02402, Statistik

BMI undersøgelse

Projekt 2

S224465 – Aniq K. Shamim

07-11-2023

Indhold

Statistisk analyse

a) Beskrivelse af datamaterialet

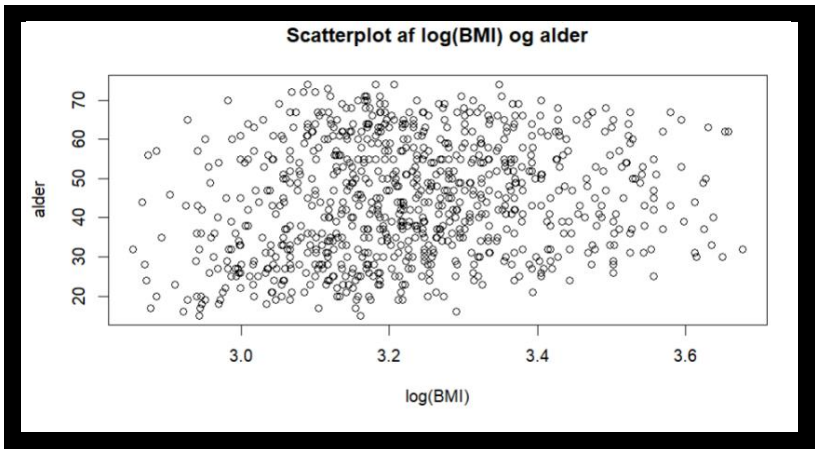
I dette projekt, vil der analyseres data fra BMI-undersøgelsen, hvor målet er at opstille en passende multipel lineær regressionsmodel for BMI. Det givne datasæt indeholder 847 observationer, hvor der herunder fokuseres på fire variable, som er:

- ID: *Respondentens nr. (kan bruges til identifikation)*.
- BMI: *Respondentens BMI (i kg/m^2)*.
- Alder: *Respondentens alder (i år)*.
- Fastfood: *Hyppighed af respondentens besøg ved fastfood restauranter (dage/år)*.

Derudover tilføjes også en variabel mere, som er log(BMI), hvilket blot er logaritmen til BMI. Denne variabel gør det muligt at udføre udregninger lettere, senere hen.

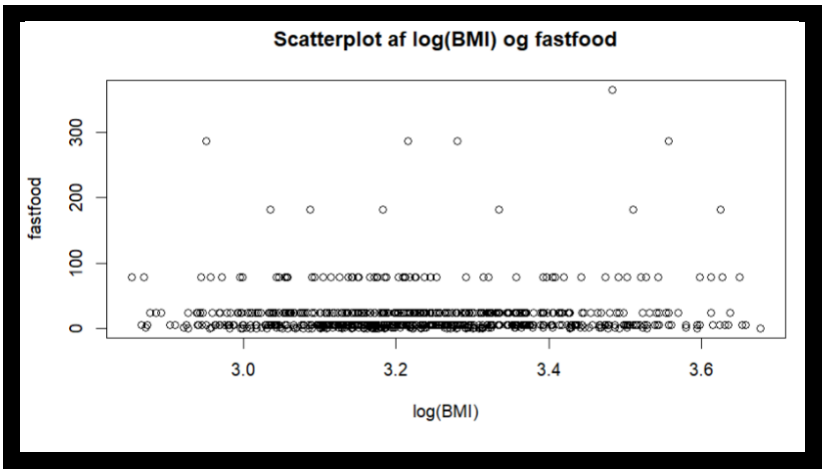
Dette er en komplet stikprøve, da det er alle værdier, som er givet. Alle nævnte variable er desuden kvantitative, da dataet er numerisk, samt både kan kvantiseres og måles. Af de ovennævnte variable, udvælges nu 3, som er log(BMI), alder og fastfood. For de 3 variable, laves nogle scatterplots, hvor log(BMI), stilles mod alder og fastfood, for at undersøge sammenhængen mellem dem.

På følgende scatterplot af log(BMI) overfor alder, ser det ud til, at der ikke er en egentlig sammenhæng mellem de to faktorer. Dette kan ses på det spredte data.



