02402 - Statistics Projekt 2: Bmi undersøgelse

Daniel F. Hauge - s2011687

Dette projekt består af en statistisk analyse af et datasæt med mennesker. Projektet forsøger at kaste lys på BMI med statistik. Projektet henvender sig til læserer som er familiær med projekt beskrivelsen fra statistik kurset 02402 fra DTU og befærder sig komfortabelt i statistiske begreber og metoder.

Projektet er lavet som en R notebook, og indeholder derfor områder hvor R er brugt som redskab til udregning og plot optegninger. Bokse med skriften `Code` indikere at der er blevet kodet noget i R til at udregne, tegne eller gemme værdier til senere brug. R koden kan findes på følgende måder:

- Se den medfølgende ".R" fil, eller via på:
- Se den medfølgende ".rmd" fil, eller på:
- Projektet kan også findes i renderet form med kode afsnit på: https://htmlpreview.github.io/?https://github.com/DanielHauge/02402-statistics/blob/master/Project1/project.nb.html

a) Descriptiv analyse

Data)

Datamaterialet indeholder 847 observationer med 5 egenskaber på mennesker for projektets problemstilling. Variable på hver observation er som følger:

Variable Navn	Måle type	Måle enhed	Forklaring
id	N/A	Heltal	Et unikt tal der kan bruges som identification for observationen.
age	Kvantitativt	År	Personens alder målt i år.
fastfood	Kvantitativt	dage pr. år	Antal dage per år personen har spist fastfood.
bmi	Kvantitativt	Bmi	Dette er en enhed som prøver at beskrive kroppens størelses forhold, baseret på vægt og højde. Bmi står for "Body Mass Index".
logbmi	Kvantitativt	Log(Bmi)	Dette er bmi'en der er blevet log transformeret.

Nedenfor ses første og sidste observation som et præsenterende eksempel:

	id <int>	bmi <dbl>	age <int>	fastfood <dbl>	logbmi <dbl>
1	1	21.2963	44	0.0	3.058533
847	847	21.2963	24	78.2	3.058533

2 rows

Fordeling

Nøgle tal

```
fastfood:
     Antal Observationer: 847
     Middelværdi: 19.0446280991736
     Variation: 1066.10341969013
     Standard afvigelse: 32.6512391754146
     Nedre kvartil: 6
     Median: 6
     Øvre kvartil: 24
```

```
age:
     Antal Observationer: 847
     Middelværdi: 44.6221959858323
     Variation: 211.202249072655
     Standard afvigelse: 14.5327990790713
     Nedre kvartil: 32
     Median: 44
     Øvre kvartil: 57
```

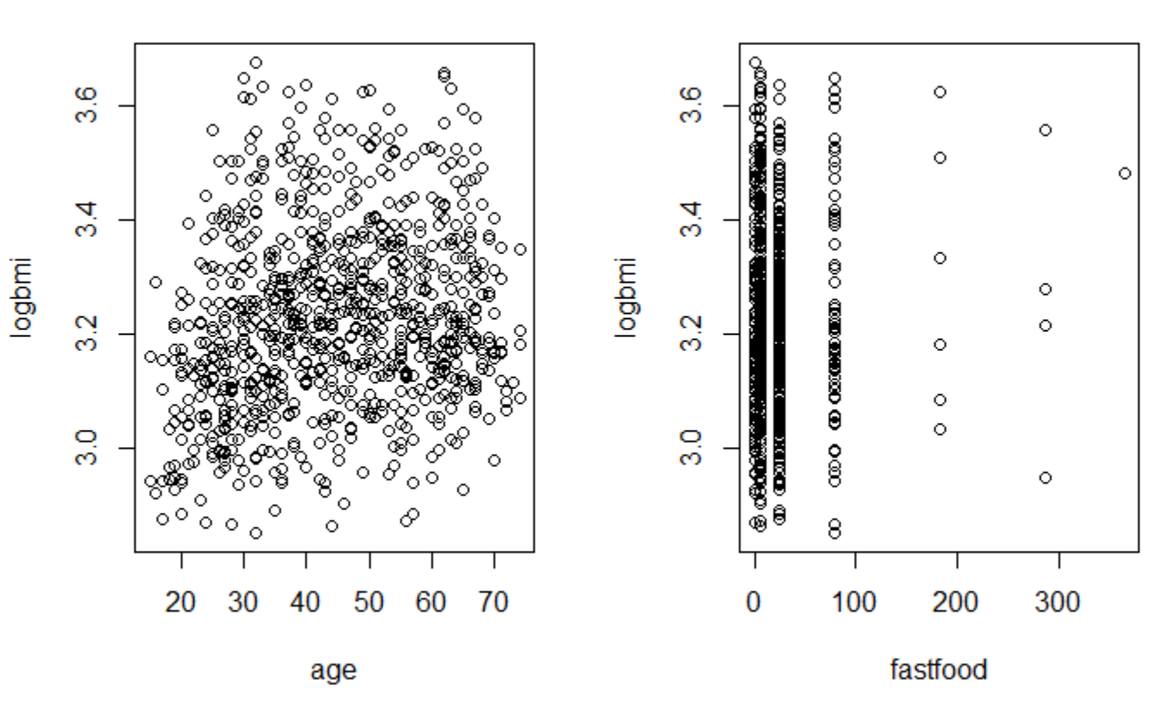
```
logbmi:
     Antal Observationer: 847
     Middelværdi: 3.22849470250382
     Variation: 0.0257192700029911
     Standard afvigelse: 0.160372285644968
     Nedre kvartil: 3.11961976167143
     Median: 3.2161018978877
     Øvre kvartil: 3.33360177544307
```

