Lösningsförslag, tentamen SDA1 del 1
2024-03-26

Uppgift 1

a) Bland variablerna ovan, välj ut en kategorisk, en numerisk och en ordinal variabel. Motivera ditt svar.

Vi skulle kunna välja ut kön (kategorisk), ålder (numerisk) och betyg (ordinal).

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

- Stapeldiagram: Visar fördelningen för en kategorisk variabel. Varje stapel representerar ett antal eller en andel av observationerna som hör till en viss kategori.
- Histogram: Visar fördelningen för en numerisk variabel. Varje stapel visar antalet eller andelen observationer som ligger inom ett visst intervall.

c) (c.) Vilken percentil i en fördelning mostsvarar var och en av följande?

- Den första kvartilen motsvarar den 25:e percentilen.
- Den tredje kvartilen motsvarar den 75:e percentilen.
- Medianen (den andra kvartilen) motsvarar den 50:e percentilen.

d) En tidning ska publicera en artikel om ett politisk parti som har gått uppåt i opinionen. Tidningen har tillgång till två olika diagram, som båda visar hur partiets opinionssiffror har stigit under de senaste fyra månaderna. Vilket av diagrammen nedan bör tidningen använda om de vill att diagrammet ska leva upp till areaprincipen? Motivera.

De bör välja diagram 2.

Areaprincipen säger att arean av en stapel ska vara proportionell till det antal eller den andel som stapeln representerar. Om staplarna är lika breda, som i detta fall, betyder det att höjden på en stapel ska vara proportionell till antalet eller andelen observationer som representeras.

I diagram 1, som börjar vid 10 procent, kan vi se exempelvis att stapeln för juni är mer än dubbelt så hög som stapeln för maj, trots att ökningen från maj till juni är en ökning på mindre än 10 procent. Det bryter mot areaprincipen, och ger intrycket av en större ökning i väljarstödet än vad som faktiskt är fallet.

I diagram 2 är höjden på varje stapel, och därmed arean, proportionell till det procentuella väljarstödet.

Uppgift 3

a) När du har anpassat (fitted) en regressionsmodell kan du vilja studera residualerna. Vad är en residual, och vilka är två av de modellantaganden som du kan verifiera genom att studera residualerna?

En residual är skillnaden mellan mellan det estimerade värdet på en observation och det verkliga värdet:

$$e = y_i - \hat{y}_i$$

Genom att studera residualgrafer kan vi verifiera bland annat regressionsmodellens antaganden om att

1. att residualerna är normalfördelade
2. att residualerna har en varians som är konstant, dvs oberoende av $\hat{y}$.

b) En vanlig regressionsmodell anpassas genom minsta kvadrat-metoden. På vilket sätt är minsta kvadrat-metoden kopplad till residualerna?

Minsta kvadratmetoden säger att vi väljer den regressionslinje (dvs de regressionskoefficienter) som minimerar kvadraten av residualerna. Uttrycket som minimeras kan skrivas

$$\sum_{i=1}^{n} e^2$$

c) Du ska göra en regressionsmodell som predikterar inkomst med hjälp av en eller flera förklaringsvariabler. Du överväger två alternativa modeller[...] Nämn två metoder som du kan använda för att jämföra de båda modellerna och avgöra vilken som är lämpligast.

Vi kan jämföra modellernas adjusted r-squared. En modell med högre adjusted r-squared kan anses bättre enligt den metoden.

Ett annat sätt är att dela upp datamaterialet i träningsdata och testdata. Modellen tränas med hjälp av träningsdata och utvärderas genom att RMSE räknas ut för testdata. Lägre RMSE är bättre.

Ett mer sofistikerat sätt att dela in datamaterialet i tränings- och testdata är att göra en korsvalidering. Vid korsvalidering kan vi räkna ut RMSE för respektive modell. Lägre RMSE talar för att en modell är bättre.

d) En analytiker vill ta reda på medelinkomsten i Sverige, och har i det syftet samlat in inkomstuppgifter från 1000 slumpvis valda invånare i Sollentuna kommun. Analytikerns chef säger att undersökningen måste göras om eftersom urvalsmetod som användes kommer att medföra bias. Vad är bias, och har chefen rätt? Motivera svaret.

Bias är ett systematiskt fel som innebär att vi kan förvänta oss en felaktig skattning, oavsett hur stort urvalet är.

Chefen har rätt i att man kan förvänta bias i detta fall. Invånarna i Sollentuna kommun (eller i vilken annan kommun som helst) är knappast representativa för hela Sveriges befolkning i fråga om inkomst.

Uppgift 3

Vi har vår korstabell i en matris som vi kallar m.

```
          Ja   Nej
Annons A 506 31500
Annons B 532 33300
Annons C 287 28800
```

a) Hur många annonsvisningar gjordes totalt under kampanjen?

Antalet annonsvisningar är det summan av observationerna i korstabellen.

```
sum(m)
```

```
[1] 94925
```

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

Marginalfördelningarna är fördelningarna av variablerna *Köp* respektive *Annons* när vi betraktar dem var för sig.

Summera varje kolumn respektive varje rad, och dela sedan med det totala antalet annonsvisningar.