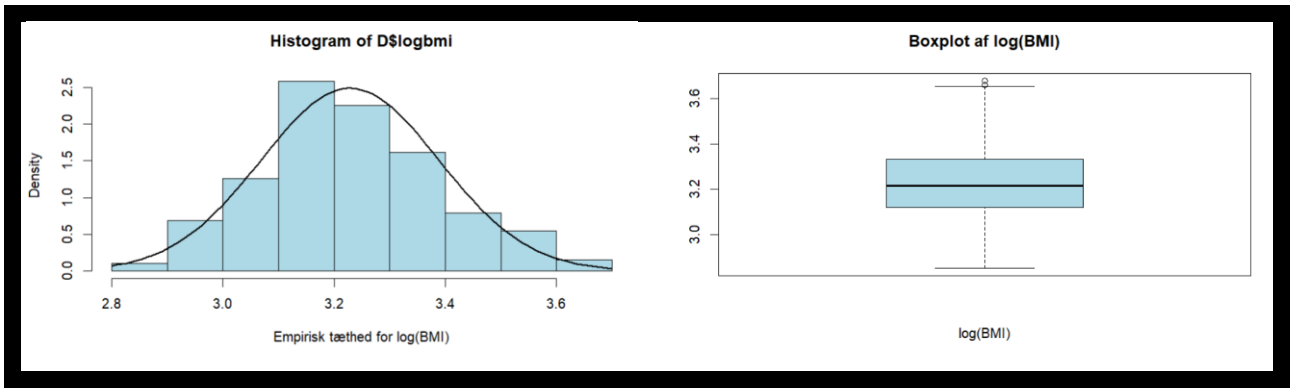
Figur 1: Scatterplot af log(BMI) og alder.

Det samme gør sig gældende på dette scatterplot af log(BMI) overfor fastfood. Igen ses meget spredt data, hvorved det ud fra denne figur, kan siges, at der ikke er nogen sammenhæng, selvom at det går imod den virkelige forventning man har om fastfood og BMI.



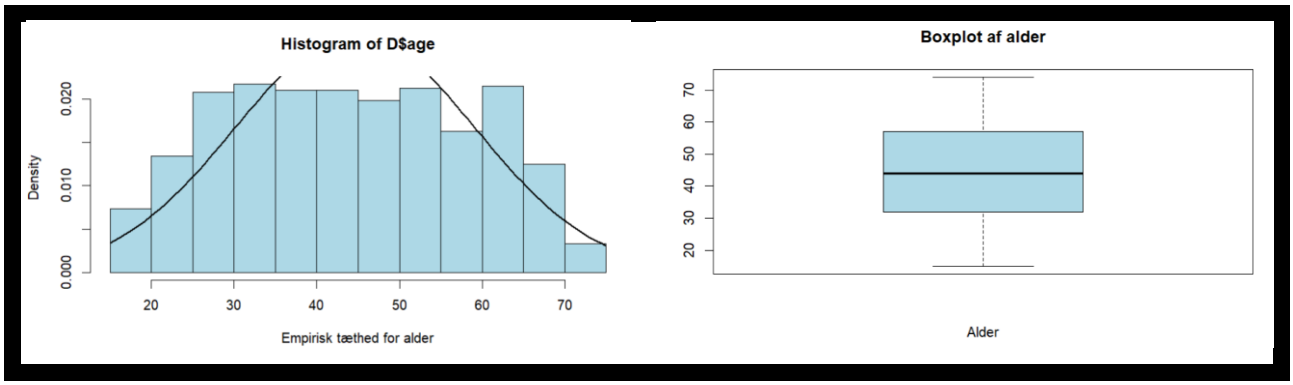
Figur 2: Scatterplot af log(BMI) og fastfood.

Fordelingen af observationerne for log(BMI), alder og fastfood undersøges nu. For at gøre dette, opstilles densitetshistogrammer, samt boxplots, for hver af de tre variable.