Plots

I denne sektion kastes der lys på data'en med plots.

Scatterplots

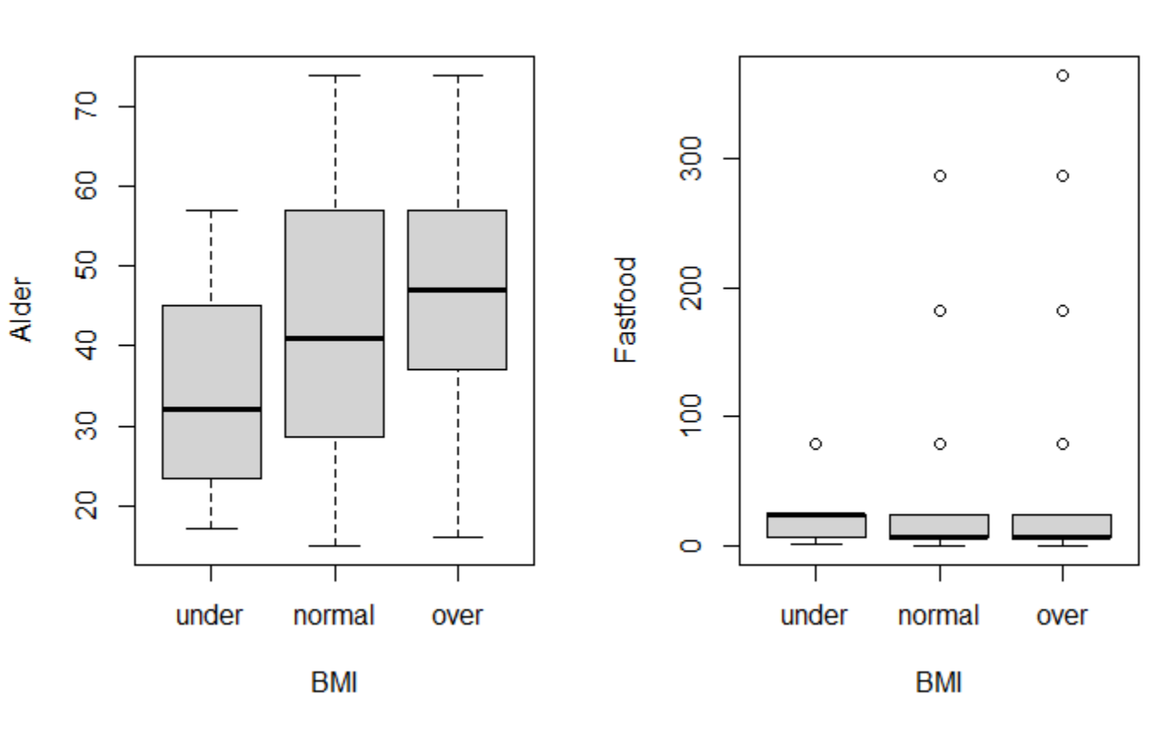
Nedenfor ses simple scatter plots af logbmi på y aksen, år og fastfood på x aksen.



