Lösningsförslag, tentamen SDA1, del 1, 20230922

Uppgift 1

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

En numerisk variabel anges i en bestämd enhet. Värden på numeriska variabler kan rangodnas. En kategorisk variabel kan ha icke-numeriska värden och används ofta för gruppindelning av observationerna.

Exempel på en numerisk variabel kan vara ålder. Exempel på en kategorisk variabel kan vara kön.

(b.) Vad är en outlier? (5p)

En outlier är ett extremvärde som avviker från det övergripande mönstret i en fördelning.

(c.) Vad är bias? Ge ett exempel, verkligt eller påhittat, på hur bias kan uppkomma när data samlas in. (5p)

Bias är ett systematiskt fel som uppstår när vi tar ett stickprov. Det kan till exempel uppkomma i en survey om personer med vissa åsikter är mindre svarsbenägna och därför underrepresenterade i stickprovet. Det kan också uppstå om alla observationer i stickprovet kommer ur en grupp som inte representerar hela populationen med avseende på det vi vill mäta, exempelvis om vi vill mäta befolkningens medelinkomst och väljer ut enbart personer som har högskoleutbildning.

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

En dold variabel är en variabel som vi inte har i vår data men som orsakar ett samband mellan variabler som vi studerar. Exempel: Om vi ser ett samband mellan antalet tv-apparater per hushåll och kvalitén på sjukvården i ett land så kan landets ekonomiskt välstånd vara en dold variabel som förklarar varför variablerna samvarierar.

Uppgift 2. (20 poäng)

(a.) Vad är en förklaringsvariabel (explanatory variable) respektive en responsvariabel (response variable) i en regressionsmodell? (4p)

Responsvariabeln är variabeln som vi vill estimera med hjälp av förklaringsvariabeln.

(b.) Vad är en residual i en regressionsmodell? Hur beräknas den? (4p)

En residual är skillnaden mellan det observerade värdet och det estimerade värdet på responsvariabeln för en observation. Den räknas ut $y - \hat{y}$, där y är det observerade värdet och $\hat{y}$ det estimerade värdet.

(c.) När du har anpassat en regressionsmodell kan du studera residualerna för att se om de modellantaganden (model assumptions) som gäller för en regressionsmodell är uppfyllda. Nämn två sådana antaganden. (6p)

Residualer som är normalfördelade. Residualer med konstant varians.

(d.) En grupp ingenjörer diskuterar vilken av två regressionsmodeller som är bäst. En ingenjör föredrar modell 1, som har lägre RMSE när den utvärderas genom korsvalidering. En annan ingenjör invänder och hävdar att en modell med högre R-kvadrat (R-squared) alltid är bättre, och eftersom den mer flexibla modell 2 har högre R-kvadrat bör den väljas. Har den andra ingenjören rätt i sitt resonemang on R-kvadrat? Motivera ditt svar. (6p)

Ingenjören har inte rätt i sitt resoneamng om R-squared. Värdet på R-squared är ofta högt för flexibla modeller, som anpassar sig väl till träningsdata. Om RMSE är lägre än för en enklare modell så talar det för att den enklare modellen har bättre generaliserbarhet och ger bättre skattningar när vi gör prediktioner med ny data.

Uppgift 3. (20 poäng)

En butikskedja har tre butiker, som ligger i Stockholm, Göteborg och Malmö. Butikskedjan har gjort en kundundersökning, där kunder i de tre butikerna har tillfrågats om de är nöjda eller inte med sitt senaste köp. Resultatet av undersökningen visas i tabellen.

Variabeln *Stad* anger i vilken butik en kund har handlat och variabeln *Nöjd* anger om kunden är nöjd eller inte.

Här är en lösning i R. Om du använder denna tenta som övningstenta, försäkra dig gärna om att du får samma resultat med en miniräknare.