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

```
      Ja       Nej
1.395839 98.604161
```

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

```
Annons A Annons B Annons C
33.71715 35.64077 30.64209
```

c) Företaget vill i efterhand utvärdera hur väl de olika annonserna fungerade. För att göra detta, bör vi räkna ut fördelningen av variabeln Köp betingad på variabeln Annons, eller bör vi räkna ut fördelningen av variabeln Annons betingad på variabeln Köp? Motivera ditt svar

Vi betingar *Köp* på *Annons*, dvs vi ser hur stor andel av användarna som gör ett köp *givet* att de har sett en viss annons. I praktiken innebär det att vi undersöker varje annons för sig. För varje annons tar vi reda på hur stor andel av annonsvisningarna som resulterar i ett köp.

d)

```
# Samtliga fördelningar i procent när vi betingar på Annons
100 * m / rowSums(m)
```

```
                Ja      Nej
Annons A 1.5809536 98.41905
Annons B 1.5724758 98.42752
Annons C 0.9866951 99.01330
```

Exempeluträkning: Andelen i procent som gör ett köp givet att de ser annons A är:

$$100 \cdot 506/(506 + 31500) \approx 1.58$$

e) Tolka resultatet från deluppgift d.

Ungefär 1.6 procent av de som såg annons A eller annons B gjorde ett köp. Av de som såg annons C var det knappt en procent som gjorde ett köp, så den annonsen tycks ha fungerat något sämre än de andra.

Uppgift 4

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

Interceptet kan tolkas som att det tar 3.2 sekunder att läsa en text som innehåller 0 ord. Detta ska inte ses som en bokstavlig tolkning.

Lutningskoefficienten kan tolkas som att modellen estimerar att varje ytterligare ord i texten medför att den tar ytterligare 0.34 sekunder att läsa.

b) Förklara notationen och verifiera koefficienterna.

r_{xy} är sample-korrelationen mellan textens längd (x) och tiden i sekkunder som det tar att läsa texen (y).

s_x och s_y är sample-standardavvikelsen för respektive variabel.

$\bar{x}$ och $\bar{y}$ är variablernas medelvärden.

För att verifierakoefficienterna kan vi använda följande formler:

$$b_1 = r_{xy} \cdot \frac{s_y}{s_x} = 0.9 \cdot \frac{34}{90} = 0.34$$

$$b_0 = \bar{y} - b_1\bar{x} = 248 - 0.34 \cdot 720 = 3.2$$

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

- Interceptet: För en person som inte har deltagit i snabbläsningskursen tar det 4.1 sekunder att läsa en text som innehåller 0 ord. Tolka ej bokstavligt.
- För varje ytterligare ord i texten tar det ytterligare 0.4 sekunder att läsa den, *givet värdet på* x_2.
- Det tar 24 sekunder mindre att läsa en text för den som gått snabbläsningskursen, *givet antalet ord i texten.*

d) Hur många sekunder estimerar denna modell att det tar att läsa en text som består av 700 ord?

$\widehat{\log(y)} = 0.5 + 0.007x = 0.5 + 0.007 \cdot 700 = 5.4$

$\hat{y} = e^{\widehat{\log(y)}} = e^{7.5} \approx 221$

Modellen estimerar att det tar ungefär 221 sekunder att läsa en text som innehåller 700 ord.

e) Räkna ut R-squared för modellen.

$$R^2 = 1 - \frac{SSE}{SST}$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 30000$$

$$SSR = SST - SSE = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 25000$$

$$SSE = SST - SSR = 30000 - 25000 = 5000$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{5000}{30000} \approx 0.83$$

Uppgift 5

a Leah sprang loppet på 255 minuter. Hur stor andel av löparna hade en lägre tid än Leah?

Vi räknar ut z-värdet som motsvarar 255 minuter, och sedan ser vi vilken andel av alla observationer som har ett lägre z-värde i en standard normalfördelning.

```
ybar <- 235
sdy <- 40
y0 <- 255
z <- (y0 - ybar) / sdy
z
```

```
[1] 0.5
```

```
pnorm(q=z)
```

```
[1] 0.6914625
```

Tiden 255 minuter motsvaras av z-värdet 0.5. I en normalfördelningstabell kan vi se att ungefär 69 procent av alla observationer i en normalfördelning har ett lägre z-värde än 0.5. Därmed hade 69 procent av löparna i loppet en lägre tid än 255 minuter.

b) Hur många minuter tog loppet att springa för en löpare vars tid befann sig vid den 10:e percentilen i fördelningen?

```
z <- qnorm(p=0.1)
z
```

```
[1] -1.281552
```

```
ybar + z * sdy
```

```
[1] 183.7379
```

Om 10 procent av löparna har en lägre tid så är z-värdet ungefär -1.28. Det motsvarar på ett ungefär tiden 184 minuter.

c)

```
# z25 är z-värdet för första kvartilen.
z25 <- qnorm(p=0.25)
z25
```

```
[1] -0.6744898
```

```
# z75 är z-värdet för tredje kvartilen.
z75 <- qnorm(0.75)
z75
```

```
[1] 0.6744898
```

```
# y25 är tiden i minuter för första kvartilen.
y25 <- ybar + z25 * sdy
y25
```

```
[1] 208.0204
```

```
# y75 är tiden i minuter för tredje kvartilen.
y75 <- ybar + z75 * sdy
y75
```

```
[1] 261.9796
```

```
# Kvartilavståndet
y75 - y25
```

```
[1] 53.95918
```

Den första kvartilen ligger vid 208 minuter och den tredje kvartilen vid 262 minuter.

Kvartilavståndet blir ungefär $262 - 208 = 54$ minuter.

d

- **a** är den första kvartilen, som vi vet från uppgift c är 208 minuter.
- **b** är medianen, som i en normalfördelning är samma som medelvärdet, dvs 235 minuter.
- **c** är den tredje kvartilen, som vi i uppgift c räknade ut är 262 minuter.