BMI-undersøgelse II
Projekt 2

Mads Yar - s193992

Den 09. november 2021

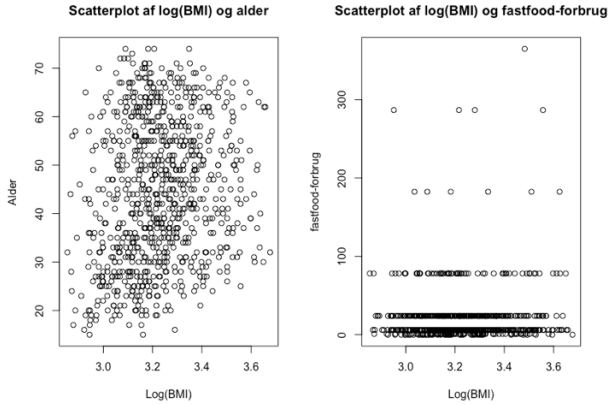
Indholdsfortegnelse

Deskriptiv analyse

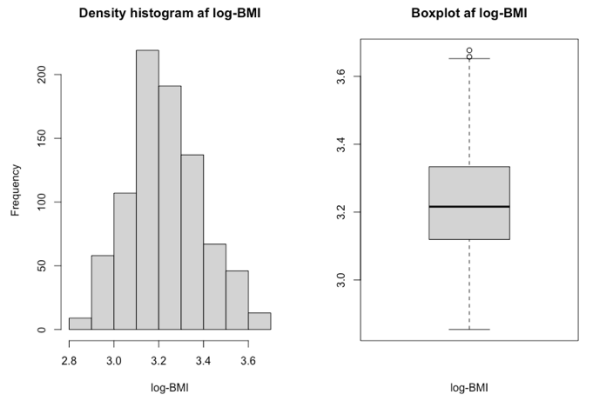
Dette projekt handler omkring Body Mass Index (BMI). Derfor vil der blive analyseret en datamængde, som beskriver forskellige ting vedrørende respondenten og deres BMI. BMI er defineret ud fra en persons højde og vægt med følgende formel:

$$BMI = \frac{vægt}{højde^2}$$

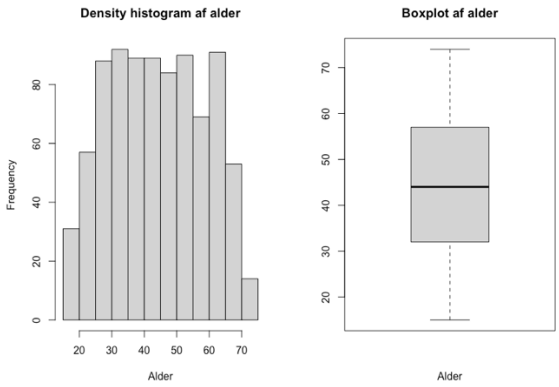
Datasættet består af 847 datapunkter (observationer), og det er inddelt i fire variable, som er: ID, BMI, alder og fastfood-forbrug. ID er et nummer (kvantitativ variabel) mellem 1 og 847, som beskriver hver respondent. Hver respondent har fået deres eget ID-nummer. Alder er en kvantitativ variabel, som beskriver hver respondentens alder. Fastfood-forbrug var tidligere en kategoriserende variabel, men er nu en kontinuert variabel, som beskriver hyppigheden af respondentens besøg ved fastfood restauranter. BMI er en kvantitativ variabel, som er blevet log-transformeret, for at gøre fremtidige beregninger nemmere. Datasættet er komplet, da der er ingen manglende værdier for variablerne.



