Faglig aktivitet II

November 2020

1 Formål

Opgaven er løsning af et datasæt i R.
Datasættet er vedhæftet.
Selve opgaven:

- Opgaven skal løses individuelt.
- På forsiden skal være studienummer.
- Opgaverne skal løses i R.
- For hver delopgave skal der være en sammenfattende konklusion.
- Der må gerne være kommentarer til koden, men ikke et krav.
- Det hele afleveres i én pdf fil.

2 Opgaven

Baggrund I lyset af nedlukningen af samfundet i foråret 2020 har fokus været rettet mod sundhedsvæsnets udgifter til behandling. Denne opgave beskæftiger sig med udgifter, der er forbundet med indlæggelse på hospital. Materialet hidrører fra en større stikprøveundersøgelse af indlæggelsesforløb.
Data for indlæggelse af 138 patienter, der omfatter tre danske regioner, findes i Excel-filen $aktivitet_I I.xlsx$ og indeholder følgende variabler:
Variabel Beskrivelse.

- TREATCOST Samlede udgifter i DKK til behandling af patienten
- MEDICINE Udgifter i DKK til medicin
- LAB Udgifter til laboratorieundersøgelser i DKK
- XRAY Udgifter til røntgenundersøgelser i DKK
- INHALATOR Udgifter til respirator i DKK

- STATUS Patientens tilstand ved udskrivelse:
  0 = patienten døde; 1 = patienten overlevede
- CAREDAYS Antal dage patienten var indlagt på hospital
- INTENSIVEDAYS Antal dage patienten var indlagt på intensiv
- AGE Patientens alder i år
- GENDER Patientens køn: 0 = kvinde; 1 = mand
- INSURANCE Forsikring der har betalt for indlæggelsen:
  0 = offentlig sygesikring; 1 = privat forsikring
- REGION Klassifikationsvariabel af geografisk lokation for den indlagte patient:
  1 = Region Hovedstaden; 2 = Region Midtjylland; 3 = Region Syddanmark

2.1 Opgave

Udarbejd en deskriptiv statistisk analyse af de samlede udgifter til behandling af patienten TREATCOST. Analysen skal indeholde et histogram. Definer middelværdien, modus, medianen og standardafvigelsen. Kommenter på formen og symmetrien af fordelingen og undersøg, om der er ekstreme observationer.

- Opstil og fortolk et 95 % konfidensinterval for middelværdien.
  Antag at variablen TREATCOST er normalfordelt med middelværdi og standardafvigelse, som fundet i opgaven.
- Hvad er sandsynligheden for, at de samlede udgifter til behandling af en tilfældigt valgt patient ligger over 95.000 DKK?
- Hvad er sandsynligheden for, at de samlede udgifter til behandling af en tilfældigt valgt patient ligger mellem 40.000 DKK og 65.000 DKK?

2.2 Opgave

Udgiften til en indlæggelse TREATCOST kan tænkes at være positivt relateret til antallet af dage, patienten har været indlagt på hospital CAREDAYS. Opstil, for at undersøge denne problemstilling en simpel regressionsmodel, af formen:

$$TREATCOST_i = \beta_0 + \beta_1 CAREDAYS_i + \epsilon_i \quad i = 1, \ldots, 138$$

Redegør for den anvendte metode og dens forudsætninger. Gennemfør en modelkontrol. Er residualerne "hvid støj", og hvordan ser normalfordelingsplottet ud? Er dette en god model, når der ses på determinationskoefficienten? Observeres den postulerede sammenhæng?

2.3 Opgave

Udvid analysen og anvend multipel regression til at undersøge, hvilke variable der øver indflydelse på TREATCOST. Estimer en model af følgende from:

$$TREATCOSTi = \beta_0 + \beta_1 MEDICINE_i + \beta_2 LAB_i + \beta_3 XRAY_i + \beta_4 INHALATOR_i + \beta_5 STATUS_i + \beta_6 CAREDAYS_i + \beta_7 INTENSIVEDAYS_i + \beta_8 AGE_i + \beta_9 GENDER_i + \beta_1 0 INSURANCE_i + \epsilon_1 \quad i = 1, \ldots, 138$$

Udarbejd et pænt og letlæseligt output. Opstil en korrelationsmatrix og kommenter på denne. Foretag en selektion af variablerne, og find den mest hensigtsmæssige model. Kommenter på fortegnene, og kommenter på de relevante residualdiagrammer.

2.4 Opgave

Regionale forskelle i specialisering kan tænkes at have en indflydelse på brugen af respirator og således også på de dermed forbundne udgifter. Undersøg denne problemstilling ved anvendelse af et relevant test. Anvend variablen INHALATOR fordelt på de 3 grupper givet ved variablen REGION. Opstil forudsætningerne, formuler hypoteserne og udfør testet. Foretag om nødvendigt en supplerende analyse og kommenter på udfaldet.

2.5 Opgave

Udgiften til bestemte undersøgelser kan tænkes at være relateret til køn. Undersøg på denne baggrund en hypotese, der siger, at middeludgifterne røntgenundersøgelse er højere for mandlige end for kvindelige patienter. Brug variablen GENDER til at opdele variablen XRAY i 2 grupper. Formuler hypoteserne for testet og udfør dette. Hvad er udfaldet af testet, og hvordan defineres p-værdien.

Opstil og udfør i forlængelse af analysen et test til at undersøge om varianserne i de 2 dataserier er identiske. Hvorfor er dette test relevant at foretage i relation til opgavens første del?