STOCKHOLMS UNIVERSITET
Statistiska institutionen
Karl Sigfrid

Tentamen i Dataanalys och Regression, 4.5 hp

Kurs: ST1101, Statistik och Dataanalys I, 15 hp
Tentamensdatum: 2024-11-07

Skrivtid: 08.00-12.00 (4 timmar).
Godkända hjälpmedel: Miniräknare utan lagrade formler och text. Lexikon.
Bifogade hjälpmedel: Formelsamling och statistiska tabeller.

Tentamen består av 5 uppgifter, uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.

Svar med fullständiga redovisningar ska lämnas. Svar ges på svenska eller engelska.

- Använd endast skrivpapper som tillhandahålls i skrivsalen.
- För full poäng på en uppgift krävs tydliga och väl motiverade lösningar.

Tentamen kan maximalt ge 100 poäng, och för godkänt resultat krävs minst 50.

Betygsgränser:

A:	90-100
B:	80-89
C:	70-79
D:	60-69
E:	50-59
Fx:	40-49
F:	0-39

OBS! Fx och F är underkända betyg som kräver omexamination. Studenter som får betyget Fx kan alltså inte komplettera för högre betyg. Lösningsförslag läggs ut på Athena efter tentamen i samband med rättningen.

`Lycka till!`

Uppgift 1. (20 poäng)

(a.) Nämn två typer av diagram som är lämpliga att använda för att visa fördelningen av en numerisk variabel. Nämn också två typer av diagram som är lämpliga för att visa fördelningen av en kategorisk variabel. (5p)

(b.) Vad är en outlier, och hur kan du agera när du upptäcker en outlier i din data? (5p)

(c.) Vad betyder det när vi säger att bias har uppstått som en konsekvens av vår metod för att samla in data? Resonera i termer av vilka individer som har möjlighet bli en del av vårt urval (sample). Ge ett exempel, verkligt eller påhittat, på hur bias kan uppkomma när data samlas in. (5p)

(d.) Vad är en dold variabel (lurking variable)? Ge ett exempel på ett scenario, verkligt eller påhittat, där en dold variabel har betydelse. (5p)

Uppgift 2. (20 poäng)

(a.) Vad är en förklaringsvariabel (explanatory variable) respektive en responsvariabel (response variable) i en regressionsmodell? (5p)

(b.) När du har anpassat en regressionsmodell kan du för varje observation räkna ut skillnaden mellan det observerade värdet och det skattade värdet av responsvariabeln med formeln

$e = y - \hat{y}$

Vad är namnet på det avstånd som e representerar i ovanstående formel, och hur är e kopplat till de modellantaganden om konstant varians och normalfördelning som vi gör för en enkel linjär regression? (5p)

(c.) Vi har två regressionsmodeller. När vi anpassar vardera modell med träningsdata, och utvärderar med separat testdata, har modell 1 lägre RMSE än modell 2. Modell 1 har även lägre R-kvadrat (R^2) än modell 2. Vad säger detta om de båda modellerna, och vilken modell är rimligast att välja? Motivera ditt svar. (5p)

(d.) Nämn en fördel med att göra en studie i form av ett experiment istället för att göra det i form av en observationsstudie. (5p)

Uppgift 3. (20 poäng)

En matvarubutik bestämde sig för att belöna sina bästa stamkunder. De kunder som belönades fick välja mellan en flaska olivolja, en chokladask och en biobiljett. Butiken kände till åldern på de kunder som belönades, och kunder som var 65 år eller äldre kategoriserades som pensionärer. De val som kunderna gjorde är sammanställda i tabellen nedan.

		Pensionär	
		Ja	Nej
Belöning	Olivolja	108	290
	Chokladask	201	447
	Biobiljett	39	60

Variabeln *Belöning* anger vilken belöning en kund valde. Variabeln *Pensionär* anger om kunden kategoriserades som pensionär eller inte.

