

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i: STK1110 — Statistiske metoder og dataanalyse 1.

Eksamensdag: 10. desember 2010.

Tid for eksamen: 14.30–18.30.

Oppgavesettet er på 4 sider.

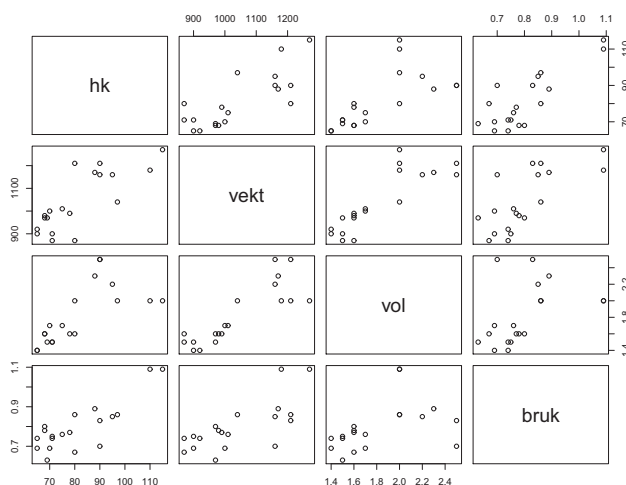
Vedlegg: 1) Data over bensinforbruk for 19 personbiler.
2) Tabell for standardnormalfordelingen.
3) Tabell for t -fordelingene.
4) Tabell for kji-kvadrat fordelingene.

Tillatte hjelpemidler: Godkjent lommeregner.
Formelsamling for STK1100 og STK1110.

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1

I denne oppgaven skal vi se nærmere på hvordan bensinforbruket til en bil avhenger av ulike tekniske data for bilen. I vedlegg 1 er det gitt tekniske data og informasjon om bensinforbruket (i liter per mil) for 19 ulike personbiler. (Dataene er tatt med for fullstendighetens skyld, og det er ikke nødvendig å se på dataene for å løse oppgaven.) De tekniske dataene er antall hestekrefter for motoren, vekten til bilen (i kg) og slagvolumet til motoren (i liter). I R-utskriftene står **hk** for antall hestekrefter, **vekt** for bilens vekt, **vol** for slagvolum og **bruk** for bensinforbruk. Plottet nedenfor gir en oversikt over dataene.



(Fortsettes på side 2.)

Resultatet av en multippel lineær regresjon med bensinforbruk som respons og antall hestekrefter, vekt og slagvolum som forklaringsvariabler er gitt nedenfor. Merk at R-utskriften er redigert (ved at noe av informasjonen er fjernet fra utskriften).

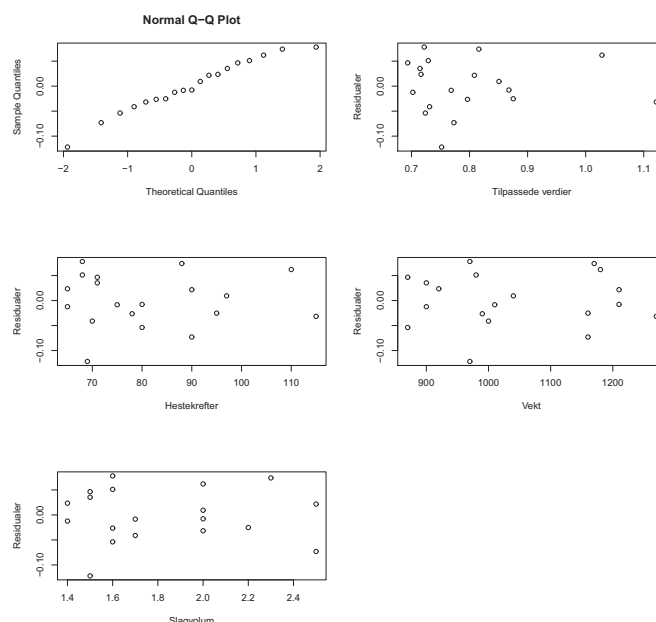
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.01033	0.11966	0.086	0.9323
hk	0.00604	0.00148	4.076	0.0010
vekt	0.00071	0.00023	3.005	0.0089
vol	-0.23979	0.07335		

Residual standard error: 0.0577

Multiple R-squared: 0.8212

- Forklar kort hva de ulike delene av R-utskriften forteller deg.
- Bruk en hypotesetest til å avgjøre om slagvolumet til motoren har signifikant betydning for bensinforbruket. Hvordan kan du forklare at den estimerte effekten av slagvolum er negativ selv om sammenhengen ser ut til å være positiv i plottet først i oppgaven?
- La θ være forventet endring i bensinforbruket når styrken til motoren økes med 10 hestekrefter (og vekt og slagvolum ikke endres). Bestem et 95% konfidensintervall for θ og forklar hva intervallet sier deg.

Nedenfor er det gitt ulike plott av residualene for den tilpassede modellen.



- Beskriv de forutsetningene regresjonsanalysen bygger på og bruk residualplottene til å vurdere om disse forutsetningene er rimelig godt oppfylt.

(Fortsettes på side 3.)