Vi börjar med att definiera en matris med de värden som ges i tabellen. Vi kallar matrisen m.

```
m <- matrix(c(138, 101, 60, 60, 47, 39), ncol=2)
rownames(m) <- c("Stockholm", "Göteborg", "Malmö")
colnames(m) <- c("Ja", "Nej")
m
```

```
           Ja Nej
Stockholm 138  60
Göteborg  101  47
Malmö      60  39
```

(a.) Hur många kunder totalt svarade på frågan? (2p)

Varje kund som svarade på frågorna finns i en av de 6 grupperna. Genom att lägga ihop antalet respondeter i varje grupp får vi det totala antalet respondenter.

```
sum(m)
```

```
[1] 445
```

(b.) Hur många kunder i undersökningen hade handlat i Malmö-butiken och var missnöjda med sitt senaste köp? (2p)

Vi ser i tabellen att det finns 39 kunder som tillhör gruppen som handlade i Malmö-butiken, och som samtidigt tillhör gruppen som var missnöjd.

(c.) Räkna ut marginalfördelningarna för båda variablerna. Uttryck marginalfördelningarna i procent. Tolka marginalfördelningarna. (4p)

Vi får marginalfördelningen för variabeln *Nöjd* i antal genom att räkna ut summan av varje kolumn. Genom att dela med det totala antalet personer i undersökningen och multiplicera med 100 får vi marginalfördelningen i procent.

```
# Marginalfördelningen i antal
colSums(m)
```

```
 Ja Nej
299 146
```

```
#Marginalfördelningen i procent
100 * colSums(m) / sum(m)
```

```
      Ja      Nej
67.19101 32.80899
```

Vi får marginalfördelningen för variabeln *Stad* i antal genom att räkna ut summan av varje rad. Genom att dela med det totala antalet personer i undersökningen och multiplicera med 100 får vi marginalfördelningen i procent.

```
# Marginalfördelningen i antal
rowSums(m)
```

```
Stockholm  Göteborg     Malmö
      198       148        99
```

```
# Marginalfördelningen i procent
100 * rowSums(m) / sum(m)
```

```
Stockholm  Göteborg     Malmö
 44.49438  33.25843  22.24719
```

(d.) Räkna ut fördelningen av variabeln Nöjd betingad på variabeln Stad. Ange de betingade fördelningarna i procent. (8p)

När vi betingar fördelningen av variabeln *Stad* betyder det att vi delar in observationerna i grupper - en grupp för varje stad. För var och en av städerna räknar vi sedan ut procentandelen som är nöjd respektive missnöjd.

Notera att summan för varje rad, dvs varje stad, är 100%.

```
# Fördelningen av Nöjd i procent betingad på Stad

# Koden nedan delar varje värde i matrisen m med radsumman
# och multiplicerar med 100
100 * m / rowSums(m)
```

```
                Ja      Nej
Stockholm 69.69697 30.30303
Göteborg  68.24324 31.75676
Malmö     60.60606 39.39394
```

(e.) Tolka resultatet från deluppgift d i termer av skillnader i kundnöjdhet. (4p)

Stockholm och Göteborg har ungefär samma andel nöjda kunder. Malmö har en något mindre andel nöjda kunder. (Detta gäller för de kunder som ingår i vår data. Om skillnaderna är statistiskt signifikanta är en fråga för nästa delkurs.)

Uppgift 4. (20 poäng)

Ett antagningsprov till en utbildning i journalistik mäter ordförståelse. Baserat på ett stort antal provresultat vet vi att provresultaten är normalfördelade. Medelvärdet är 55 poäng, och standardavvikelsen är 8 poäng.

(a.) Anna gör provet och får 59 poäng. Hur stor andel av alla som har gjort provet har fått högre poäng än Anna? (4p)

Vi kan räkna ut det z-värde som motsvarar $y = 59$ poäng med formeln

$$z = \frac{y - \bar{y}}{s_y} = \frac{59 - 55}{8} = 0.5$$

I normalfördelningstabellen kan vi se att 69.15% av alla observationer har z-väden som är mindre än 0.5, vilket betyder att $100\% - 69.15\% = 30.85\%$ av alla observationer har ett större z-värde.

