02402 Statistik

Projekt 2 – BMI-undersøgelse

7. november 2023

Josephine Rosencrone Johnsen, s204183

Indhold

Statistisk analyse 3
- Deskriptiv analyse 3
- Multipel lineær regressionsmodel 4
- Modellens parametre 4
- Modelkontrol 5
- 95% konfidensinterval 6
- Hypotesetest 7
- Backward selection 7
- Prædiktioner 8

Statistisk analyse

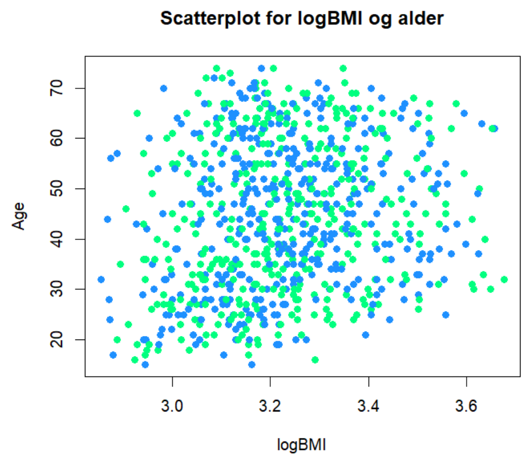
Deskriptiv analyse

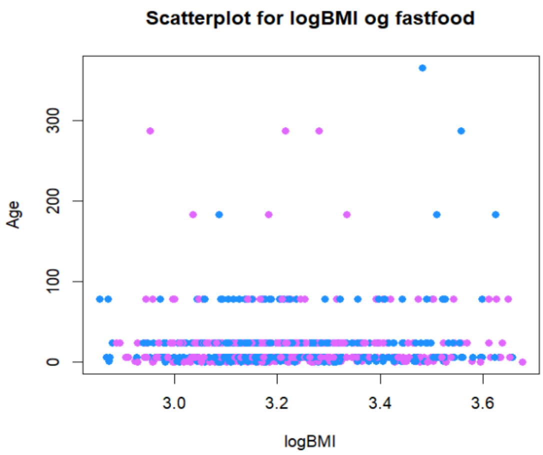
Vi arbejder med et datasæt, hvor vi ønsker at undersøge 3 af dets variable: logaritmen til BMI (logBMI), alder (age) og fastfood-indtag (fastfood). Alle disse variable er kvantitative. Vi starter med at undersøge positionsmålene for disse 3 variable:

Variabel	Antal obs.	Stikprøve-gennemsnit	Stikprøve standard-afvigelse	Nedre kvartil	Median	Øvre kvartil
	n	$(\bar{x})$	*(s)*	(Q_1)	(Q_2)	(Q_3)
Alder	847	44.62	14.53	32	44	57
Fastfood	847	19.14	32.65	6	6	24
logBMI	847	3.23	0.16	3.22	3.22	3.33

Her ses det at der er i alt 847 observationer, hvor der ikke mangler nogle observationer for nogle af variablene. Det ses at respondenterne gennemsnitligt har et BMI på 25, er 45 år gamle, og spiser fastfood 19 dage om året. Vi kan undersøge disse variable yderligere ved brug af plots.

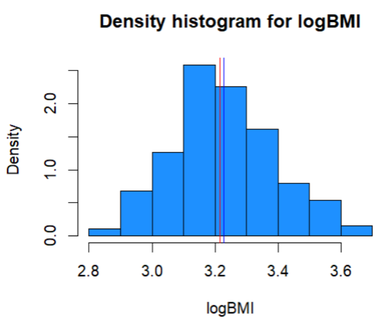
Vi opstiller først scatterplots for logaritmen til BMI mod hhv. alder og fastfood:

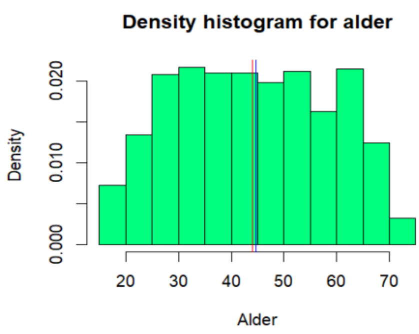


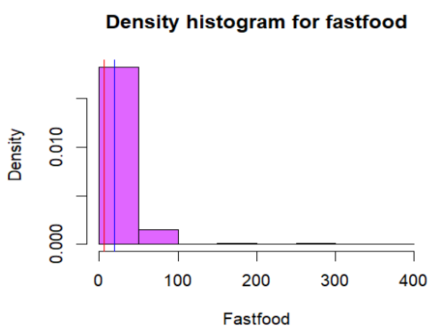


Ovenfor er logBMI markeret med blå, alder med grøn og fastfood med lilla. Det ses på disse plots at der er ingen korrelation mellem logaritmen til BMI og de to variable.

Vi undersøger nu de tre variable ved at kigge på deres density histogrammer:



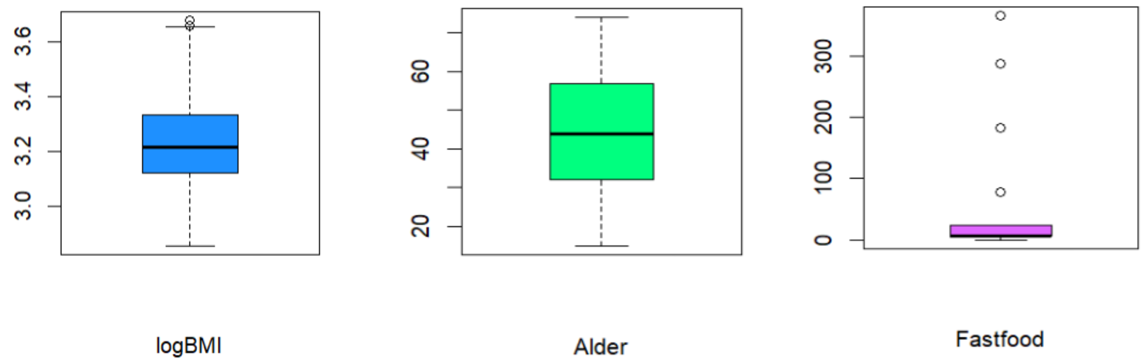




Af histogrammet for logaritmen til BMI fremgår det, at den empiriske tæthed er højreskæv, idet den længste "hale" ligger til højre for midten, samt at stikprøvegennemsnittet (blå streg) ligger til højre for medianen (rød streg). Det ses også at observationerne er nogenlunde samlet omkring midten.

Histogrammet for alder viser at fordelingen af observationer er næsten uniform, idet der ikke er noget tydeligt toppunkt – dog er den stadig højreskæv, men idet stikprøvegennemsnittet (blå streg) ligger til højre for medianen (rød streg). Histogrammet for fastfood viser en meget tydelig højreskæv empirisk fordeling. Det bemærkes at der må være nogle ekstreme observationer, eftersom halen næsten ikke kan ses på histogrammet.

Vi undersøger nu de tre variable ved brug af boksplots:



På boksplottet for logBMI ses det at fordelingen er positivt skæv, idet den øverste halvdel (median til maxværdi) er længere end den nederste halvdel (median til mindsteværdi). Der ses også 2 ekstreme værdier. På boksplottet for alder ser den empiriske tæthed næsten symmetrisk ud, men den er positivt skæv idet den øverste halvdel er længere end den nederste. Her er der dog ingen ekstreme observationer. Boksplottet for fastfood viser at der er 4 ekstreme observationer, og en fordeling som er tydeligt positiv skæv.