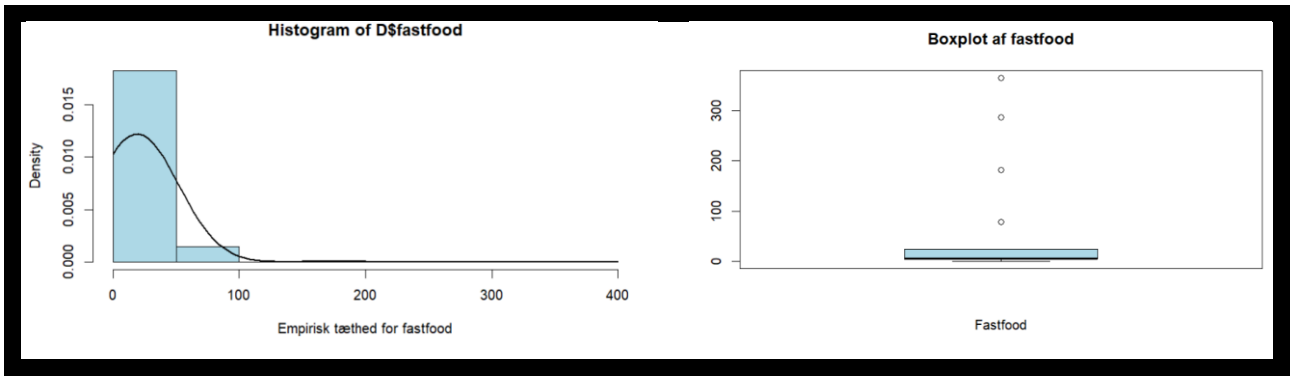
Figur 3: Histogram og boxplot for log(BMI).

På dette histogram for log(BMI), kan det ses, at dataet er delvist normalfordelt. Derudover kan det ses på boxplottet, at medianen af dette data, er på omtrent 3,2 og derfor ligger lidt længere mod det nedre kvartil. Det skal også bemærkes, at der kan ses to ekstreme værdier ved de højeste værdier af boxplottet.



Figur 4: Histogram og boxplot for alder.

Dette histogram for alder viser, at observationerne er meget spredte. Dette gør også, at standardafvigelsen er høj. På boxplottet kan det aflæses, at medianen er på cirka 45, hvor den også befinder sig nogenlunde i midten af 25%-kvartilet og 75%-kvartilet. Boxplottet afspejles dog ikke så godt på histogrammet i dette tilfælde, da dét viser en symmetrisk fordeling, hvorimod histogrammet, viser mere ligeligt fordelt data.



Figur 5: Histogram og boxplot for fastfood.

Ovenstående histogram af fastfood, er en klar illustration af data, som er meget spredt. Der kan ses, at der er en masse observationer under 100 og meget få over 100. På boxplottet, kan dette også ses tydeligt, hvor en stor del af observationerne kan aflæses til at ligge under 100. Medianen ses også værende langt nede, nær 25%-kvartilet. Ydermere er der også fire ekstreme værdier, som er mod den højere ende af boxplottet. Disse fire ekstreme værdier er med til at forårsage en høj standardafvigelse, grundet deres beliggenhed ift. medianen.

Al relevant data, som disse figurer indeholder, indsættes nu i en tabel med opsummerende størrelser således:

Variabel–navn	Antal observationer	Stikprøve–gennemsnit	Stikprøve–standardafvigelse	Nedre kvartil (25%)	Median (50%)	Øvre kvartil (75%)
BMI	847	25.57	4.22	22.64	24.93	28.04
Log(BMI)	847	3.228	0.160	3.120	3.216	3.334
Alder	847	44.62	14.532	32	44	57
Fastfood	847	19.04	31.651	6	6	24

Tabel 1: Opsummerende størrelser for variablene.

Denne tabel opgiver præcise tal, hvor boksplots og histogrammer fra tidligere, er med til at give et visuelt overblik.

b) Multipel lineær regressionsmodel

Nu opstilles en multipel lineær regressionsmodel med logaritmen til BMI som responsvariabel, beskrevet ved γ_i, og med alder og fastfood-forbrug, som forklarende variable, beskrevet ved hhv. $x_{1,i}$ og $x_{2,i}$. Dette gøres for at undersøge sammenhængen mellem γ og x_1, $x_{2,i}$ ... x_p, med følgende:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

ε_i er en beskrivelse af residualer, hvor fejlene er uafhængige og ensfordelte med $\varepsilon_i \sim N(0, \sigma^2)$. Gennemsnittet er lig 0, og der er en varians, som er ukendt. Derudover fortæller β_0, hvor der er skæring med y-aksen og β_1 og β_2 repræsenterer hældningen af de to variable. Ydermere er der x_1 og x_2, som begge er forklarende variable, der hhv. beskriver alderen og fastfood-indtaget.

c) Estimering af modellens parametre

Nu estimeres modellens parametre, som består af regressionskoefficienterne, kaldet β_0, β_1 og β_2 og residualernes varians, kaldet σ^2. Dette gøres ved hjælp af R, hvorfra der fås følgende output.

```
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
```

Derfra kan det aflæses, at estimaterne for β_0, β_1 og β_2 er:

$$\beta_0 = 3.1124298$$
$$\beta_1 = 0.0023744$$
$$\beta_2 = 0.0005404$$

Det kan her ses, at både β_1 og β_2 er positive, hvilket må betyde, at det er en voksende funktion. Dog en langsomt voksende funktion, grundet de lave værdier.

