Løsning Eksamen vår 2015, stat100

Oppgave 1

A) $Y_i = \alpha + \beta x_i + e_i$ der e_i-ene er uavhengige og normalfordelt $(0, \sigma)$, der Y_i er kalorimengde og x_i er fettmengde for hamburger nummer i, i = 1, 2, . . .18.
β er forventet (gjennomsnittlig) økning i kilokalori (pr 100 gram) for hver gang fettprosenten øker med 1.
α er forventet (gjennomsnittlig) kalorimengde for hamburgere uten fett, dette er meningsløst av to grunner, slike hamburgere finnes neppe og våre data er ikke hentet fra områder nesten uten fett.
σ er et mål for variasjon (i betydning standardavvik) i kalorimengde for alle hamburgere med samme fettprosent.

Estimater:
$$\hat{\alpha} = 126{,}92$$
$$\hat{\beta} = 10{,}25$$
$$\hat{\sigma} = 33{,}08$$

B) $R^2 = \dfrac{\text{SSRegression}}{\text{SSRegression} + \text{SSError}} = \dfrac{\text{SSFett}}{\text{SSFett} + \text{SSResiduals}} = \dfrac{98172}{98172 + 17507} = 0{,}849$
I dette datasettet er det slik at omtrent 85 % av variasjonen i kalorier, kan forklares av variasjonen i fettprosent.

C) 95 % KI er gitt ved $(\hat{\beta} \pm t_{0,025,16} SE(\hat{\beta})$, eller (10.25 ± 2,12*1,082) = (7,96; 12,54)
Vi ser at 9 ligger inne i intervallet, dermed har vi ikke klart å påvise noen motsetninger med ekspertene. Vi kan ikke forkaste H_0 på 5 % nivå der H_0: $\beta = 9$ mot H_1: $\beta \neq 9$.

D) Generelt: Skal teste $H_0: \theta = \theta_0$ mot H_1: $\theta \neq \theta_0$, for en eller annen parameter θ på nivå α. Dersom du lager et $(1 - \alpha)$*100% KI for θ og dette inneholder θ_0 kan du ikke forkaste H_0. Omvendt: Hvis intervallet ikke inneholder θ_0, så kan du forkaste H_0 og tro på H_1. Dermed blir et konfidensintervall en samling av nullhypoteser som ikke kan forkastes.

Spesielt her vi lager et 95 % KI for β. Dersom du vil teste H_0: $\beta = \beta_0$ mot H_1: $\beta \neq \beta_0$, på nivå 0,05, og lager 95% KI for β og dette inneholder verdien β_0, kan du ikke forkaste H_0.

E) $\hat{Y} = \hat{\alpha} + \hat{\beta} * 16{,}6 = 126{,}92 + 10{,}25 * 16{,}6 = 297{,}07$
95 % PI for antall kilokalorier er gitt ved:
$$\hat{Y} \pm t_{0,025,16}\hat{\sigma}\sqrt{1 + \frac{1}{n}} = (297 \pm 2{,}12 * 33{,}08\sqrt{\frac{19}{18}}) = 297 \pm 72 = (225;\ 369)$$
Generelt er bredden: $2t_{0,025,16}\hat{\sigma}\sqrt{1 + \dfrac{1}{n} + \dfrac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$, men når ny x er lik gjennomsnittet faller siste brøk under rottegnet ut og vi får minimal bredde.