Multipel lineær regressionsmodel

Vi ønsker nu at opstille en lineær regressionsmodel med logaritmen til BMI som vores responsvariabel Y_i og variablene alder og fastfood som forklarende variable $x_{1,i}$ og $x_{2,i}$ .For at kunne opstille denne lineære model skal observationerne være uafhængige, hvilket vi så at de er, i vores deskriptive analyse da vi undersøgte korrelation.

Vi opstiller vores model:

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

Her er de deterministiske variable $x_{1,i}$ og $x_{2,i}$, hvor de svarer til henholdsvis alder og fastfood, og de stokastiske variable er Y_i og ε_i. For at der er varians homogenitet, kræver det at residualen ε_i har et gennemsnit på 0, er en normalfordelt stokastisk variabel, samt har en konstant varians.

Modellens parametre

Vi benytter nu R til at estimere modellens parametre, bestående af modellens regressionskoefficienter β_0, β_1 og β_2, samt residualernes varians σ^2. Vi aflæser følgende i R:

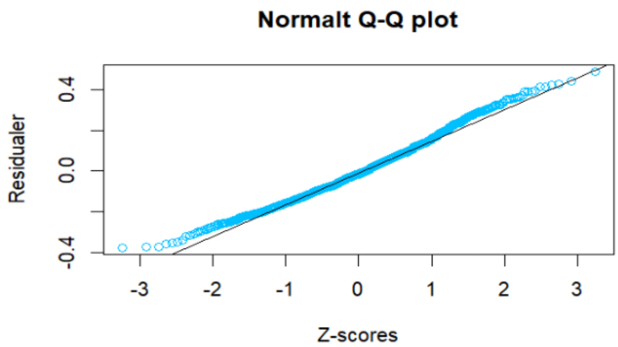
	Estimat af parametre	**Estimat af standardafvigelse**
β_0	$\hat{\beta}_0 = 3.1124298$	$\hat{\sigma}_0 = 0.0193517$
β_1	$\hat{\beta}_1 = 0.0023744$	$\hat{\sigma}_1 = 0.0003890$
β_2	$\hat{\beta}_2 = 0.0005404$	$\hat{\sigma}_2 = 0.0001732$

Derudover aflæses estimatet af residual varians til at være $\hat{\sigma}^2 = 0.1573^2$ med 837 frihedsgrader, samt modellens forklarende varians til at være $R^2 = 0.04487$.

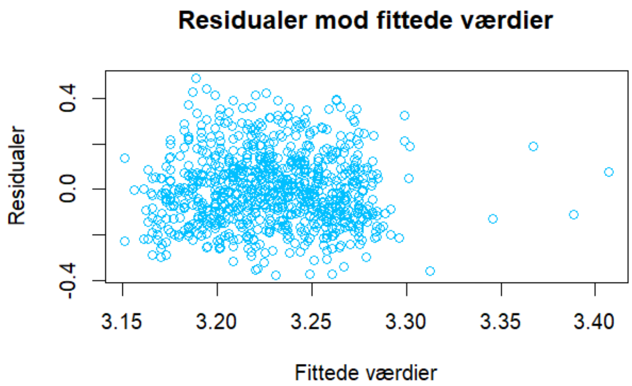
De estimerede parametre er et udtryk for flere ting, herunder den forventede ændring i y når x_i ændres én enhed, effekten af x_i korrigeret for de øvrige variables effekt og effekten af x_i når de andre variable forbliver uændrede (uge 9, slide 19). Her er $\hat{\beta}_1$ estimatet af alders påvirkning på logaritmen til BMI og $\hat{\beta}_2$ er estimatet af fastfoods påvirkning på logaritmen til BMI. Begge er positive og små, og vil derfor ikke have stor effekt på y.