De estimerede standardafvigelser for β_0, β_1 og β_2, kan også aflæses til nedenstående, gennem samme R-kode:

$$\text{Std. Error } (\beta_0) = 0.0193517$$
$$\text{Std. Error } (\beta_1) = 0.0003890$$
$$\text{Std. Error } (\beta_2) = 0.0001732$$

Nu findes frihedsgraderne anvendt til estimatet af residualernes varians σ^2. Dette gøres ved formlen:

$$DF = n - (p + 1)$$

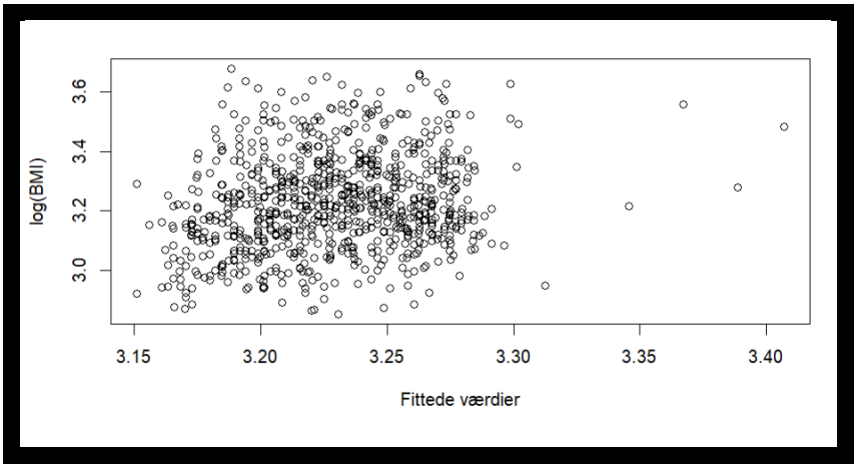
Her indsættes tallene, hvor n angiver antallet af observationer, som i dette tilfælde er 840 og p angiver antallet af variable, som er 2. Dette bliver således.

$$DF = 840 - (2 + 1) = 837$$

Herfra vides det så også, at residualernes varians (σ^2) er på 0.1573 og modellens forklarede varians (R^2) er på 0.0448.

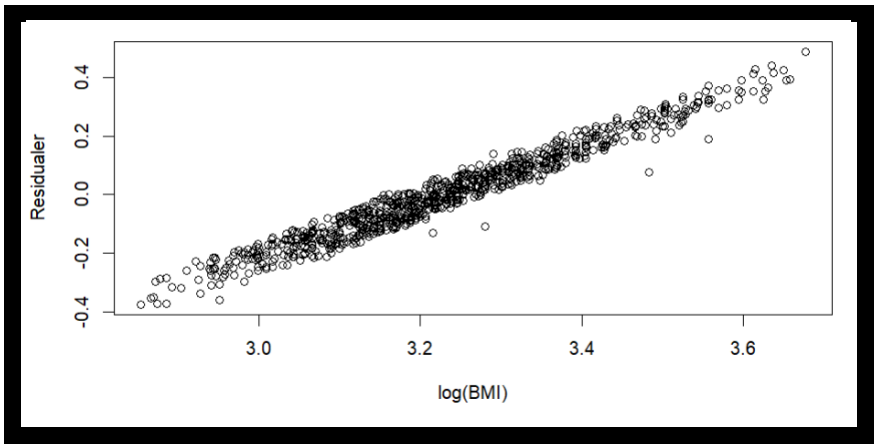
d) Modelkontrol af forudsætningerne for modellen

Gennem en modelkontrol, undersøges nu, om forudsætningerne for modellen er opfyldte. Dette gøres ved at se nærmere på – og analysere plots, som er lavet via R. Disse plots er opstillet, hvor observationer, fittede værdier og residualer er sat op mod hinanden.