Det är alltså ungefär 31% som har ett högre resultat.

Så här kan vi räkna ut samma sak i R med funktionen *pnorm*:

```
z <- (59 - 55) / 8
z
```

```
[1] 0.5
```

```
1 - pnorm(q=z)
```

```
[1] 0.3085375
```

(b.) För att bli antagen till utbildningen krävs att ditt provresultat tillhör de högsta 10 procenten, dvs att resultatet ligger vid den 90:e percentilen eller högre. Hur många poäng behöver du för att uppnå detta? (Avrunda svaret uppåt till närmaste heltal.) (4p)

Vi börjar med att leta upp det z-värde som motsvarar den 90:e percentilen. Vi hittar inte exakt 0.9, men vi hittar 0.8997 som är det närmaste värdet. Det motsvarar ett z-värde som är 1.28.

För att omvandla vårt z-värde till antal poäng använder vi formeln

$$y = \bar{y} + z \cdot s_y = 55 + 1.28 \cdot 8 = 65.24$$

Avrundat uppåt behöver alltså 66 poäng för att vårt resultat ska vara bland de 10 procent högsta resultaten.

Så här kan vi räkna ut samma sak i R med funktionen *qnorm*:

```
z <- qnorm(p=0.9)
z
```

```
[1] 1.281552
```

```
ceiling(55 + z * 8) # ceiling betyder att vi avrundar uppåt
```

```
[1] 66
```

(c.) Räkna ut kvartilavståndet (IQR) för fördelningen av provresultat. Gör detta med hjälp av normalfördelningen. (5p)

Kvartilavståndet räknas ut $Q3 - Q1$, där $Q3$ är den tredje kvartilen och $Q1$ den första kvartilen.

Kvartilerna hittar vi genom att först hitta z-värden som motsvarar den 75:e respektive den 25:e percentilen. Därefter räknar vi ut avståndet $Q3 - Q1$. Här använder vi funktionerna qnorm och pnorm i R. Försäkra dig gärna om att du får samma svar genom att använda normalfördelningstabellen.

```
z1 <- qnorm(p=0.25)
z1
```

```
[1] -0.6744898
```

```
q1 <- 55 + z1 * 8
q1
```

```
[1] 49.60408
```

```
z3 <- qnorm(p=0.75)
z3
```

```
[1] 0.6744898
```

```
q3 <- 55 + z3 * 8
q3
```

```
[1] 60.39592
```

```
q3-q1
```

```
[1] 10.79184
```

Kvartilavståndet blir alltså ungefär 10.8.

(d.) Låddiagrammet (box plot) är ett sätt att illustrera fördelningen av provresultaten. Skalan i diagrammet är log(poäng), alltså logaritmerade poäng. Räkna återigen ut kvartilavståndet (IQR) för poäng-fördelningen, denna gång med hjälp av informationen som ges i diagrammet. Kvartilavståndet ska anges i originalskala, så kvartilerna måste transformeras från den logaritmerade skalan. Notera att värden för de streckade vertikala linjerna är inskrivna i diagrammet. (7p)

Notera: Eftersom du räknade ut kvartilavståndet även i deluppgift c bör du här få ungefär samma svar, men eftersom siffrorna i diagrammet är avrundade blir svaret inte exakt samma.

Kvartilerna representeras av gränserna för lådan, dvs $\log(Q1) = 3.9$ och $\log(Q3) = 4.1$.

Vi använder formeln $y = e^{\log(y)}$ för att hitta kvartilerna.

$Q1 = e^{\log(Q1)} = e^{3.9}$

$Q3 = e^{\log(Q3)} = e^{4.1}$

Det betyder att $IQR = Q3 - Q1 = e^{4.1} - e^{3.9} = 60.34 - 49.40 = 10.94$.

