Statistiska metoder

Inlämningsuppgift

Iris virginica

Introduktion

Denna inlämningsuppgift kommer att behandla Edgar Anderson's Iris Dataset.
Herr Anderson har mätt (i centimeter) längden och bredden på folderblad och kronblad på 150 Irisblommor av arterna är setosa, versicolor och virginica.

Datasetet är en dataframe av storleken 150 rader och 5 kolumner
Kolumnerna innefattar variablerna foderbladslängd (Sepal.Length), foderbladsbredd Sepal.Width), kronbladslängd (Petal.Length), kronbladsbredd (Petal.Width) och art (Species).

Datatyper är för foderbladslängd, foderbladsbredd, kronbladslängd, kronbladsbredd kvantitativ data, med undergrupp: oavbrutet/ ratio eftersom det är mätbart, en längd och innehåller decimal.
Datatyp för art är kvalitativt, med undergrupp nominal eftersom det är namnet på en art och kan inte jämföras matematiskt.

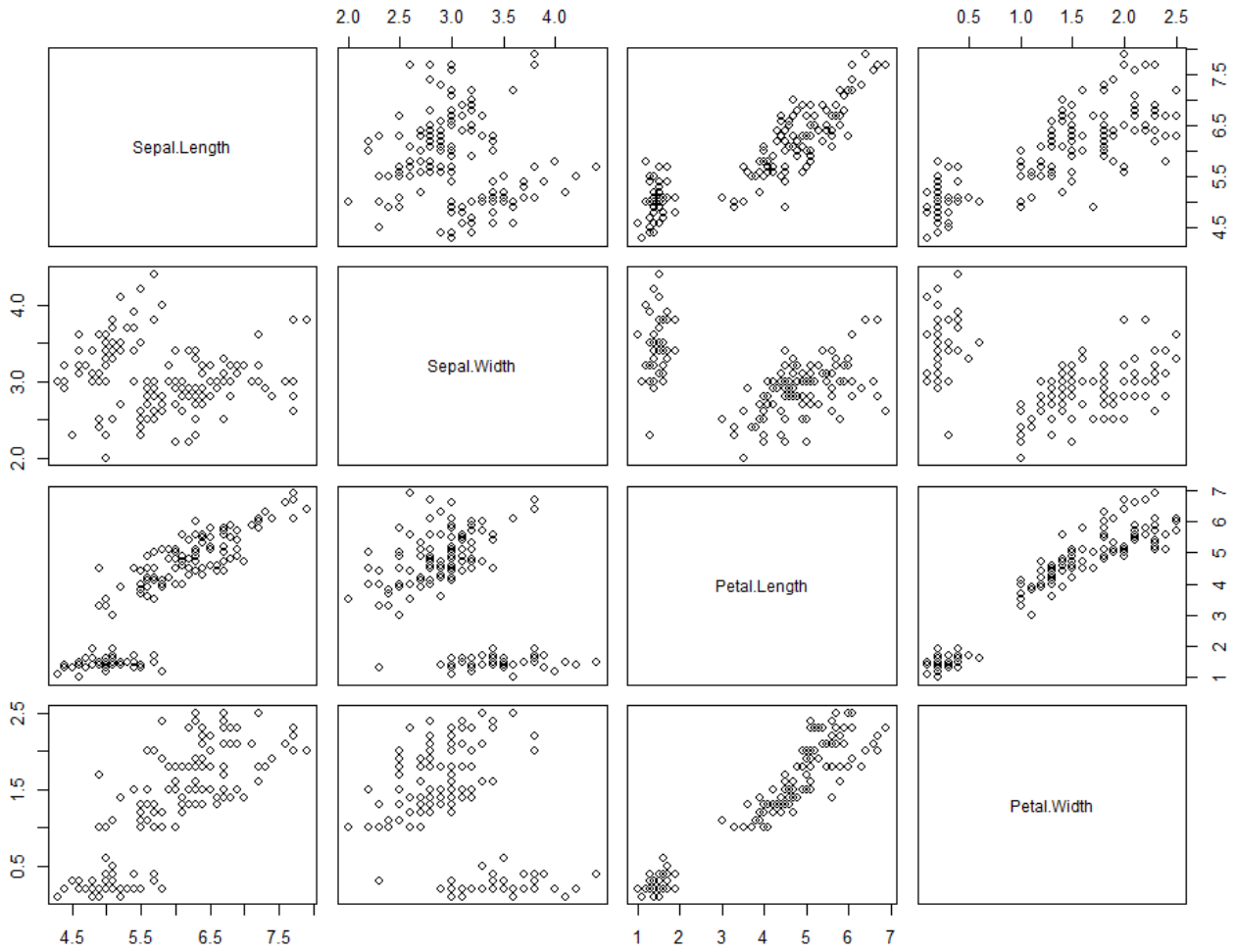
Databasen finns tillgänglig på Rstudios egna bibliotek men kan även hittas på denna webbadress:
https://www.kaggle.com/datasets/arshid/iris-flower-dataset
Iris dataset är väl analyserat och används av data analytiker för att övningsmaterial. Det har inte manipulerats av mig då den redan är processad och har inga saknade värden.