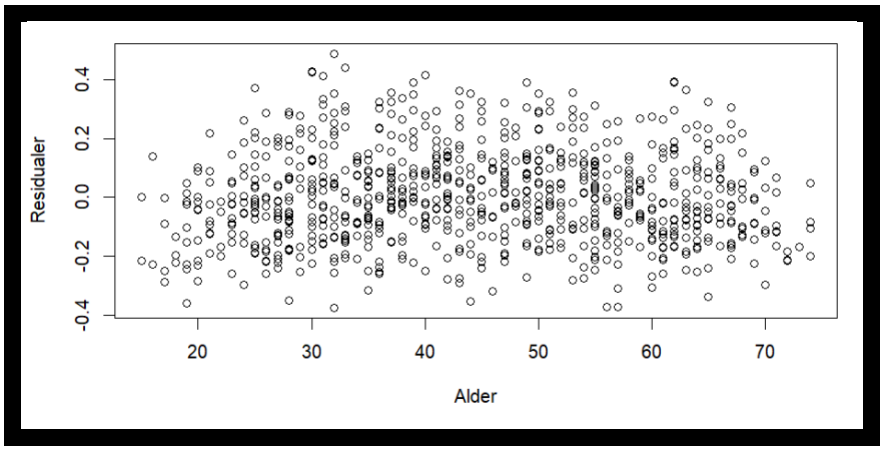
Figur 6: log(BMI) op imod de fittede værdier.

I ovenstående plot, af log(BMI) op imod de fittede værdier, kan det aflæses, at på trods af størstedelen af punkterne, som befinder sig mellem cirka 3.15 og 3.30 på x-aksen, er stadig der en stor spredning af datapunkterne. Dog med enkelte, mere ekstreme afvigelser. Dette fortæller, at der er meget lille – eller ingen sammenhæng mellem log(BMI) og de fittede værdier.



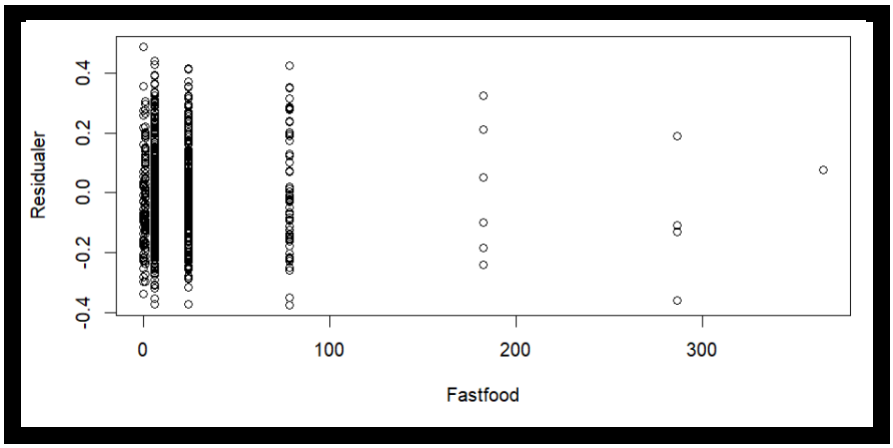
Figur 7: Residualer op imod log(BMI).

Kigges der derimod på et plot af residualer op imod log(BMI), ses en tydelig lineær sammenhæng af datapunkterne, da de ligger tæt, og nærmest på linje. Dog er der igen enkelte punkter, der afviger lidt. På baggrund af disse antagelser, kan det konkluderes, at residualer og log(BMI), har en systematisk lineær sammenhæng.



