Danmark Tekniske Universitet

02323 Introduktion til statistik

Projekt 2: BMI

09.november.2020

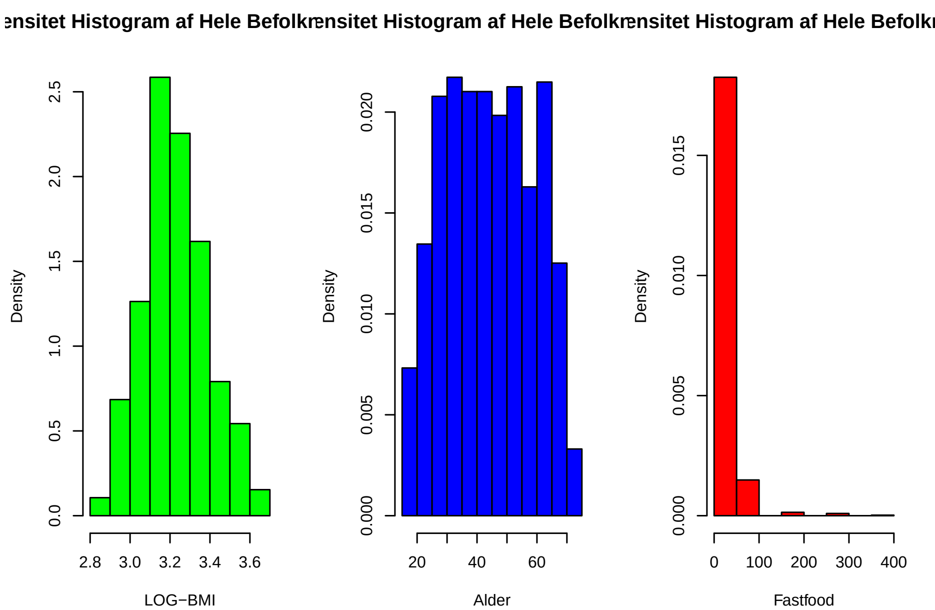
Volkan Isik, s180103

a) Lav en kort deskriptiv analyse og opsummering af data for variablene logbmi, age og fastfood. Inkluder scatterplots af logaritmen til BMI mod de to andre variable, samt histogrammer og boxplots af alle tre variable. Der skal også være en tabel med opsummerende størrelser, som for hver variabel inkluderer antal observationer, gennemsnit, standardafvigelse, median, samt 25%- og 75%-fraktiler.

Der er i alt 847 respondenter med 4 forskellige variabler og 3388 observationer:

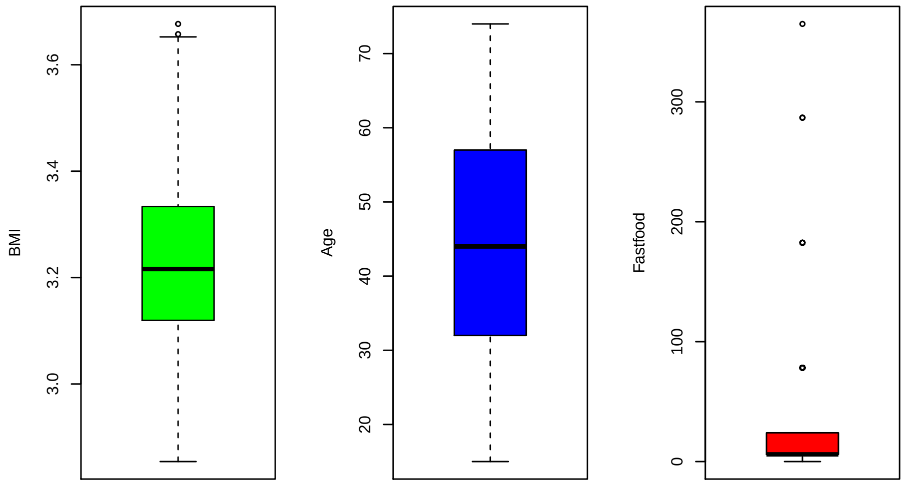
- id: Respondentens vægt
- bmi: Respondentens bmi i højde/kg^2. Kvantativ variable.
- age: Respondentens alder. Kvantativ variable.
- fastfood: Respondentens fastfoor-forbrug. Kontinuert variable.

