tager værdierne fra Figur 5:

$$\hat{\beta}_i = 0.002377 \approx 0.0024$$
$$\hat{\sigma}_{\beta_i} = 0.003890 \approx 0.00039$$
$$df = 837$$

Kvantil kan findes ved hjælp af følgende R-command $qt\left(\left(1-\frac{\alpha}{2}\right), df\right)$:

$$qt((0.975,837) = 1.962802 \approx 1.96$$

95% konfidensinterval for koefficienten alder kan nu beregnes :

$$CI_i = 0.0024 \pm 1.96 * 0.00039$$
$$CI_i \approx 0.0016356; 0.0031644$$

Det betyder at 95% konfidensintervallet for alders koefficienten ligger mellem 0.0016 og 0.0031.

Overstående resultat kan sammenlignes med en anden resultat der skal estimeres ved hjælp af en R-Command *confint* der beregner konfidensinterval for fastfood og age og intercept hvilket betyder skæring med y-aksen.

	2.5%	*97.5%*
Age	0.0016108861	0.0031378342
Fastfood	0.0002003159	0.0008803957
Intercept	3.0744463234	3.1504132672

f) Hypotese test

For at finde ud af om β_1 kan være 0.001, skal der opstilles en null hypotese med et 5% signifikansniveau:

$$H_0: \beta_1 = 0.001$$
$$H_1: \beta_1 \neq 0.001$$

Først skal der beregnes t-statistik ved hjælp af følgende formel :

$$t_{obs,\beta 1} = \frac{\hat{\beta}_1 - \beta H_0}{\hat{\sigma}_{\beta_1}}$$
$$\approx \frac{0.0024 - 0.001}{0.00039}$$
$$\approx 3.59$$

P-værdien kan regnes ved hjælp af følgende formel :

$$p - værdi = 2.P(T > |t_{obs,\beta 1}|)$$

Hvor $P(T > |t_{obs,\beta 1}|)$ er sandsynligheden for $t_{obs,\beta 1}$er mindre end t-fordeling som vi fandt i opgave c ved at bruge n -1 frihedsgrader. Ved hjælp af R-command pt() kan man finde p-værdien:

$$p - værdi \approx 0.0003$$

Da p-værdien 0.0003 < 0.05, kan det konkluderes at nulhypotesen kan forkastes, hvilket betyder at alder koefficienten ikke er lig med 0.001 med et signifikansniveau.

g) Backward selection

Backward selection er en metode i regression hvor man kan eliminere de ikke betydelige variabler fra modellen så kun de betydelig variable er tilbage. I figuren 5 kan det der ses p-værdierne for variabler:

$$p_0 = 2 * 10^{-16}$$
$$p_1 = 1.58 * 10^{-9}$$
$$p_2 = 0.00188$$

Da $\alpha = 0.05$ $som\ er\ signifikansniveau$ og p-værdier for andre variabler er mindre end 0.05, kan det konkluderes at alle de tre variabler var betydlige da $\alpha = 0.05$ og derfor behøves der ikke at gøre brug af backward selection for at reducere modellen.

h) 95% prædiktionsintervaller

Table 5: Fit tabel

ID	LogBMI	Fit	Nedre	Øvre
841	3.143436	3.236993	3.236993	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

Det fremgår af tabel 5 , at modellen er præcis da alle observationer i LogBMI kolonnen ligger indenfor 95% prædiktionsintervallet. Det betyder at modellen rammer præcist indenfor prædiktions konfidensinterval set i kolonnerne “Nedre” og “Øvre”