STOCKHOLMS UNIVERSITET
Statistiska institutionen
Karl Sigfrid

Tentamen i Dataanalys och Regression, 4.5 hp

Kurs: ST1101, Statistik och Dataanalys I, 15 hp
Tentamensdatum: 2025-02-13

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.) Nämn en skillnad mellan kategoriska variabler och numeriska variabler. Ge ett exempel på en kategorisk respektive en numerisk variabel. (5p)

(b.) Stapeldiagram och histogram är två typer av diagram som ser ungefär likadana ut. Vad är skillnaden mellan stapeldiagram och histogram? (5p)

(c.) Vad är en outlier? (5p)

(d.) Ett företag som tillverkar telefoner ville undersöka telefonernas batteritid. De testade därför hur länge batteriet räckte på 11 telefoner. På 2 av telefonerna räckte batteriet i 27 timmar. På 4 av telefonerna räckte batteriet i 29 timmar. På 5 av telefonerna räckte batteriet i 30 timmar. Beräkna typvärdet (mode), medianen och medelvärdet för batteritiden hos de telefoner som testades. (5p)

Uppgift 2. (20 poäng)

(a.) Vad betyder det att en modell är överanpassad (overfitted)? Vad betyder det att en modell är underanpassad (underfitted)? (5p)

(b.) När du anpassar och utvärderar en statistik modell, som exempelvis en regressionsmodell, vad är syftet med att dela in datamaterialet i träningsdata och testdata? (5p)

(c.) På en universitetskurs gick det att delta antingen på plats eller på distans. Universitetet ville undersöka om det fanns en skillnad i hur nöjda de som gick kursen på distans var jämfört med dem som gick kursen på plats. En enkät bland studenterna resulterade i korstabellen nedan. Räkna ut marginalfördelningarna *i procent* för de båda variablerna *Studieform* och *Nöjd*. Tolka fördelningarna. (4p)

		Nöjd	
		Ja	Nej
Studieform	På plats	157	60
	Distans	96	47

(d.) Med samma korstabell som i deluppgift (c.), räkna ut fördelningen av variabeln *Nöjd* betingad på *Studieform*. Uttryck de betingade fördelningarna i procent. Tolka resultatet. (6p)

Uppgift 3. (20 poäng)

Ekonomiassistenten på ett litet förtag har ett dataset med information om de anställda. För varje anställd finns flera variabler. En variabel anger den anställdes lön i kronor per månad. En annan variabel anger hur många år den anställde har arbetat på företaget.
För att undersöka om det finns ett samband mellan lön och anställningstid anpassar ekonomiassistenten regressionsmodellen

$$\hat{y} = 27669 + 577x,$$

där x är antalet år personen har varit anställd och $\hat{y}$ är den estimerade månadslönen i kronor.

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

(b.) Hur hög estimerar modellen att lönen är för en anställd som har jobbat på företaget i 8 år? (4p)

(c.) Ekonomiassistenten har också information om vilka av företagets anställda som har en högskoleexamen, och lägger till en dummyvariabel för det i modellen. Modellen är nu

$$\hat{y} = 25521 + 489x_1 + 3408x_2,$$

där x_1 är antalet år personen har varit anställd. Variabeln x_2 är en dummyvariabel som har värdet 1 om personen har en högskoleexamen, och annars värdet 0. $\hat{y}$ är den estimerade månadslönen i kronor. Tolka modellens samtliga tre koefficienter. För poäng på den här deluppgiften måste tolkningen ta hänsyn till att det nu är en multipel linjär regressionsmodell. (6p)

(d.) I deluppgift (c.) fick variabeln x_2 värdet 1 om en anställd hade högskoleexamen, och annars värdet 0. Formulera regressionsmodellen så som den skulle se ut om kodningen var den omvända, dvs om x_2 fick värdet 1 för anställda som *saknar* högskoleexamen och 0 för anställda som har en högskoleexamen. Visa med ett exempel att prediktionerna blir samma oavsett hur dummyvariabeln kodas. (6p)

Uppgift 4. (20 poäng)

I uppgift 3 såg vi den enkla linjära regressionsmodellen

$$\hat{y} = 27669 + 577x,$$

där x är antalet år som en person har varit anställd på ett företag och $\hat{y}$ är personens estimerade månadslön i kronor. Företaget är ett litet företag med bara 5 anställda.

I tabellen nedan visar y de anställdas löner, och x visar hur många år de har varit på företaget. Tabellen visar också z-värden, dvs standardiserade värden, för x och y.

	y	x	z_y	z_x
Anställd 1	28000	2	-1.35	-1.35
Anställd 2	32000	10	-0.79	-0.64
Anställd 3	43000	19	0.76	0.16
Anställd 4	43000	25	0.76	0.69
Anställd 5	42000	30	0.62	1.13

(a.) Standardavvikelsen för y är ungefär 7092 kronor. Använd informationen i tabellen för att verifiera att det stämmer. (4p)

(b.) Verifiera med hjälp av de värden för variabel y som står i tabellen att $z_y = -1.35$ för *Anställd 1*. (4p)

(c.) Förklara i ord vad det betyder att $z_y = -1.35$ i termer av hur lönen för *Anställd 1* förhåller sig till övriga löner på företaget. (4p)

(d.) Visa att korrelationskoefficienten $r_{x,y} = 0.92$. (4p)

(e.) Vi vet att $r_{x,y} = 0.92$. Vi känner också till standardavvikelsen för respektive variabel: $s_y = 7092$, $s_x = 11.3$. Använd den här informationen, och annan information ur tabellen om det behövs, till att verifiera att $b_0 = 27669$ och att $b_1 = 577$ i den enkla linjära regressionsmodellen. (4p)

(För de deluppgifter som kräver beräkningar är det ok om du på grund av avrundningar kommer fram till svar som skiljer sig något från de värden som ska verifieras.)

Uppgift 5. (20 poäng)

Ett bussbolag transporterar dagligen resenärer mellan Norrköping och Linköping. Enligt bussbolaget följde restiderna under förra året en normalfördelning med medelvärdet 47 minuter och standardavvikelsen 8 minuter.

(a.) Hur stor andel av resorna tog mer än 60 minuter? (5p)

(b.) En bussförare får höra att restiden för dennes senaste resa mellan Norrköping och Linköping låg vid den 30:e percentilen. Vad betyder det, och ungefär hur lång tid tog förarens senaste resa? (5p)

(c.) Räkna ut kvartilavståndet (IQR) för restidernas fördelning. (5p)

(d.) Bussbolaget funderar på att köra en ny rutt, som skulle minska tiden för varje resa med exakt 4 minuter. Om varje bussresa förra året hade gått 4 minuter snabbare, vad hade då värdet varit vid den tredje kvartilen av tidsfördelningen? (5p)