Histogramer

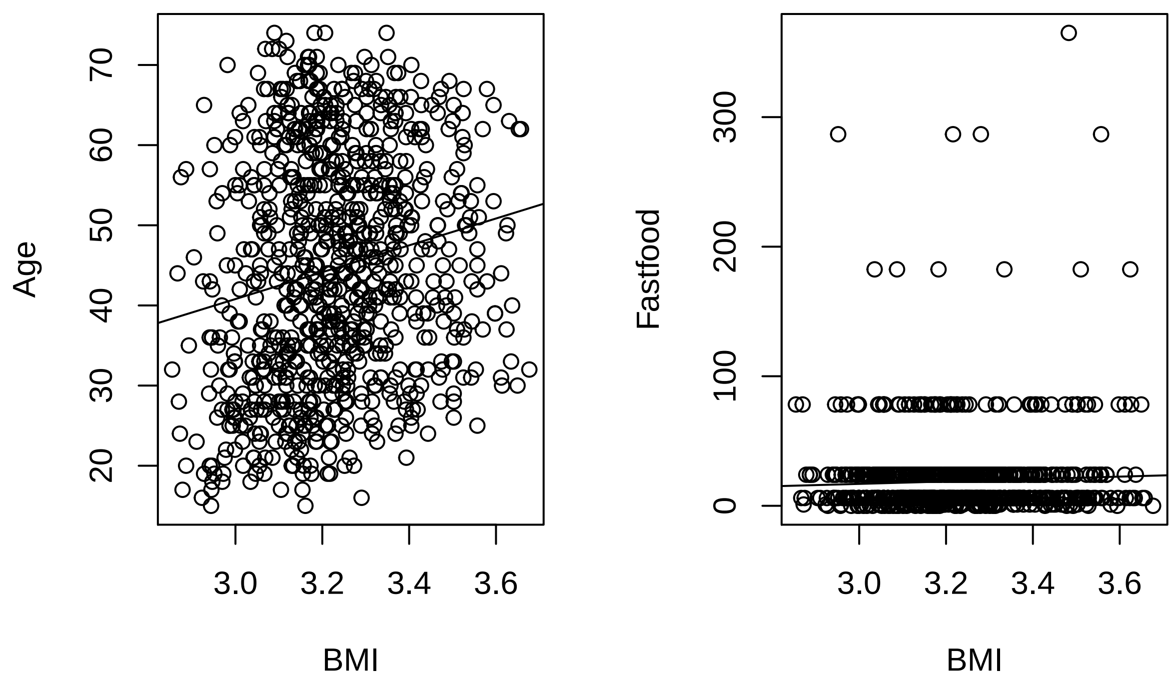


De tre histogram viser fordelingen af observationerne. Log-Bmi histogramen viser at observationerne er normalfordelt(logtransformeret). Alder histogrammen er normalfordelt og der findes mange forskellige observationer i forskellige alder gennemsnit ligger omkring 45. Fastfood observationerne har en højre skæv fordeling. Man kan se at der er en del ekstremer i denne fordeling

Boxploter



Bmi boxplotten bekræfter normal fordeling hvor man kan se medianen som står midten af sin boks.Den har en nedrekvartil ca. 3,1 og en øvrekvartil lidt over 3,3. Alder bloxplotten har normalfordeling.Medianen står ved 45, nedrekvartil ved 32 og øvre kvartil ved 57. Fastfood ser ud til at være højre skæv og har en del ekstremer.



Scatterplotten undersøger vi om der er sammenhæng imellem BMI og de andre variabler alder og fastfood.

