Danmark Tekniske Universitet

02323 Introduktion til statistik

Projekt 2: BMI

10.november.2020

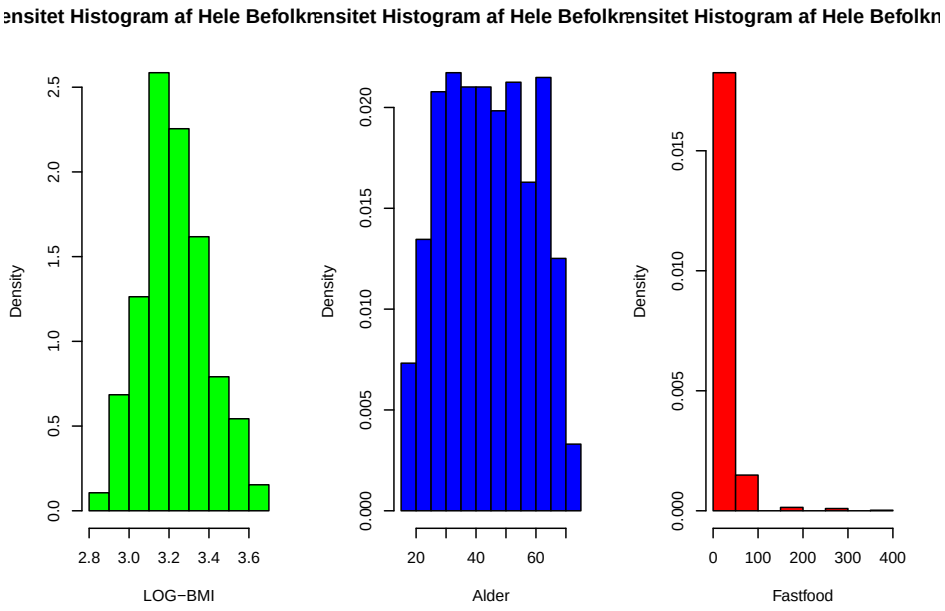
Volkan Isik, s180103

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:

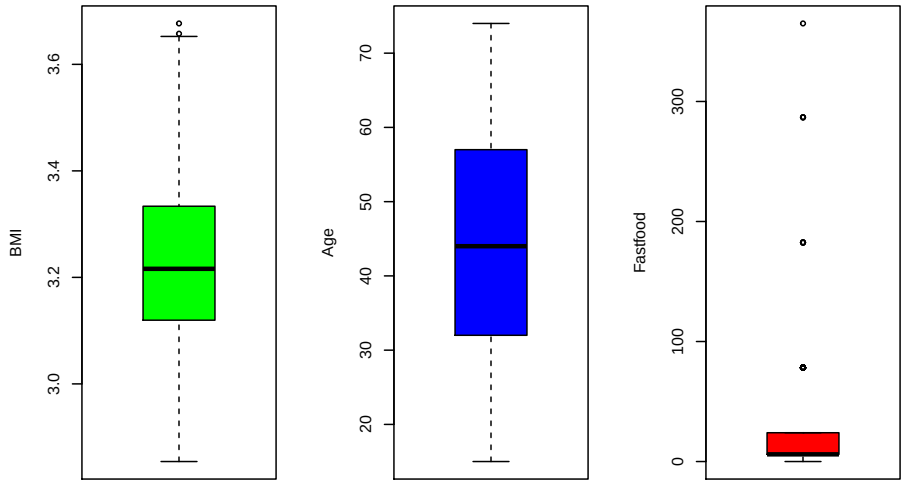
- id: Respondentens vægt
- bmi: Respondentens bmi i højde/kgˆ2. Kvantativ variable.
- age: Respondentens alder. Kvantativ variable.
- fastfood: Respondentens fastfoor-forbrug. Kontinuert variable.