En llle ting der bør nævnes er at fastfood har en katagoriserende natur, men er i kvantitativ form, derfor ses observationerne af fastfood at ligge på linjer.

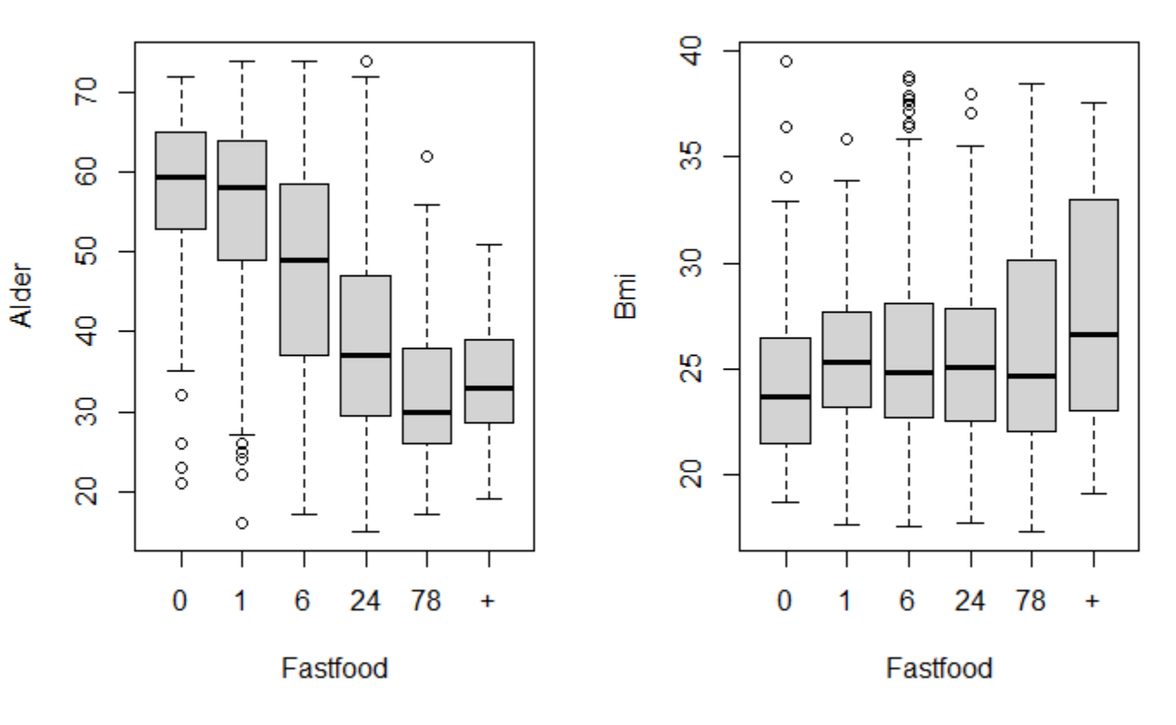
Segmentering & Sammenligning

Først kan vi segmentere efter BMI. Nedenfor ses 2 boxplots der segmentere BMI i under, normal og overvægtig.