Der ser ud til at der er linær sammenhæng imellem BMI og alder. Der kan ikke ses en sammenhæng i mellem BMI og Fastfood. De fleste observationer i fastfood er ligger omkring 0 - 10. Det betyder at populationen har lav fastfood tal for de meste.

```
           n      mean          var         sd       Q1        Q2        Q3
BMI      847  3.228495 2.571927e-02  0.1603723  3.11962  3.216102  3.333602
AGE      847 44.622196 2.112022e+02 14.5327991 32.00000 44.000000 57.000000
FASTFOOD 847 19.044628 1.066103e+03 32.6512392  6.00000  6.000000 24.000000
```

Tabellen viser de nøjagtige tal fra populationen . Variansen er meget lave i de forskellige variabler. Man kan se at nedrekvartil og medianen er 6 for fastfood hvilket betyder de meste af befolkning har fastfood tal omkring dette. Man kan se de helt store forskel i mellem gennemsnittet og medianen for fastfood hvilket bekræfter den højreskæve fordeling.

b) Opstil en multipel lineær regressionsmodel med logaritmen til BMI som responsva- riabel (Y i), og med alder og fastfood-forbrug som forklarende variable (hhv. x 1,i og x 2,i). Husk at angive forudsætningerne/de statistiske antagelser for modellen. (Se bemærkning 5.6, ligning (6-1) og eksempel 6.1).

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon \sim N(0, \sigma^2) \text{ and i.i.d where } i = 1, ..., n$$

Hvor:

- $\mathrm{Y_i}$ er log BMI for måling i
- $\mathrm{x_{1,i}}$ er alder for måling i
- $\mathrm{x_{2,i}}$ er fastfood for måling i

c) Estimer modellens parametre, som består af regressionskoefficienterne, her kaldet β_0 , β_1 , β_2 , og residualernes varians, μ^2.

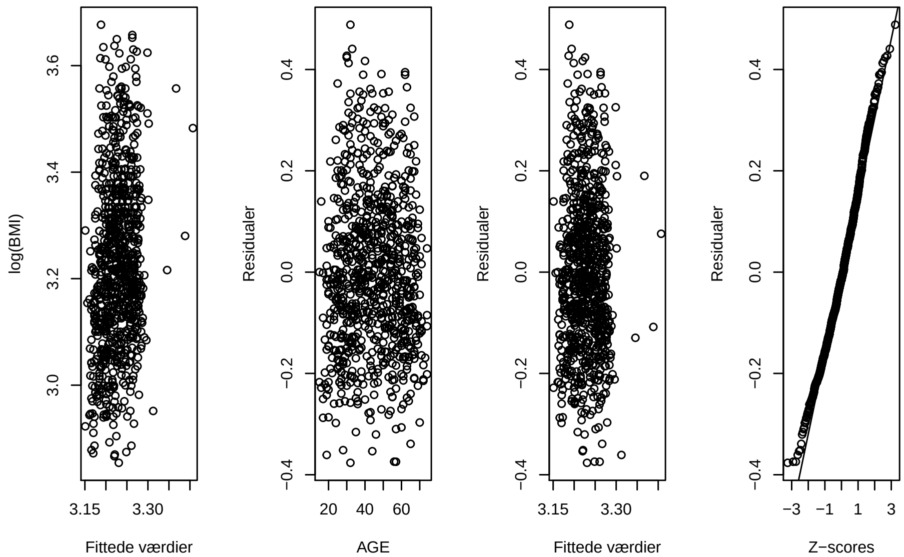
```
Call:
lm(formula = logbmi ~ age + fastfood, data = D_model)

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
```

Udefra summary kan vi se at:

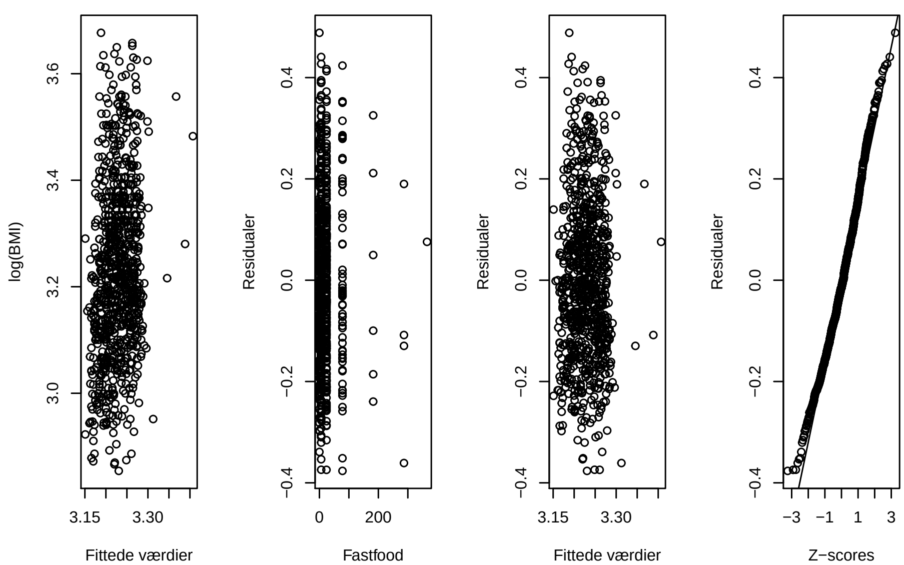
- β_0 $estimat = 3.1124298$ skæring i linjen $\sigma^2 = 0.0193517$
- β_1 $estimat = 0.0023744$ hældning for alder $\sigma^2 = 0.0003890$ variansen for alder
- β_2 $estimat = 0.0005404$ hældning for fastfood $\sigma^2 = 0.0001732$ variansen for fastfood
- residualernes varians, $\sigma^2 = 0.1573$ $Frihedsgrader = 837$

- $R^2 = 0.04487$ Den forklarende variabel: Man kan se at kun 5% af alle observationer kan forklares med denne regression.

d) Foretag modelkontrol for at undersøge, om forudsætningerne for modellen (mo- dellens antagelser) er opfyldte. Benyt de plots, der kan laves ved hjælp af R-koden nedenfor, som udgangspunkt for din vurdering. (Se afsnit 6.4 om residualanalyse).



Graferne viser at der er ingen linær sammenhæng i mellem residualerne og alder.



Man kan igen se at der er ikke noden linærsammmenhæng i mellem residualerne og fastfood obsevationerne.

e) Angiv formlen for et 95% konfidensinterval for koefficienten for alder, her kal- det β_1 . (Se metode 6.5). Indsæt tal i formlen og beregn konfidensintervallet. Benyt derefter nedenstående R-kode til at kontrollere resultatet og til at bestemme konfidensintervaller for de to andre koefficienter i modellen.

Formlen for konfidensintervaller $\hat{\beta_1}$

$$\hat{\beta_1} \pm t_{1-a/2}\sigma_{\hat{\beta_1}}$$

```
## [1] 1.962799
## [1] 0.001610871
## [1] 0.003137929
```

$$0.002374 \pm 1.96278 * 6.104$$

$$[0.001611; 0.003138]$$

f) Man er interesseret i, om β_1 kunne have værdien 0.001. Opstil den tilsvarende hypotese. Anvend signifikansniveauet $\alpha = 0.05$. Angiv formlen for den relevante teststørrelse (se metode 6.4), indsæt tal og beregn teststørrelsen. Angiv fordelingen af teststørrelsen (inkl. frihedsgrader), beregn p-værdien og konkluder.

Hypotesen:

$$H_{0,i} : \beta_i = 0.001$$

$$H_{1,i} : \beta_i \neq 0.001$$

$$t_{obs,\beta_i} = \frac{\hat{\beta_i} - \beta_{0,i}}{\sigma_{\hat{\beta_i}}}$$

$$p - value_i = 2P(T > |t_{obs,\beta_i}|)$$

$$p - value_i = 2P(T > |3.53316|)$$

$$p - value_i = 0.0004329$$

Der blev påvist signifikant effekt.