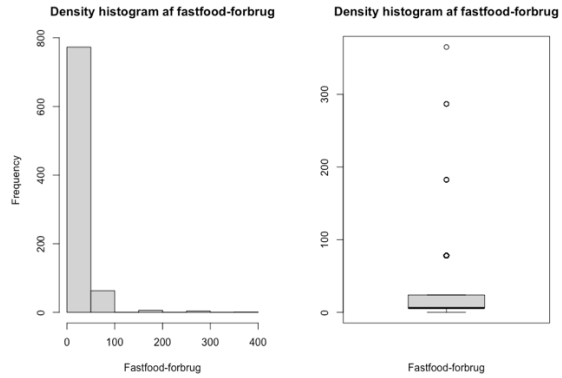
Der er blevet udført scatterplots (i R) af log-BMI sammenlignet med alder og fastfood-forbrug. Her ses det, at der ikke er sammenhæng mellem log-BMI og alder eller log-BMI og fastfood-forbrug, da observationerne er meget spredte i begge tilfælde. Det er især iøjnefaldende ift. BMI og fastfood-forbrug, da man normalt ville associere høj BMI med hyppig indtagelse af fastfood.



Der er blevet udført density histogram og boxplot af log-BMI for at undersøge fordelingen af log-BMI. Den empiriske tæthed for log-BMI er næsten normalfordelt. Dette kan ses på boxplottet, da dataen for log-BMI er næsten symmetrisk, men der er stadig to outliners.



Her kan det ses, at der er varians i datasættet ift. alder, når man kigger på den empiriske tæthed for alder. Mange af aldersgrupperne er stærkt repræsenteret på histogrammet, hvilket giver en stor spredning i datapunkterne. Dette er dog ikke understøttet i boxplottet for alder, da dette viser en symmetrisk fordeling af dataen ift. alder.



Det kan ses på histogrammet at datapunkterne for fastfood-forbrug er meget spredte med en overvægt af datapunkter i bunden af datasættet. Dette bliver også understøttet i boxplottet, da man kan se at størstedelen af dataen befinder sig under ca. 60. Den store spredning vises i form af de fire outliners/ekstremer.

Nedenstående er en tabel med opsummerende størrelser, som bekræfter ovenstående udmeldinger ift. histogrammer, boxplots og scatterplots.

Variabel:	Antal obs.	Stikprøve gennemsnit	Stikprøve standard afvigelse	Nedre kvartil	Median	Øvre kvartil
	n	$(\bar{x})$	(s)	$(Q1)$	$(Q2)$	$(Q3)$
BMI	847	25.6	4.21	22.6	24.9	28.0
Alder	847	44.6	14.5	32.0	44.0	57.0
Log-BMI	847	3.23	0.16	3.12	3.22	3.33
Fastfood	847	19.0	32.7	6.00	6.00	24.0

Multipel lineær regressionsmodel

Der vil nu blive opstillet en multipel lineær regressionsmodel, hvor der er blevet taget udgangspunkt i den generelle lineære model. Her vil responsvariablen Y_i være log-BMI. De forklarende variabler

x_1 og x_2 vil være hhv. alder og fastfood-forbrug. β_0 vil være skæring med y-aksen. β_1 og β_2 vil være hældningen for x_1 og x_2. ε_i vil være residual med en ukendt varians og en middelværdi på nul. Hermed kan man opstille følgende model:

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

Estimer modellens parametre

Modellens parametre kan estimeres i programmet R. Det er kun de første 840 datapunkter, der vil blive brugt til at estimere parametrene, da de sidste 7 vil blive brugt til at validerer parametrene.

Her blev der fundet frem til følgende med R: $\hat{\beta}_0 = 3.1124$ med en varians på 0.0193^2. $\hat{\beta}_1 = 0.0024$ med en varians på 0.0004^2 $\hat{\beta}_2 = 0.0005$ med en varians på 0.0002^2
De tre ovenstående parametres relation mellem log-BMI og de to forklarende variable blev beskrevet i sidste opgave. Det kan dog ses på $\hat{\beta}_1$ og $\hat{\beta}_2$ at de vil stige langsomt, da de er positive men små.