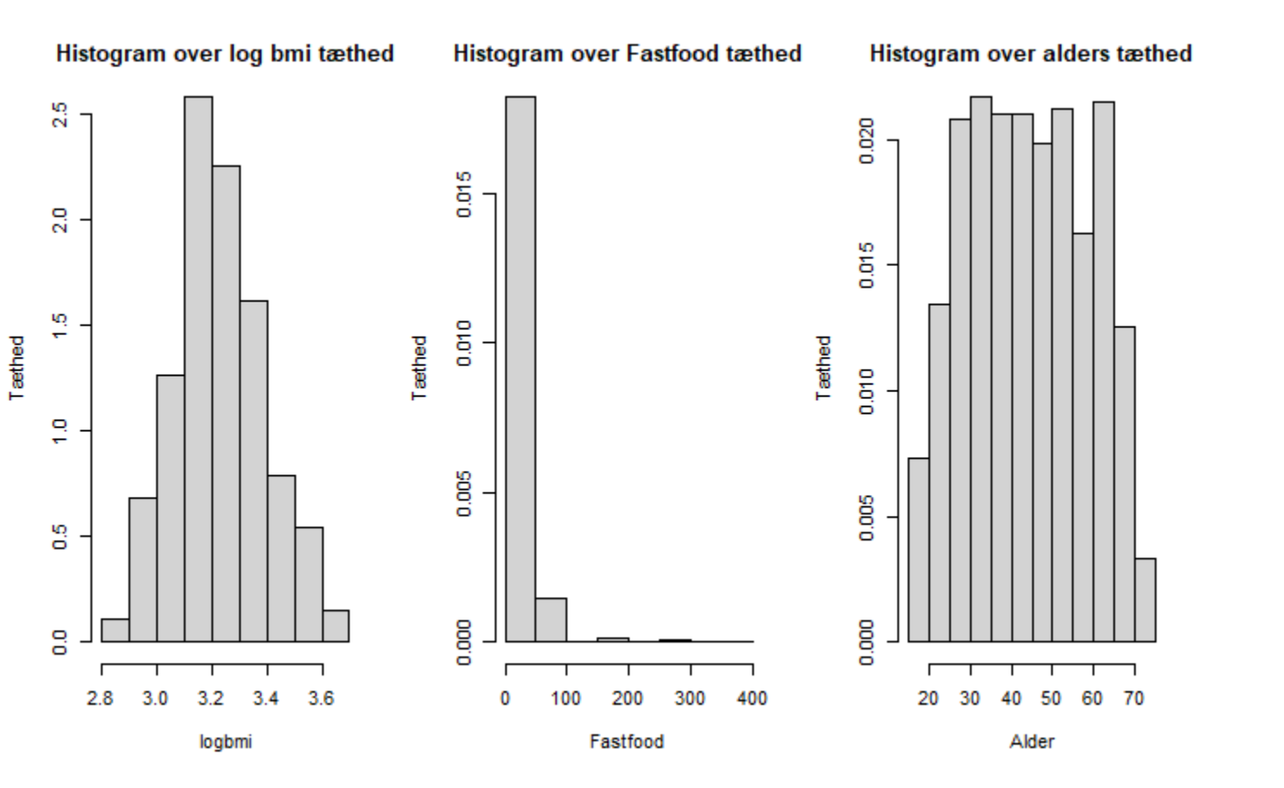
Man kan se at der er en tendens til at mennesker med højere BMI typisk er ældre, samt at der med højere BMI også findes flere ekstreme tilfælde af fastfood.

Nedenfor ses 2 boxplots hvor segmenteringen følger fastfood.



Med disse boxplots kan man se at det generelt er den yngre del af stikprøven der spiser mest fastfood. Det ses også at med højere fastfood forbrug er spredningen af BMI en del stører og generelt højere bmi, hvilket kunne indikere at fastfood har en øgenede indflydelse på bmi. Dog bør der kigges på flere faktorer for at konkludere, da det nok ikke kun er fastfood eller alder der har indflydelse på bmi.

Nedenfor ses histogrammer for bmi, fastfood og alder.



Histogrammerne viser en højreskæv normalt fordelt logbmi. Histogrammet viser også at fastfood generelt ikke er en daglig dags ting, men mere til seværdigheder. Dog er der sjælnde tilfælde hvor fastfood er tæt på en dagligdags ting. Stikprøven Lader også til at følge en semi uniform fordeling. Hvis vi ser bort fra helt unge og helt gamle mennesker, har vi nogenlunde lige mange mennesker i alders grupperne fra slut 20'erne til start 60'erne, pånær slut 60'erne.

b) Multipel lineær regressionsmodel

Det forudsættes at residualerne er normal fordelt. Denne forudsætning er vigtig for den multi linære regression.

Multipel lineær regressionsmodel:

- $logbmi_i = \beta_0 + \beta_1 \times alder_i + \beta_2 \times fastfood_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$

c) Model parametre

Bruger R til udregning af modellens parametre.