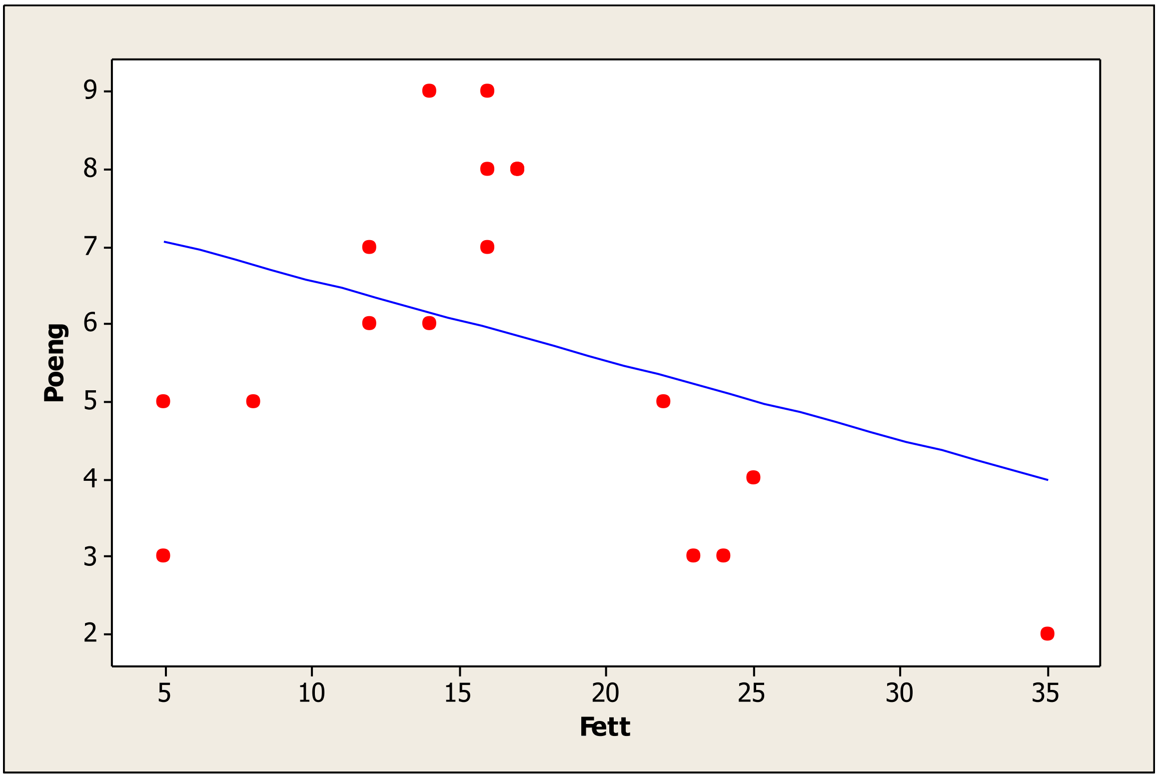
F) Residual er:

Observasjon – Tilpasset verdi,
For første hamburger blir det:
289 – (126,92 + 10,25*17) = 289 - 301 = -12.

Denne observasjonen inneholder 12 kilokalorier mindre enn hva vi anslår den til ut fra fett-inneholdet.
Kan ikke se noe fra figur 1 eller figur 2 som tyder på alvorlige brudd med modellantagel-sene. Det være seg linearitet, konstant varians eller normalfordeling.

G Se på figur 3, to fenomener synes klare. Linja har nesten ikke fall eller stigning, der-med betyr det at protein ikke har særlig betydning. I tillegg er R^2 nesten 0, dermed er svært lite av den observerte variasjon i kalori-innehold forklart gjennom variasjon i protein. Sam-menhengen mellom protein og kalori er derfor heller ikke signifikant (p-verdi er 0,826).
Hvis $\beta = 0$, da er $P(|\hat{\beta}| \geq 0{,}7815) = 0{,}826$. Det er dette som er definisjonen av en p-verdi.

H En rett linje kan ikke fange opp dette utsagnet: *Folk liker hamburgere med noe fett, men ikke for mye.*
En må ha et toppunkt omtrent mellom maksimal og minimal fettverdi. Figur 4 viser at en har negative residualer ved lave og høye verdier av tilpasset poeng, mens de er positive ved middels verdier. Positive residualer tyder på at folk gir mer poeng enn det en lineærmodell skulle tilsi, og dette gjelder da for middels feite hamburgere. For de magre og de feite gir modellen laver poeng enn det som lineærmodellen tilsier. Altså må det være en ikke-lineær sammenheng ute og går, som har et toppunkt på middels feite hamburgere. For eksempel vil muligens en annengradskurve fang opp sammenhengen. Under ser du det eksakte spredningsplottet, du trenger bare i skissere sånn omtrent hvordan det ser ut.



Opp-gave	**A**	**B**	**C**	**D**	**E**	**F**
1				O		
2	O					
3						O
4		O				
5		O				
6						O
7			O			
8						O
9	O					
10	O					
11	O					
12						O
13	O					
14					O	
15		O				
16					O	
17						O
18				O		
19		O				
20		O				