Modelkontrol

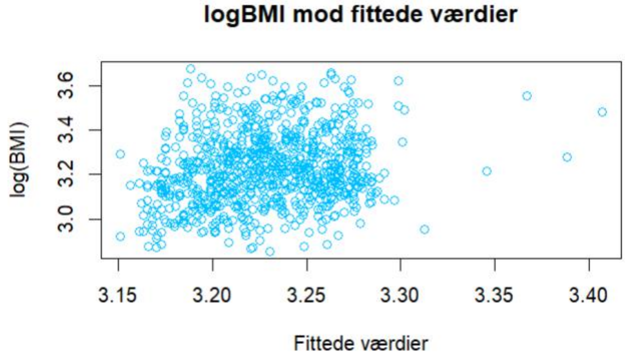
Vi foretager nu en modelkontrol for at undersøge hvorvidt modellens antagelser er opfyldte. Vi starter med at lave et normalt Q-Q plot af residualerne, for at undersøge hvorvidt de er normalfordelte, ligesom antaget i vores model:



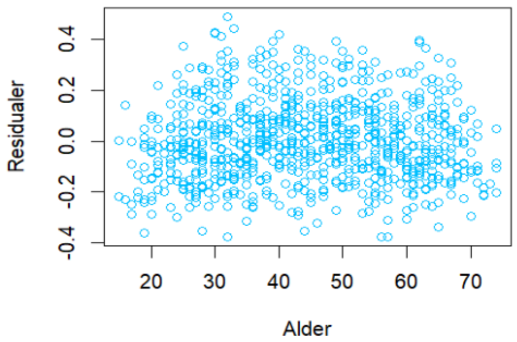
På det ovenstående Q-Q plot ser residualerne relativt normalfordelt ud, idet punkterne fordeler sig nogenlunde på en ret linje. Vi ønsker nu at undersøge hvorvidt residualerne er systematisk fordelte, hvilket kan gøres ved at plotte residualerne mod de fittede værdier $\hat{y}_i$:

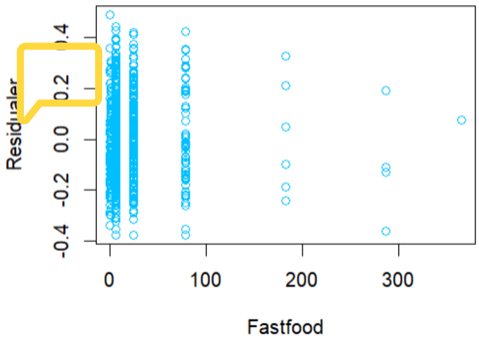


På det ovenstående plot ser fordelingen usystematisk ud og det ligner ikke der er systematisk afhængighed. Vi undersøger også hvorvidt der er systematisk fordeling mellem logaritmen til BMI og de fittede værdier, ved at plotte dem mod hinanden:



Her ser fordelingen for residualerne også usystematisk ud, hvilket betyder at de er indbyrdes uafhængige og overholder vores antagelse om linearitet. Vi plotter nu residualerne mod de forklarende variable (alder og fastfood indtag). Her undersøger vi lineariteten og hvorvidt der er systematiske afvigelser fra lineariteten.





På de ovenstående plots ses det at der ikke er nogle systematiske afvigelser, kun små uafhængige afvigelser, og derfor accepterer vi dette som værende lineært.

95% konfidensinterval

Vi undersøger nu konfidensintervallet, hvor et (1-α) konfidensinterval for β_i er givet ved følgende:

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

Hvor $t_{1-\alpha/2}$ er $(1-\alpha/2)$ kvartilen af en t-fordeling med $n-(p+1)$ frihedsgrader. Vi ønsker at bestemme et 95% konfidensinterval for koefficienten for alder β_1, hvor antallet af frihedsgrader vil være 840-(2+1) = 837. I et 95% konfidensinterval vil α=0.05, hvilket betyder at $t_{1-\alpha/2}$ her vil være $t_{0.975}$. Vi benytter R til at bestemme t-fraktilen, hvilket giver os: $t_{0.975} = 1.96$.