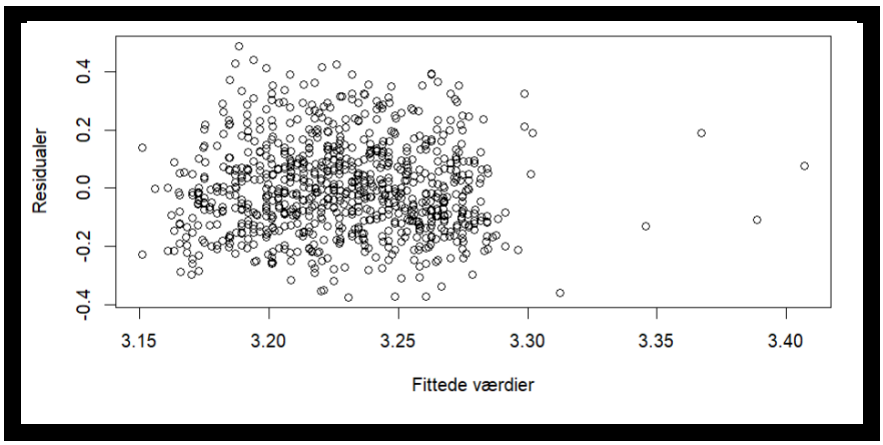
Figur 8: Residualer op imod alder.

Nu er et plot af residualer op imod alder udarbejdet. Dette plot viser en stor spredning af dataet, fordelt ud over hele figuren. Herfra vides det, at der er en meget lille – eller ingen sammenhæng mellem residualer og alder.



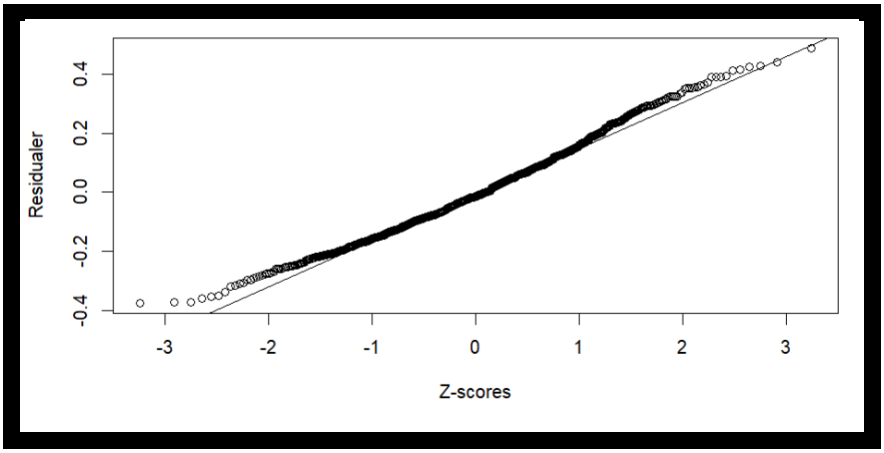
Figur 9: Residualer op imod fastfood.

I ovenstående plot af residualer op imod fastfood, ser man en tendens til dataansamlinger i den lodrette retning. Dog er disse datapunkter stadigvæk spredte i x-aksens retning – og uden den rette lineære form til, at en systematisk lineær sammenhæng kan ses. Derfor har residualer og fastfood en meget lille – eller ingen sammenhæng.



Figur 10: Residualer op imod de fittede værdier.

Plottet herover illustrerer residualer op imod de fittede værdier. Som det før er set, så ses her en stor spredning af datapunkterne med mindre afvigelser, hvilket fortæller, at der må være en meget lille – eller ingen sammenhæng mellem residualer og de fittede værdier.



Figur 11: Q-Q-plot.

Ovenfor er opstillet et Q-Q-plot. Dette bruges til at undersøge om, hvorvidt residualerne er normalfordelte eller ikke er. Det kan aflæses, at langt størstedelen af datapunkterne, befinder sig på den såkaldte Q-Q-line, hvilket er en indikator for, at dataet er normalfordelt. Det skal dog bemærkes, at der ved "halerne", er mindre afvigelser. Residualerne kan stadig være normalfordelt, på trods af dette.

e) Undersøgelse af konfidensinterval for alder

For at kunne beregne et 95% konfidensinterval for koefficienten for alder, i dette tilfælde kaldet β_1, bruges formlen:

$$\hat{\beta}_i \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_i}$$

I denne formel er $t_{1-\alpha/2}$, svarende til $(1-\alpha/2)$-fraktilet af en t-fordeling med $n-(p+1)$ frihedsgrader. Så $t_{0.975}$ udregnes, hvor $\alpha = 5\% = 0.05$ for et 95% konfidensinterval. Frihedsgraden er fundet til $\mathrm{DF} = 837$, som vist i opgave c. Nu bruges den indbyggede qt-funktion i R, hvorfra at følgende resultat fås:

```
> # Fraktil til udregning af konfidensinterval
> qt(0.975, 837)
[1] 1.962802
```