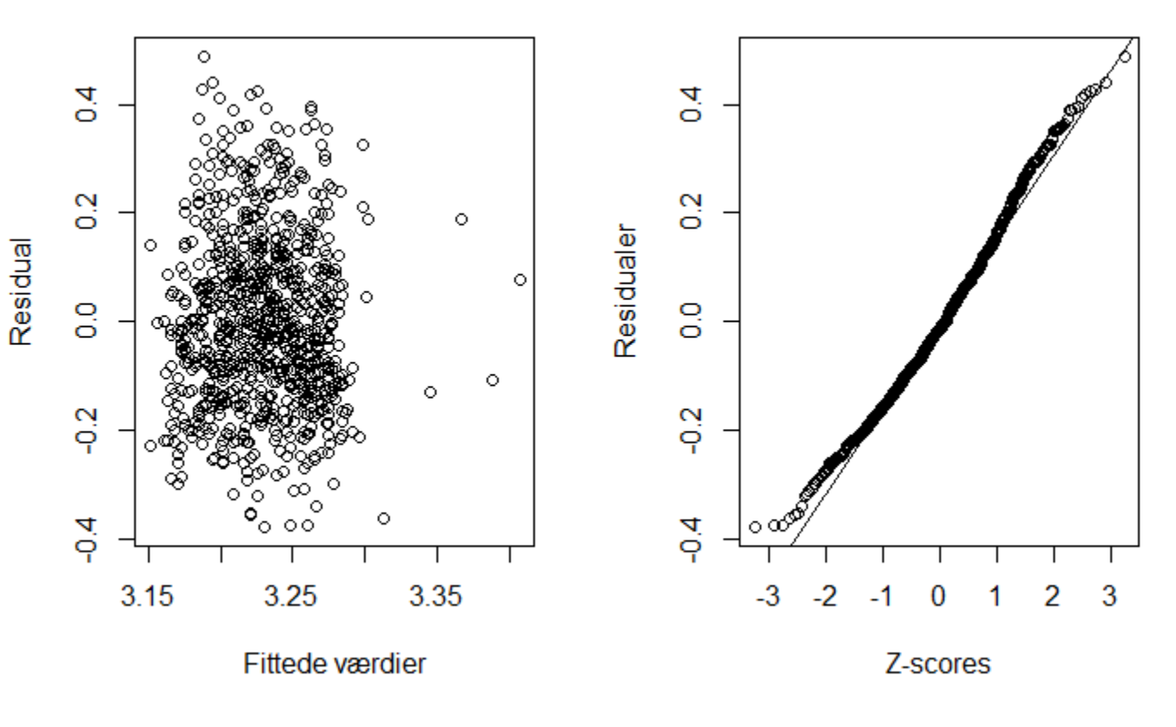
```
Call:
lm(formula = logbmi ~ age + fastfood, data = D_model)

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
```

- $\beta_0 = 3.1124298$ er intercepten og angiver en start kontext for tilpasning for de faktorer der ikke betragtes i modellen.
- $\beta_1 = 0.0023744$ er kooficienten der angiver indfyldelses størelsen af første forkarlings variable. (I dette tilfælde hvor stor en indfyldelse alder har på modellen)
- $\beta_2 = 0.0005404$ er kooficienten der angiver indflydelses størelsen af anden forklarings variable. (I dette tilfælde hvor stor en indflydelse fastfood har på modellen).
- $\sigma = 0.1573$ er standard residual fejl.

Denne fortolkning burde ikke bruges hvis forklarings variablerne er kollineare. Men i dette tilfælde ser det ud til at være fint. Det kan blandt andet ses ved at finde correlationen mellem forklaringsvariablerne (vises senere) og ved P værdien i tidligere R output.