Vi indsætter vores værdier og bestemmer 95% konfidensintervallet for koefficienten for alder β_1:

$$\hat{\beta}_1 = 0.0023744 \pm 1.98 \cdot 0.0003890 = [0.001611\,, 0.003137].$$

Vi benytter R til at bestemme 95% konfidensintervallerne for alle parametre og sammenligner intervallet for β_1 med det vi har beregnet:

	2,5%	97,5%
β_0	3.0744463234	3.1504132672
β_1	0.0016108861	0.0031378342
β_2	0.0002003159	0.0008803957

Her ser vi at det beregnede konfidensinterval for β_1 er det samme som det fundet ved brug af R-funktionen. Derudover ses det af disse konfidensintervaller at ingen af dem vil ramme 0, hvilket betyder at de alle er signifikante. Derudover vil der være lille usikkerhed i vores stikprøve estimater.

Hypotesetest

Vi undersøger nu om β_1 kan have værdien 0.001, hvilket gøres ved at opstille en hypotese:

$$H_{0,1}: \beta_1 = \beta_{0,1}$$

$$H_{1,1}: \beta_1 \neq \beta_{0,1}$$

Hvor $\beta_{0,1} = 0.001$. Vi undersøger denne hypotese ved signifikansniveau $\alpha=0.05$. Vi bestemmer teststørrelsen ved følgende formel:

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

Vi indsætter vores værdier og beregner:

$$t_{obs,\beta_1} = \frac{0.0023744 - 0.001}{0.0003890} = 3.533162$$

Vi kan nu bestemme p-værdien, hvilket gøres ved brug af følgende formel:

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

Vi indsætter vores værdier og beregner:

$$p - værdi_1 = 2P(T > |3.53|) = 0.00045$$

Her ses det altså at vores p-værdi er mindre end α ($0.00045 < 0.05$), hvilket betyder at vi skal forkaste vores nulhypotese. β_1 kan altså ikke antage værdien 0.001.

Backward selection

Vi undersøger nu ved brug af backward selection hvorvidt modellen kan reduceres. Ved backward selection starter vi med den fulde model, og reducerer trinvist de mindst signifikante værdier fra. Vi kigger på p-værdien for hhv. alder og fastfood:

$$p - værdi\ (alder) = 1.58 \cdot 10^{-9}$$

$$p - værdi\ (fastfood) = 0.00188$$

Her ses det at p-værdien for begge er signifikante, og vi ved ud fra deres konfidensintervaller at de begge er signifikante fordi de ikke indeholder 0. Vi undersøger hvorvidt det vil gøre en forskel at reducere modellen, ved at undersøge hvorvidt R^2 bliver større eller mindre.

Vi fjerner fastfood, eftersom den har den største p-værdi og undersøger hvordan dette påvirker R^2. Dette giver os en $R^2=0.03377$, hvilket er lavere end før reduktionen, hvor $R^2=0.04487$. Vi prøver i stedet at fjerne alder, og ser at den nye $R^2=0.00235$, hvilket er mindre end begge de tidligere R^2 værdier. Vi kan derfor konkludere at modellen ikke kan reduceres ved brug af backwards selection, og den endelige slutmodel er modellen vi startede med:

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

Estimaterne af denne model er altså identiske med de tidligere fundne estimater, idet modellen er den samme.

Prædiktioner

Vi ønsker nu at bestemme prædiktionerne og 95% prædiktionsintervallerne for logaritmen til BMI for de 7 sidste observationer i datasættet. Vi benytter R til at gøre dette, og får følgende:

ID	logBMI	Prædiktion	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

På den ovenstående tabel sammenligner vi logBMI og prædiktionerne, som vi kan se, ikke er identiske. Dog ligger logBMI stadig inden for 95% prædiktionsintervallet, hvilket betyder at vi ikke behøver at forkaste vores model, hvis vi blot tager højde for prædiktionsintervallet.