Utifrån låddiagrammet har vi beräknat kvartilavståndet till ungefär 10.9.

Uppgift 5. (20 poäng)

En glasskiosk har samlat data som visar medeltemperaturen i grader Celsius och antalet sålda glassar för ett antal dagar. De har använt denna data till att anpassa regressionsmodellen

$$\hat{y} = 132 + 17x,$$

där x är medeltemperaturen i grader Celsius en viss dag och $\hat{y}$ är det estimerade antalet sålda glassar den dagen.

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

Interceptet: Om det är noll grader estimerar vi att de säljer 132 glassar. (Tolkningen av interceptet bör i det här fallet inte ses som en realistisk prediktion. Förmodligen är glasskiosken stängd när temperaturen är noll grader.)

Lutningen: För varje ytterligare grad Celsius estimerar modellen att de säljer ytterligare 17 glassar.

(b.) Hur många glassar estimerar modellen att glasskiosken säljer en dag då det är 24 grader Celsius? (4p)

$$\hat{y} = 132 + 17x = 132 + 17 \cdot 24 = 540$$

Modeller estimerar att de säljer 540 glassar en dag då det är 24 grader varmt.

(c.) Förklara notationen och verifiera sedan med hjälp av denna information att $b_0 = 132$ och att $b_1 = 17$, avrundat till heltal.

$r_{x,y}$ är korrelationskoefficienten, s_x är standardavvikelsen för x, s_y är standardavvikelsen för y, $\bar{x}$ är medelvärdet av x, $\bar{y}$ är medelvärdet av y.

Vi kan räkna ut b_1 och b_0 på följande sätt:

$$b_1 = r_{x,y} \cdot \frac{s_y}{s_x} = 0.7 \cdot \frac{190}{7.8} = 17.05 \approx 17$$

$$b_0 = \bar{y} - \bar{x} \cdot b_1 = 439 - 18 \cdot 17.05 = 132.1 \approx 132$$

(d.) Tolka modellens samtliga tre koefficienter, dvs interceptet och de två lutningskoefficienterna. (4p)

Interceptet: Kiosken säljer 290 glassar en dag då det är noll grader och inte regn.

b1: Givet att det regnar, eller givet att det inte regnar, säljer de 8 ytterligare glassar för varje ytterligare grad Celsius.

b2: Givet en viss temperatur säljer de 206 glassar färre under dagar med regn. Alternativt, om vi håller temperaturen konstant säljer de 206 glassar färre under dagar med regn.

(e.) Om det regnar en dag har x_2 värdet 1 och annars värdet 0. Formulera regressionsmodellen så som den skulle se ut om kodningen var den omvända, dvs om dagar med regn kodades som 0 och regnfria dagar som 1. Visa med ett exempel att prediktionerna blir samma oavsett hur dummyvariabeln kodas. (4p)

Att vi 206 glassar *färre* de dagar då det regnar, givet en viss temperatur, är samma sak som att vi säljer 206 *fler* glassar dagar då det *inte* regnar. Koefficienten b_2 ska alltså vara *plus* 206.

För att resultatet ska bli samma måste vi minska interceptet med 206. Det ger modellen

$$\hat{y} = 84 + 8x_1 + 206x_2$$

För att illustrera att prediktionen för en viss dag blir samma oavsett hur dummyvariabeln är kodad kan vi exempelvis räkna ut det estimerade antalet sålda glassar för en dag med regn och 24 grader med var och en av formlerna.

```
# Med regn kodat som 1: En dag då det är 24 grader och regn.
290 + 8*24 - 206 * 1
```

```
[1] 276
```

```
# Med regn kodat som 0: En dag då det är 24 grader och regn
84 + 8 * 24 + 206 * 0
```

```
[1] 276
```

Oavsett kodning estimerar vi då att 276 glassar är sålda.