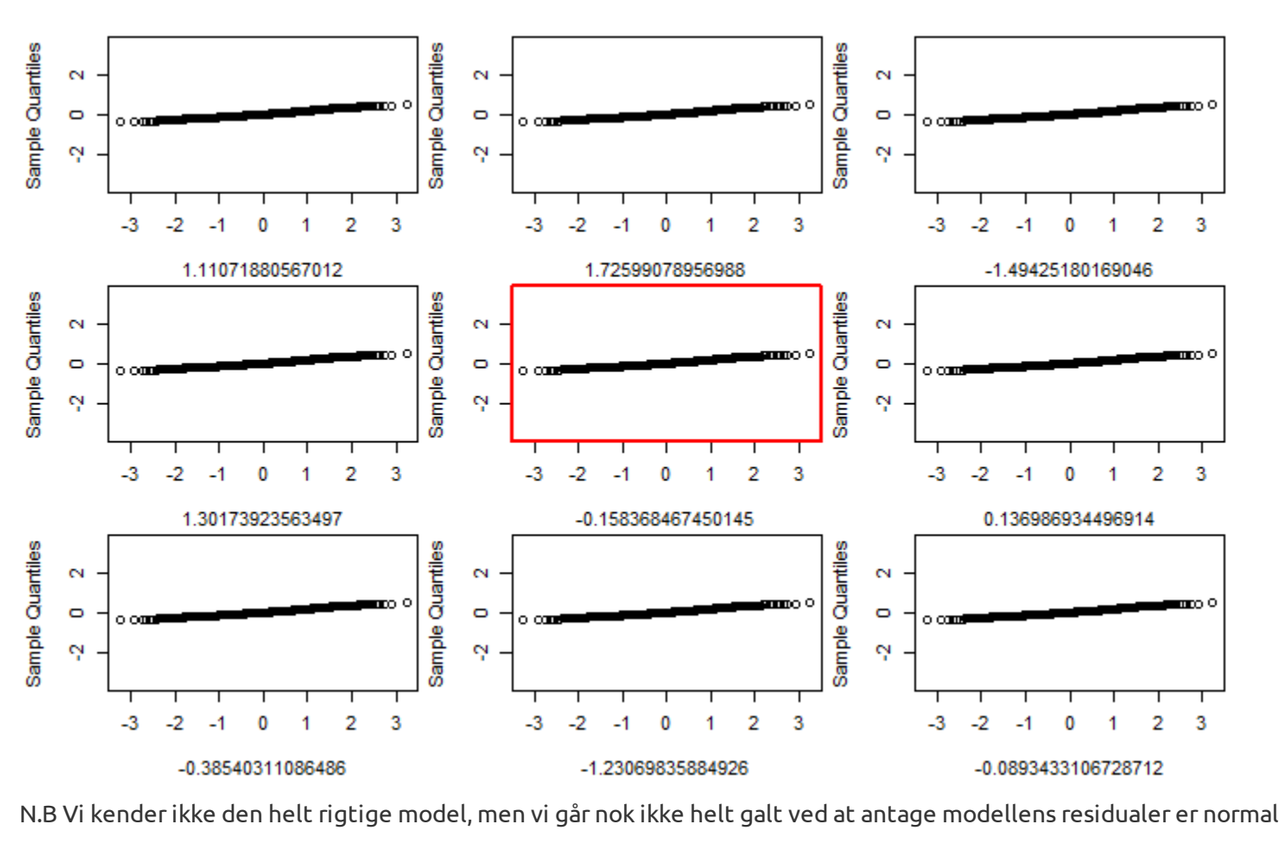
d) Model kontrol

Det er vigtigt at residualerne er normal fordelt for at den multipel liniære regression passer. Nedenfor ses et plot af det fittede værdier imod dets residual. Hvis der ses bort fra de enkelte ekstreme residualer, er der ikke umiddelbart noget system over residualerne, samt at residualerne følger qq plottet rimelig tæt.



Udover ovenstående, kan vi også teste med et "find wally" experiment. Nedenfor ses 9 qq plots som experimentet er lavet på. Experiementet forsøger at udpege en eventuel afvigning fra en normal fordeling. Der kan ikke umiddelbart ses nogen afvigning fra at residualerne er normal fordelt.

```
package 惏牠MESS惏牜 was built under R version 4.0.5
```



N.B Vi kender ikke den helt rigtige model, men vi går nok ikke helt galt ved at antage modellens residualer er normal fordelt.

e) Alders koefficient konfidensinterval

Bruger formel:

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

Værdier aflæst fra tidligere R output:

- $\hat{\beta}_1 \approx 0.0023743602$
- $\hat{\sigma}_{\beta_1} = 0.0003889714$
- $t_{1-\alpha/2} = 1.9628022725$
- $0.0023743602 \pm 1.9628022725 \times 0.0003889714$

Hermed konfidens intervallet: [0.001610886, 0.003137834] Det er det samme resultat der fås med R's funktion 'confint'.

```
                   2.5 %       97.5 %
(Intercept) 3.0744463234 3.1504132672
age         0.0016108861 0.0031378342
fastfood    0.0002003159 0.0008803957
```

f) Hypotesetest

Givet null-hypotesen:

- $H_0 : \beta_1 = 0.001 \implies \beta_1 - 0.001 = 0$

Den alternative hypotese:

- $H_1 : \beta_1 - 0.001 \neq 0$

Givet signifikansniveauet $\alpha = 0.05$ kan vi teste hypotesen ved at udregne p værdien.

