STOCKHOLMS UNIVERSITET
Statistiska institutionen
Karl Sigfrid

Tentamen i Dataanalys och Regression, 4.5 hp

Kurs: ST1101, Statistik och Dataanalys I, 15 hp
Tentamensdatum: 2024-02-09

Skrivtid: 14.00-18.00 (4 timmar).
Godkända hjälpmedel: Miniräknare utan lagrade formler och text. Lexikon. Linjal.
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.
- Om du inte lyckas lösa en deluppgift och behöver det svaret för en senare deluppgift så kan du hitta på värdet för att kunna göra beräkningarna i de efterföljande uppgifterna.

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.) Vad är metadata? (5p)

(b.) Vad är en ordinal variabel? Ge exempel på en ordinal variabel, verkligt eller påhittat. (5p)

(c.) På en liten lågstadieskola går det 33 elever. 12 elever går i årskurs 1, 14 elever i årskurs 2 och 7 elever i årskurs 3. Anta att vi har ett dataset där varje observation representerar en av skolas elever, och den kategoriska variabeln *årskurs* anger vilken årskurs eleven går i. Rita **två olika typer av diagram** som båda visar hur variabeln *årskurs* är fördelad. Ange namnet på respektive typ av diagram. Diagrammen behöver inte vara snyggt ritade, men du bör välja lämpliga typer av diagram, och diagrammen bör se korrekta ut på ett ungefär. (5p)

(d.) På lågstadieskolan som beskrivs i deluppgift **c** vet vi att alla 12 elever i årskurs 1 är 7 år gamla, alla 14 elever i årskurs 2 är 8 år gamla, och alla 7 elever i årskurs 3 är 9 år gamla. Beräkna typvärdet (mode), medianen och medelvärdet för åldersfördelningen hos skolans elever. (5p)

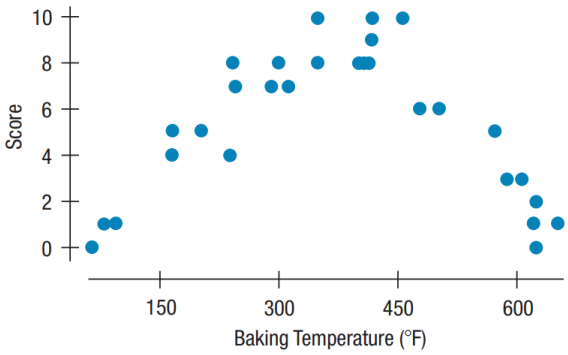
Uppgift 2. (20 poäng)

(a.) När vi anpassar (fit) en regressionsmodell brukar modellens R-squared redovisas. Vad mäter R-squared? (4p)

(b.) När vi utvärderar en regressionsmodell kan vi vilja dela in datamaterialet i träningsdata och testdata. Vad är syftet med att göra det? (6p)

(c.) Två typer av felkällor när du drar generella slutsatser utifrån ett datamaterial är **slumpmässig variation** och **bias**. Förklara vad orsaken till var och en av dessa två felkällor kan vara. Ge gärna exempel, verkliga eller påhittade. (6p)

(d.) En grupp vänner har provat att baka kladdkakor med olika ugnstemperaturer, och sedan har de poängsatt varje kladdkaka. Grafen nedan visar att det finns ett tydligt samband mellan baktemperatur (x-axeln) och smakpoäng (y-axeln). När du räknar ut korrelationskoefficienten mellan temperatur och smakpoäng blir den nära noll, trots det tydliga sambandet. Förklara varför. (4p)



Figur 1: Smakpoäng vs. Ungstemperatur

Uppgift 3. (20 poäng)

En biltillverkare har samlat ihop ett antal testpersoner för en filmvisning. Deltagarna delas in i tre grupper, och grupperna får se varsin reklamfilm: film A, film B eller film C. Efter att ha sett en av de tre filmerna får varje deltagare ange om de har blivit mer positivt inställda till det aktuella bilmärket. Resultatet visas i tabellen nedan.

		Positiv	
		Ja	Nej
Film	Film A	35	101
	Film B	37	107
	Film C	32	56

Variabeln *Film* anger vilket film en viss deltagare såg, variabeln *Positiv* anger om en deltagare blev mer positivt inställd eller inte efter att ha sett filmen.