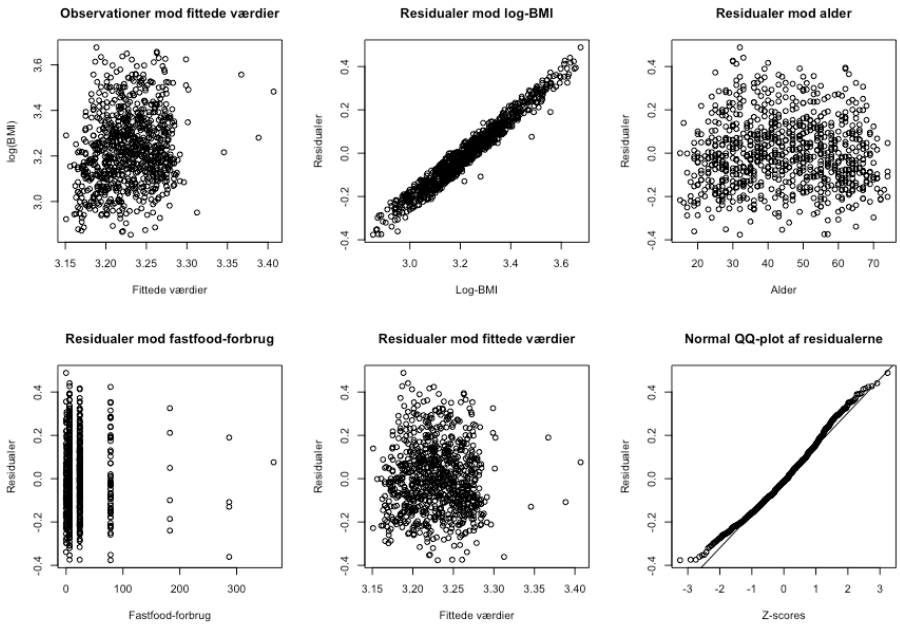
Man kan finde antallet af frihedsgrader brugt til estimatet af residualernes varians og modellens forklarende varians ved hjælp af følgende formel:

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

Det vil sige, at der er blevet brugt 837 frihedsgrader til at estimere residualernes varians og modellens forklarende varians. Disse er hhv. $\hat{\sigma}^2 = 0.1573^2$ og $R^2 = 0.0449^2$.

Modelkontrol

Når man har opstillet en model, skal man også kontrollere om ens model kan valideres. For at gøre dette, kan man benytte sig af forskellige plots udført i R og udføre sin vurdering ud fra dem.



Fra plottet ”Observationer mod fittede værdier” kan man se, at der ikke er et sammenhæng mellem datapunkterne undtaget at størstedelen ligger i intervallet [3.15, 3.30]. Dette viser blot, at der er

chance for at datapunkterne er uafhængige værdier, hvilket vil tyde på at datasættet er normalfordelt. Dette kan også ses på plottet "Residualer mod log-BMI", hvor datapunkterne danner en ret linje, hvilket viser at der er et lineært sammenhæng mellem de fittede værdier og log-BMI. Hvorimod, hvis man kigger på graferne "Residualer mod alder" og "Residualer mod fastfood-forbrug" kan man se at der ingen sammenhæng mellem er mellem hhv. residualer mod fastfood-forbrug eller residualer mod alder. Kigger man på plottet "Residualer mod fittede værdier" kan man se, at der heller ikke er et sammenhæng mellem residualerne undtagen at der er en overvægt af dem i intervallet [3.15, 3.30]. Dette tyder igen på at residualerne er uafhængige værdier, hvilket kun ville tyde på at datasættet er normalfordelt. Kigger man til sidst på "Normal QQ-plot af residualerne" kan man se at de følger den rette linje til en vis grad med få undtagelser. Dette betyder blot at datasættet er normalfordelt. Derfor kan man konkludere at forudsætningerne for modellen er opfyldte.