Beregner test størelsen med formel:

$$t_{obs,\beta_1} = \frac{\beta_1 - 0.001}{\sigma_{\beta_1}}$$

Derefter kan test størelsen bruges til at slå op i t fordelingen for at finde p værdien med formel:

$$p = 2P(T > |t_{obs,\beta_1}|$$

Finder test størelsen for indflydelses størelsen af alder ved værdien 0.001: $t_{obs,\beta_1} = \frac{0.0023743602 - 0.001}{0.0003889714} = 3.5333194163$

Slår op i t fordelingen med test størelsen, 837 frihedsgrader og får: $p = 2P(T > 3.5333194163|) = 0.0004328512$

P værdien er meget lav, hvilket betyder der er stærk evidens imod hypotesen. Vi afviser derfor hypotesen med et signifikans niveauet på 0.05, da p værdien er mindre.

g) Backward selection

Med backward selection forsøger vi at reducere modellen ved at ignorere variable der ingen indflydelse på modellen har. Vi starter med modellen som beskrevet i opgave del b.

Nedenfor ses udregninger med R.

Correlation:

De følgende 3 correlationer er hhv. bmi-fastfood , bmi-age , fastfood-age og er udregnet med R's cor funktion.

```
[1]  0.06064365  0.16724871 -0.28567253
```

I blandt de 3 variable har ingen af dem en betydelig klar correlation, derfor kan vi ikke på det grundlag fjerne dem. Hvis der havde været en variable der angav minutter siden fødsel, havde den haft en høj correlation med alder og kunne derfor ignoreres.

Confidens intervaller

De følgende confidens intervaller er for den multipel lineære regressions models parametre.

```
                   2.5 %       97.5 %
(Intercept) 3.0744463234 3.1504132672
age         0.0016108861 0.0031378342
fastfood    0.0002003159 0.0008803957
```

På R's output ses det at både alder og fastfood har en meget lille indfyldelse på logbmi, der må altså derfor være andre faktorer der spiller ind, som eventuelt kunne være: køn, fysisk aktivitet, gener mm.

MLR Summary

Det følgende R output er en info summering af regressions modellen.

```
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
```

Givet signifikans niveauet 0.05 ses det at alle variablerne er statistisk signifikant med de lave p-værdier, hvilket betyder at der er evidens imod ide'en om at variablerne alder og fastfood ikke har en betydning på logbmi. Af denne grund påvises det altså at alder og fastfood har en statistisk effekt på logbmi, og bør derfor ikke fjernes fra modellen. Med andre ord har alder og fastfood en unik indflydelse på logbmi.

Slut model

Hermed den endelige model med dets parametre:

- $logbmi_i = 3.0744463234 + 0.0023744 \times alder_i + 0.0005404 \times fastfood_i + \epsilon_i, \epsilon_i \sim N(0, 0.1573^2)$

h) Prædiktioner

I dette afsnit er R's predict funktion brugt til beregninger følgende.

Prædiktion

```
         fit      lwr      upr
841 3.236993 2.927972 3.546015
842 3.210875 2.901802 3.519949
843 3.232245 2.923231 3.541258
844 3.232245 2.923231 3.541258
845 3.229870 2.920857 3.538883
846 3.229641 2.920601 3.538681
847 3.211670 2.901898 3.521443
```

Confidence

```
         fit      lwr      upr
841 3.236993 3.225973 3.248014
842 3.210875 3.198481 3.223270
843 3.232245 3.221437 3.243052
844 3.232245 3.221437 3.243052
845 3.229870 3.219089 3.240651
846 3.229641 3.218106 3.241176
847 3.211670 3.187454 3.235886
```

Vurdering

Givet test data er modellens prædiktion. Alle test punkternes fittet værdi (værdi tilpasset med indflydelse fra alder og fastfood) passer i både prædiktion og confidence intervallerne. Confidence intervallet reflectere middelværdien hvor prædiktions intervallet reflectere en enkelt værdi. Hvis vi udregner den rigtige bmi ved at tage værdien til exponential funktionen af de første test intervaller fås følgene:

```
Prædiktion: [18.6896893473547 , 34.6748624722567]
```

```
Confidence: [25.1780604941437 , 25.739171129542]
```

Det giver meget god mening at modellen prædiktere at en eventuel næste værdi observeret vil være mellem 18 og 34 bmi. Det er måske et lidt stort interval, men det giver meget god mening. Udover prædiktionen viser den også at middelværdien burde ligge mellem 25,16 og 25,73 hvilket også er meget plausibelt.

Givet residualernes fejl ser ud til at være normalt fordelt med en konstant variance som vist i opgave del d, kan modellen umiddelbart godt bruges til at prædiktere.