Med denne værdi fundet, samt værdien af estimatet – og standardafvigelsen (std. error) for alder fundet (begge fundet i opgave c), kan formlen for konfidensintervaller nu bruges.

$$0.0023744 + 1.962802 \cdot 0.0003890 = 0.001610$$
$$0.0023744 + 1.962802 \cdot 0.0003890 = 0.003137$$

Konfidensintervallet er hermed fundet til at være [0.001610, 0.003137] for β_1. Samme metode kan anvendes til at finde konfidensintervallet for de to andre koefficienter i modellen. Konfidensintervallerne kan derudover også findes ved blot at benytte R.

```
> # Konfidensintervaller for modellens koefficienter
> confint(fit, level = 0.95)
                    2.5 %       97.5 %
(Intercept) 3.0744463234 3.1504132672
age         0.0016108861 0.0031378342
fastfood    0.0002003159 0.0008803957
```

Hermed er den manuelle udregning af alder også blevet valideret, da resultaterne, som var udregnet med formlen, giver det samme som R-outputtet. Desuden kan konfidensintervallerne for de andre to koefficienter også ses på dette R-output. Her er det [3.074446, 3.150413] for β_0 og [0.000200, 0.000880] for β_2.

f) Hypotesetest

Da man nu er interesseret i, om β_1 kunne have værdien 0.001, udføres en hypotesetest. Dette gøres ved at teste hypotesen:

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

Det anvendte signifikansniveau er $\alpha = 5\% = 0.05$. Dette signifikansniveau bruges til at afgøre om en hypotese accepteres eller forkastes. Det sker i samspil med p-værdien. En p-værdi, der er større end signifikansniveauet, vil føre til, at hypotesen accepteres og en p-værdi, der er mindre end signifikansniveauet, vil føre til, at hypotesen forkastes.

P-værdien skal findes for at kunne lave denne hypotesetest. For at finde denne p-værdi, skal teststørrelsen beregnes. Det kan gøres ved brug af formlen:

$$t_{obs,\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}}$$

Tallene fra tidligere skal blot indsættes i denne formel for teststørrelsen:

$$t_{obs,\beta_i} = \frac{0.0023744 - 0.001}{0.0003890} = 3.533162$$

P-værdien kan herfra udregnes, hvilket gøres således:

$$p - værdi = 2 \cdot P\left(T > \left|t_{obs,\beta_i}\right|\right) = 2 \cdot P(1.962802 > 3.533162) = 0.0004$$

Da p-værdien nu er fundet, kan det ses, at den er lavere en signifikansniveauet på 0.05, og at der dermed er meget stærk evidens imod H_0. Det kan herfra konkluderes, at nulhypotesen forkastes, da β_1 ikke kan have værdien 0.001 ved med et signifikansniveau på 5%.

g) Undersøgelse med backward selection

Backward selection er en måde, hvorpå en model kan forsimples. Ved at udelade de mindre betydelige variable og beholde de betydelige variablen, kan dette gøres.

I opgave c, blev der fundet p-værdier for hver variabel, gennem R. Værdierne er gengivet her:

P-værdi (β_0): $2 \cdot 10^{-16}$

P-værdi (β_1): $1.58 \cdot 10^{-9}$

P-værdi (β_2): 0.00188

Tages udgangspunkt i disse p-værdier, kan de alle ses værende under signifikansniveauet på de 5%, hvilket betyder, at der er statistisk signifikans og disse p-værdier dermed må være betydelige. Af denne årsag, er backward selection ikke relevant at lave, i dette tilfælde.

h) Prædiktioner

Nogle af de statistiske modeller i projektet, er kun opstillet på baggrund af 840, ud af de 848 observationer. Grunden til dette er, at de sidste 7 observationer skal bruges, hvor der bestemmes prædiktioner og 95% prædiktionsintervaller for logaritmen til BMI. Dette gøres via R-funktionen, kaldet ”predict”, som giver følgende tabel.

id	logbmi	fit	lwr	upr
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

Fra denne model, kan det aflæses, at alle de observerede log(BMI)-værdier for de syv sidste observationer, ligger indenfor 95% prædiktionsintervallerne. Modellens prædiktive nøjagtighed er derfor bemærkelsesværdig.