(a.) Hur många testpersoner totalt deltog i filmvisningen? (2p)

(b.) Räkna ut marginalfördelningarna (marginal distributions) för båda variablerna. Uttryck marginalfördelningarna i procent. Tolka marginalfördelningarna. (4p)

(c.) Syftet med filmvisningen är att utvärdera de olika reklamfilmerna. För att göra detta, bör vi räkna ut fördelningen av variabeln *Positiv* betingad på variabeln *Film*, eller bör vi räkna ut fördelningen av variabeln *Film* betingad på variabeln *Positiv*? Motivera ditt svar. (4p)

(d.) Räkna ut den betingade fördelning som du föreslog i deluppgift **c**. Uttryck svaren i procent. (8p)

(e.) Tolka resultatet från deluppgift **d**. (2p)

Uppgift 4. (20 poäng)

Servitören på en restaurang har sammanställt data från sina gäster. För varje sällskap som besökt restaurangen har han noterat totalbeloppet på notan och hur mycket sällskapet betalade i dricks. Med hjälp av denna data har han anpassat regressionsmodellen

$$\hat{y} = 32 + 0.09x,$$

där x är totalbeloppet på notan i kronor (exklusive dricks) och $\hat{y}$ är det estimerade beloppet som sällskapet betalar i dricks.

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

(b.) Vi har följande information: $r_{xy} = 0.8, s_x = 402, s_y = 45.2, \bar{x} = 1105, \bar{y} = 131.5$.

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 $b_0 = 32$ och att $b_1 = 0.09$ efter att du har avrundat dina resultat. (4p)

(c.) Vi lägger till dummy-variabeln *lunch*, som har värdet 1 om gästerna äter lunch och 0 om de äter middag. Modellen blir nu

$$\hat{y} = 47 + 0.14x_1 - 67x_2,$$

där x_1 är totalbeloppet på notan, x_2 är dummyvariabeln för lunch, och $\hat{y}$ är det estimerade beloppet som sällskapet betalar i dricks.

Tolka modellens samtliga tre koefficienter. Kom ihåg att vi nu har en multipel linjär regressionsmodell. (4p)

(d.) En kollega på en annan restaurang har tagit fram sin egen modell för att estimera hur mycket dricks ett sällskap betalar. Kollegan har transformerat responsvariabeln med den naturliga logaritmen (logaritmen med bas e), och fått modellen

$$\widehat{log(y)} = 4 + 0.001x,$$

där x är totalbeloppet på notan i kronor (exklusive dricks) och $\widehat{log(y)}$ är estimatet av det logaritmerade beloppet som sällskapet betalar i dricks. Hur mycket dricks estimerar denna modell att ett sällskap betalar om totalbeloppet på notan är 724 kronor? Ange det estimerade beloppet i kronor. (4p)

(e.) Servitören testar att lägga till ytterligare variabler i sin regressionsmodell, som exempelvis tiden då notan skrevs ut, hur många som ingick i sällskapet, och hur många rätter som beställdes in. När han studerar *R-kvadrat* (R-squared) ser han att värdet på *R-kvadrat* har ökat en aning som ett resultat av de tillagda variablerna. Betyder det att modellen har blivit bättre och att de tillagda variablerna därför bör behållas i modellen? Finns det något ytterligare han kan göra för att undersöka om de tillagda variablerna förbättrar modellen? (4p)

Uppgift 5. (20 poäng)

Lönerna på ett stort svenskt företag under 2023 var normalfördelade. Medellönen var 42 tusen kronor per månad, och standardavvikelsen var 7 tusen kronor per månad.

(a.) Johan, som arbetade på företaget, tjänade 53 tusen kronor i månaden. Hur stor andel av de anställda hade *högre* lön än Johan? (5p)

(b.) Hur mycket tjänade en anställd vars lön befann sig vid den 20:e percentilen av fördelningen? (5p)

(c.) Beräkna kvartilavståndet för lönefördelningen. (5p)

(d.) I slutet av 2023 konstaterade företagsledningen att affärerna hade gått oväntat bra. De beslutade att ge alla anställda en löneökning på 10 procent. Förhandlingar med facket ledde fram till att varje anställd, utöver den beslutade löneökningen på 10 procent, dessutom fick en ytterligare höjning med 500 kronor per månad. Efter den totala löneökningen, vad blir den nya medianlönen? (5p)