Histogramer

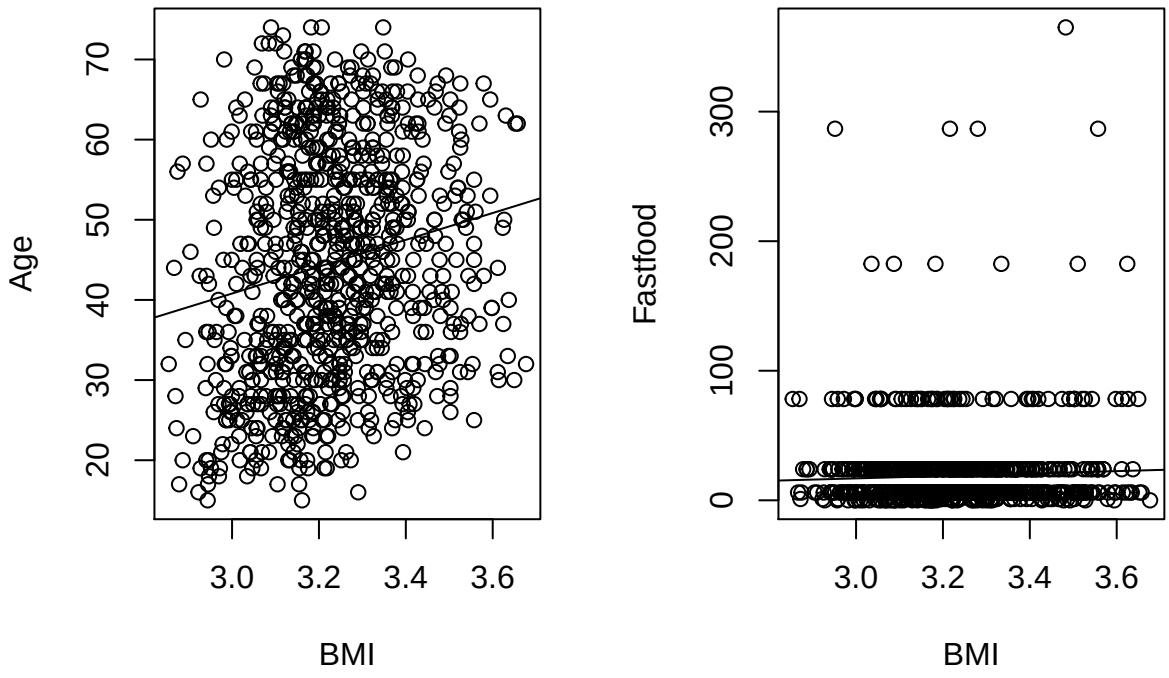


De tre histogram viser fordelingen af observationerne. Log-Bmi histogramen viser at observationerne er normalfordelt(logtransformeret). Alder histogrammen er normalfordelt og der findes mange forskellige observationer i forskellige alder gennemsnit ligger omkring 45. Fastfood observationerne har en højre skæv fordeling. Man kan se at der er en del ekstremer i denne fordeling. Fastfood er blivet ændret fra en kategoriseret til en kontinuert værdi og dette gør at fastfood fordelingen har meget 0 værdier i observationerne.

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:

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

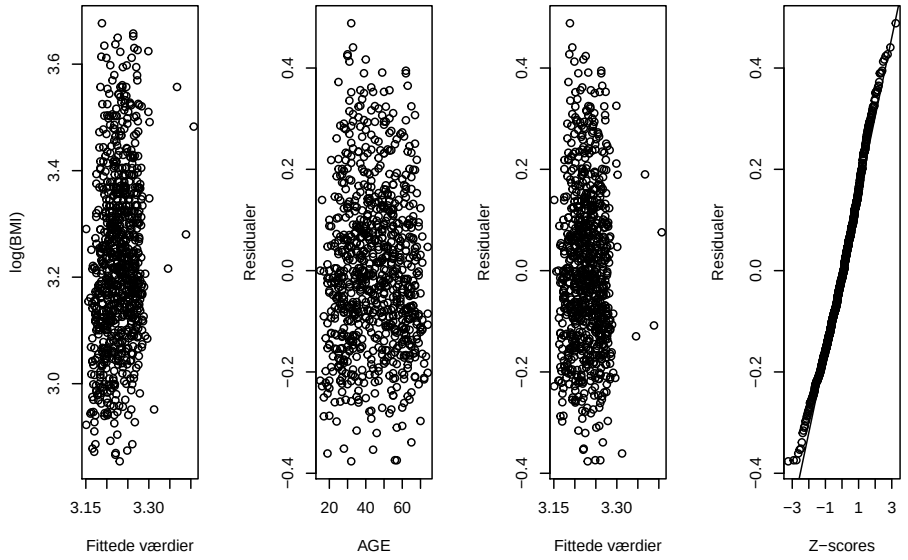
```
Call:
lm(formula = logbmi ~ age + fastfood, data = D_model)

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:

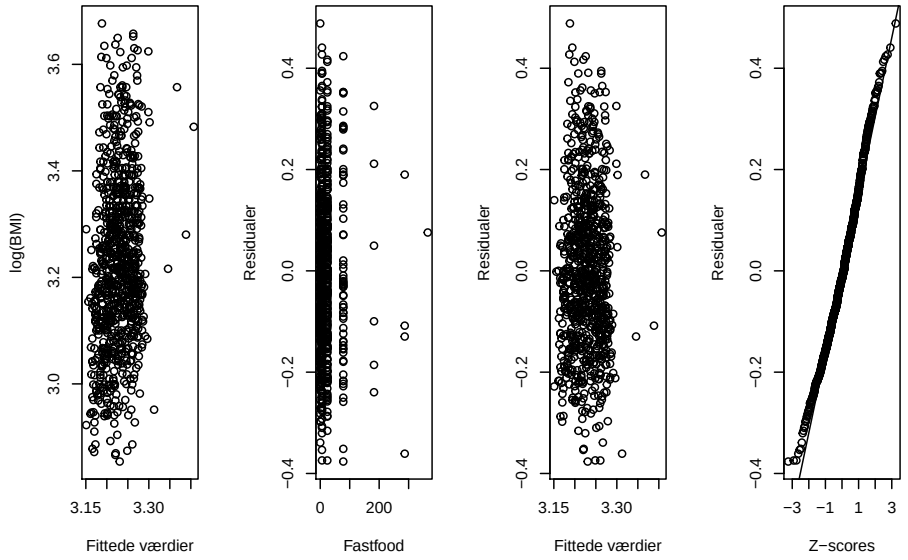
- β_0 *estimat* $= 3.1124298$ skæring i modellen. Når $x_1, i = 0$ og $x_2, i = 0$ så skærer linjen denne punkt i Y $\sigma^2 = 0.0193517$ variansen
- β_1 *estimat* $= 0.0023744$ hældning for alder forklarer sammenhæng i mellem logBMI og alder $\sigma^2 = 0.0003890$ variansen for alder

- β_2 *estimat* $= 0.0005404$ hældning for fastfood forklarer sammenhæng i mellem logBMI og 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).



Residualerne fordeler sig tilfældigt omkring nul og kan ikke konstares et system. Dette kan ses i nummer 2 graf fra venstre hvor alder som x-akse og residualer som y-akse Derfor der er ingen effekt eller afhængighed. Der behøves ikke at tag højde for en afhængighed indtil videre.



Residualerne fordeler sig tilfældigt omkring nul og kan ikke konstares et system. Derfor der er ingen effekt eller afhængighed. Der behøves ikke at tag højde for en afhængighed. Fastfood observationerne ligger omkring 0 på grund af den omformuleret fortolkning af fastfood. Graffen ude til højre viser at residualerne følger en linje hvilket viser analysen er fint. Modellen som er opstillet analyseret godkendt.

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 $\hat{\beta}_1$

$$\hat{\beta}_1 \pm t_{1-a/2}\hat{\sigma_{\beta_1}}$$

$$0.002374 \pm 1.96278 * 0.0003890$$

$$[0.001611; 0.003138]$$

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:

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

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

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

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

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

$$p - value_i = 0.0004329$$

Der blev påvist signifikant effekt.

g) Undersøg ved backward selection om modellen kan reduceres. (Se eksempel 6.13). Husk at reestimere modellen undervejs, hvis der kan foretages reduktion af modellen. Angiv slutmodellen og estimater for dens 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
```

Der er ingen grund til at fjerne nogen af variablerne. P-værdien for age er relativ lavt (under 5%) som viser god signifikant. P-værdien for fastfood under 5% som viser god signifikant. I backward selection skulle vi starte fra mest kompiliseret model og fjerne en variabel af gangen men da vi har signifikant for begge variabler fra start så er der ingen grund til reducering.

h) Tag udgangspunkt i din slutmodel fra forrige spørgsmål. Bestem prædiktioner og 95% prædiktionsintervaller for logaritmen til BMI for hver af de syv observationer i valideringsdatasættet (D test). Se eksempel 6.8, metode 6.9 og R-koden nedenfor. Sammenlign prædiktionerne med de observerede log-BMI værdier for disse syv observationer og lav en vurdering af modellens evne til at prædiktere.

```
SLUTMODEL<-lm(logbmi ~ age+fastfood , data = D_model)
# Prædiktioner og 95% prædiktionsintervaller
pred <- predict(SLUTMODEL, newdata = D_test, interval = "prediction",
level = 0.95)
# Observerede værdier sammen med prædiktioner
cbind(id = D_test$id, logbmi = D_test$logbmi, pred)
```

```
     id   logbmi      fit      lwr      upr
841 841 3.143436 3.236993 2.927972 3.546015
842 842 3.269232 3.210875 2.901802 3.519949
843 843 3.269438 3.232245 2.923231 3.541258
844 844 3.324205 3.232245 2.923231 3.541258
```

```
845 845 3.106536 3.229870 2.920857 3.538883
846 846 3.263822 3.229641 2.920601 3.538681
847 847 3.058533 3.211670 2.901898 3.521443
```

Uderfra de prædiktions intervaller kan vi nu se at de observeret værdier kan falde ind i dem. Intervallerne er ikke for bred som viser at modellen kan prædikteres og siden observationerne falder ind i det så antages der at modellens evne til at prødiktere accepteres.