(a.) Hur många kunder totalt var det som fick en belöning? (1p)

(b.) Hur många pensionärer var det som valde en biobiljett? (1p)

(c.) Räkna ut marginalfördelningarna (marginal distributions) för båda variablerna. Uttryck marginalfördelningarna i procent. Tolka marginalfördelningarna. (6p)

(d.) Räkna ut fördelningen av variabeln *Belöning* betingad (conditioned) på variabeln *Pensionär*. Ange de betingade fördelningarna i procent. (8p)

(e.) Tolka resultatet från deluppgift **d**. (4p)

Uppgift 4. (20 poäng)

En biolog undersöker hur vikten för en viss fiskart påverkas av mängden föda. För ett antal fiskar har hon samlat data som visar hur många kalorier varje fisk har ätit per dag och vad varje fisk väger i gram vid två månaders ålder. Denna data har använts för att anpassa regressionsmodellen

$$\hat{y} = 134 + 0.2x,$$

där x är antalet kalorier per dag och $\hat{y}$ är den estimerade vikten vid två månaders ålder.

(a.) Tolka modellens koefficienter (coefficients), dvs interceptet och lutningskoefficienten (slope coefficient). (4p)

(b.) Hur mycket estimerar modellen att en fisk väger vid två månaders ålder om den har ätit 21 kalorier per dag? (4p)

(c.) Vi har följande information: $r_{xy} = 0.16, s_x = 4, s_y = 5, \bar{x} = 18, \bar{y} = 137.6$.

Förklara notationen, dvs förklara vad r_{xy}, s_x, s_y, $\bar{x}$ och $\bar{y}$ står för.

Verifiera sedan med hjälp av denna information att $b_0 = 134$ och att $b_1 = 0.2$. (4p)

(d.) Vi lägger till dummy-variabeln *vitamintillskott*, som har värdet 1 för fiskar som fick ett vitamintillskott utöver maten. För fiskar som inte fick vitamintillskottet är värdet 0. Modellen blir nu.

$$\hat{y} = 126 + 0.19x_1 + 24x_2,$$

där $\hat{y}$ är fiskens vikt vid två månaders ålder, x_1 är antalet kalorier per dag och x_2 är dummyvariabeln för vitamintillskott.

Tolka modellens samtliga tre koefficienter. (4p)

(e.) Om en fisk fick vitamintillskott har x_2 värdet 1 och annars värdet 0. Formulera regressionsmodellen så som den skulle se ut om kodningen var den omvända, dvs om fiskar som fick vitamintillskott kodades som 0 och fiskar som inte fick vitamintillskott kodades som 1. Visa med ett exempel att prediktionerna blir samma oavsett hur dummyvariabeln kodas. (4p)

Uppgift 5. (20 poäng)

Ett bibliotek har sina böcker i ett register, och av registret framgår antalet sidor i varje bok. En bibliotikare som analyserat böckerna har noterat att antalet sidor i böckerna följer en normalfördelning. Det genomsnittliga antalet sidor är 340, och standardavvikelsen är 120 sidor.

(a.) En av bibliotekets böcker är "Jurtjyrkogården" av Stephen King, som är 459 sidor lång. Hur stor andel av bibliotekets böcker har *färre* antal sidor än "Jurtjyrkogården"? (5p)

(b.) Den normala lånetiden för en bok är två veckor, men biblioteket överväger att införa fyra veckors lånetid för de böcker vars sidantal ligger över den 90:e percentilen. Hur många sidor måste en bok minst ha för att sidantalet ska ligga över den 90:e percentilen? Avrunda svaret till närmaste heltal. (5p)

(c.) Räkna ut kvartilavståndet (IQR) för fördelningen av antalet sidor i bibliotekets böcker. (5p)

(d.) Bibliotikarien funderar på vad som skulle hända med fördelningen av antalet sidor om hon tog bort 10 procent av sidorna i varje bok. Vilket värde får den första kvartilen (Q1) i fördelningen om antalet sidor i varje bok minskar med 10 procent? (5p)