95% konfidensintervaller for alder

Man kan bestemme 95% konfidensintervallet for alder med følgende formel:

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

Hvor $t_{1-\alpha/2}$ er lig med $(1-\alpha/2)$ med $n-(p+1)$ frihedsgrader. $t_{1-\alpha/2}$ er blevet udregnet i R m ed kommandoen "qt()", hvor $t_{1-\alpha/2}$ blev udregnet til 1.9628.
Da man allerede kender $t_{1-\alpha/2}$ samt $\hat{\beta}_1$ og $\hat{\sigma}_{\beta_1}$ fra to opgaver siden, så kan man udregne konfidensintervallet. Derved fås:

$$KI_{\beta_1} = 0.0024 \pm 1.9628 \cdot 0.0004^2 = [0.0016, 0.0031]$$

Man kan nu bruge R-kommandoen "confint" til at kontrollere ovenstående resultat samt at udregne 95%-KI for resterende koefficienter i vores model. Her kan der konkluderes at 95%-KI for alder er korrekt udregnet. For de resterende koefficienter fås:

$$KI_{\beta_0} = [3.0744, 3.1504]$$

$$KI_{\beta_2} = [0.0002, 0.0008]$$

Hypotesetest

Man vil gerne testes om β_1 kunne have værdien 0.001. Derfor kan man opstille følgende nulhypotese med et signifikansniveau på $\alpha = 0.05$:

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

Formlen for teststørrelsen er givet ved:

$$t_{obs,\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}} = \frac{0.0024 - 0.001}{0.0004} = 3.5$$

Fordelingen af teststørrelsen er en t-fordeling, og antallet af frihedsgrader blev fundet, da modellens parametre skulle estimeres. Her blev de fundet til $DF = 837$. Når man arbejder med hypotesetest, så skal man udregne sin p-værdi, for at se om man kan acceptere eller afslå sin hypotese.
P-værdien kan udregnes således:

$$p_{værdi,i} = 2P\left(T > |t_{obs,\beta_i}|\right)$$

P-værdien er blevet udregnet i programmet R til:

$$p_{værdi} = 0.0004$$

Der gælder at nulhypotese skal afvises, hvis ens p-værdi er mindre end ens signifikansniveau. Eftersom den udregnede p-værdi for denne hypotesetest er mindre end det satte signifikansniveau på $\alpha = 0.05$, gælder der at nulhypotesen skal afvises. Derfor kan β_1 ikke have værdien 0.001 med det valgte signifikansniveau.

Backward selection

Man kan reducere ens model ved at bruge Backward selection. Det betyder at man kan fjerne de variable fra modellen, som ikke er betydelige. Derved kan man simplificere sin model. Der gælder dog at hvis p-værdierne for ens variabler er mindre end sit signifikansniveau, så er variablerne signifikante. De kan derfor ikke frasorteres. Ved hjælp af R, kan man se at:

$$p_{\beta_0} = 2.00 \cdot 10^{-16}$$

$$p_{\beta_1} = 1.58 \cdot 10^{-09}$$

$$p_{\beta_2} = 1.88 \cdot 10^{-03}$$

Man kan her se at variablernes p-værdi er mindre end det givet signifikansniveau på $\alpha = 0.05$. Derfor er alle variablerne signifikante. Dette betyder at slutmodellen er ens med den model, der blev bestemt første gang.

95% prædiktionsintervaler

Der skal bestemmes prædiktioner og 95%-prædiktionsintervaller for log-BMI til hver af de syv observationer i datasættet (D_test), som skulle bruges til validering af slutmodellen. R-funktionen ”predict” er blevet brugt til at udregne prædiktioner og 95%-prædiktionintervallet. Disse kan ses i tabellen under:

ID	Log-BMI	Prædiktioner	Prædiktionsinterval
841.0	3.143	3.197	[2.908, 3.486]
842.0	3.269	3.262	[2.928, 3.596]
843.0	3.269	3.209	[2.937, 3.480]
844.0	3.324	3.209	[2.937, 3.480]
845.0	3.106	3.215	[2.947, 3.482]
846.0	3.264	3.280	[2.993, 3.567]
847.0	3.059	3.064	[2.720, 3.408]

Kigger man på ovenstående tabel, kan man se at både prædiktionerne og de observerede log-BMI’er ligger indenfor prædiktionsintervallet for modellen. Dette er en god ting, da det viser at modellen er præcis ift. udregning af BMI eftersom den rammer indenfor det tilladte interval for alle observationerne. Der er dog en lille afvigelse på nogle af de ovenstående værdier, hvilket kunne havde været mere præcist. Alt i alt, så er dette en præcis model med en signifikant lille afvigelse.