Oppgave 2

Menneskene har hatt terninger i flere tusen år. De eldste terningene var skåret ut av et bein fra sauefoten, eller de var lagd av stein, metall eller keramikk. Vi regner i dag med at sannsynligheten er $1/6$ for å få sekser når vi kaster en terning. Men var det tilfellet også for de terningene som ble brukt i tidligere tider?

På British Museum i London fins det flere terninger fra romertiden, blant annet en terning av marmor og en terning av jern. Det har blitt gjort forsøk med å kaste disse terningene. Da marmorterningen ble kastet 204 ganger, fikk en sekser i 54 av kastene.

- La p være sannsynligheten for at et kast med marmorterningen vil gi sekser. Bestem en test med signifikansnivå 5% for testing av nullhypotesen $H_0 : p = \frac{1}{6}$ mot den alternative hypotesen $H_a : p > \frac{1}{6}$. Hva blir konklusjonen din?
- Forklar hva P -verdien til en test er. Bestem P -verdien for situasjonen i punkt a.

Terningen av jern har også blitt kastet 204 ganger. Da fikk en sekser i 42 av kastene.

- La nå p være sannsynligheten for at et kast med terningen av jern vil gi sekser. Test nullhypotesen $H_0 : p = \frac{1}{6}$ mot den alternative hypotesen $H_a : p > \frac{1}{6}$. Bruk signifikansnivå 5%. Hva blir konklusjonen din?
- Forklar hva vi mener med feil av type II. Bestem sannsynligheten for feil av type II hvis sannsynligheten for å få sekser med terningen av jern er $p = 0.20$.
- Hvor mange kast må en gjøre med terningen av jern for at sannsynligheten for feil av type II skal være 20% hvis $p = 0.20$?

Oppgave 3

Vi ser på den lineære regresjonsmodellen

$$Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i; \quad i = 1, 2, \dots, n; \quad (1)$$

der x_i -ene er gitte størrelser, ϵ_i -ene er uavhengige og $N(0, \sigma^2)$ -fordelte og β_0 , β_1 og σ^2 er ukjente parametere. Merk at vi i (1) har trukket fra gjennomsnittet av x_i -ene. Vi sier at vi har sentrert forklaringsvariabelen.

- Vis at minste kvadraters estimatorer for β_0 og β_1 kan gis på formen

$$\hat{\beta}_0 = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{og} \quad \hat{\beta}_1 = \sum_{i=1}^n \frac{x_i - \bar{x}}{S_{xx}} Y_i$$

$$\text{der } S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

(Fortsettes på side 4.)

- b) Vis at $\hat{\beta}_0$ og $\hat{\beta}_1$ er forventningsrette.
 c) Bestem $V(\hat{\beta}_0)$ og $V(\hat{\beta}_1)$ og vis at $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$.

Hint: Hvis Z_1, \dots, Z_n er uavhengige stokastiske variabler og a_1, \dots, a_n og b_1, \dots, b_n er konstanter, så er $\text{Cov}(\sum_{i=1}^n a_i Z_i, \sum_{i=1}^n b_i Z_i) = \sum_{i=1}^n a_i b_i V(Z_i)$.

Vi er interessert i å estimere $\mu_{Y \cdot x^*} = \beta_0 + \beta_1(x^* - \bar{x})$, dvs. forventet respons når $x = x^*$.

- d) Vis at $\hat{\mu}_{Y \cdot x^*} = \hat{\beta}_0 + \hat{\beta}_1(x^* - \bar{x})$ er en forventningsrett estimator for $\mu_{Y \cdot x^*}$ og bestem variansen til estimatoren.

En estimator for σ^2 er

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i - \bar{x}) \right)^2.$$

Det er kjent at S^2 er uavhengig av $\hat{\beta}_0$ og $\hat{\beta}_1$ og at $(n-2)S^2/\sigma^2$ er kji-kvadrat fordelt med $n-2$ frihetsgrader. (Du skal ikke vise det.)

- e) Forklar at

$$\frac{\hat{\mu}_{Y \cdot x^*} - \mu_{Y \cdot x^*}}{S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

er t -fordelt med $n-2$ frihetsgrader og bruk det til å bestemme et $100(1-\alpha)\%$ konfidensintervall for $\mu_{Y \cdot x^*}$.

SLUTT

Vedlegg 1: Data for personbiler

Nedenfor er det gitt ulike tekniske data og data om bensinforbruk for 19 ulike merker av personbiler. Hver linje i tabellen gir informasjon for et bilmerke, og kolonnene i tabellen gir antall hestekrefter, vekt (i kg), slagvolumet til motoren (i liter) og bensinforbruket (i liter per mil).

hk	vekt	vol	bruk
68	980	1.6	0.78
95	1160	2.2	0.85
97	1040	2.0	0.86
75	1010	1.7	0.76
115	1270	2.0	1.09
88	1170	2.3	0.89
65	900	1.4	0.69
80	870	1.6	0.67
80	1210	2.0	0.86
71	900	1.5	0.75
68	970	1.6	0.80
90	1210	2.5	0.83
90	1160	2.5	0.70
70	1000	1.7	0.69
65	920	1.4	0.74
69	970	1.5	0.63
78	990	1.6	0.77
110	1180	2.0	1.09
71	870	1.5	0.74