Frågeställning:

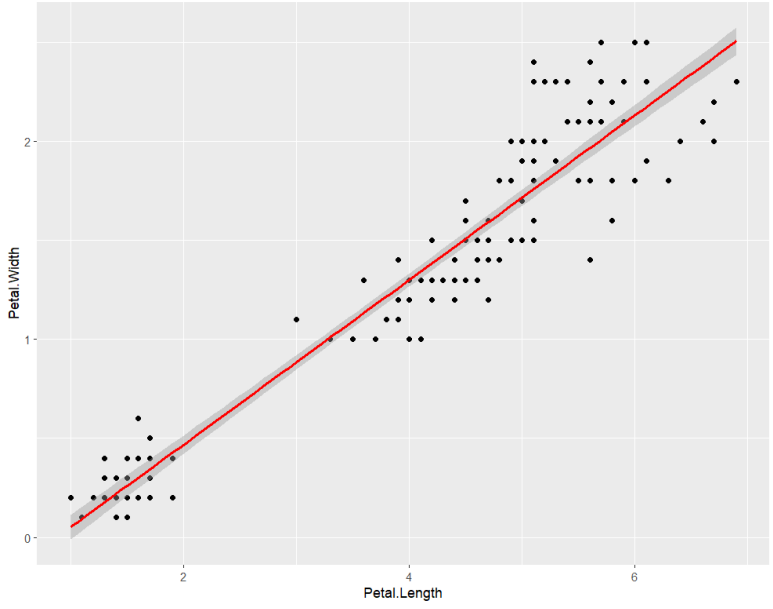
Min vetenskapliga fråga lyder:
“Finns det någon korrelation mellan kronbladens längd och kronbladens bredd?”

Scatterplot

Scatterplot skapas för att se om det finns en svag eller stark, eller ingen korrelation mellan de olika kombinationerna av variabler.

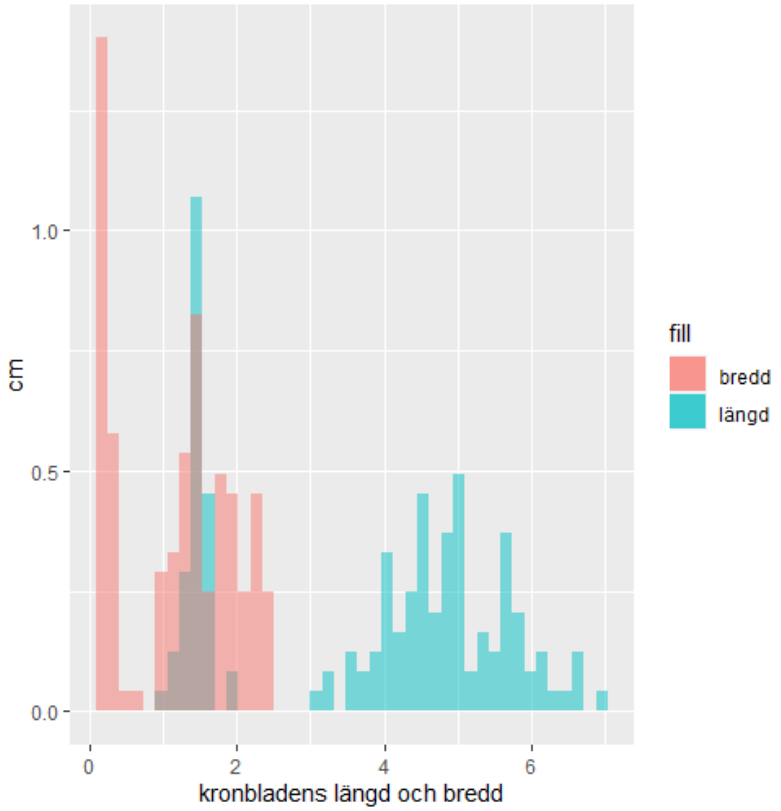


Här är en scatterplot på kronbladens längd och bredd.
Frågan är om förhållandet mellan kronbladens längd och bredd är i proportion till varandra. Med en medelvärdeslinje. Vi kan se att det finns en stark positiv korrelation.

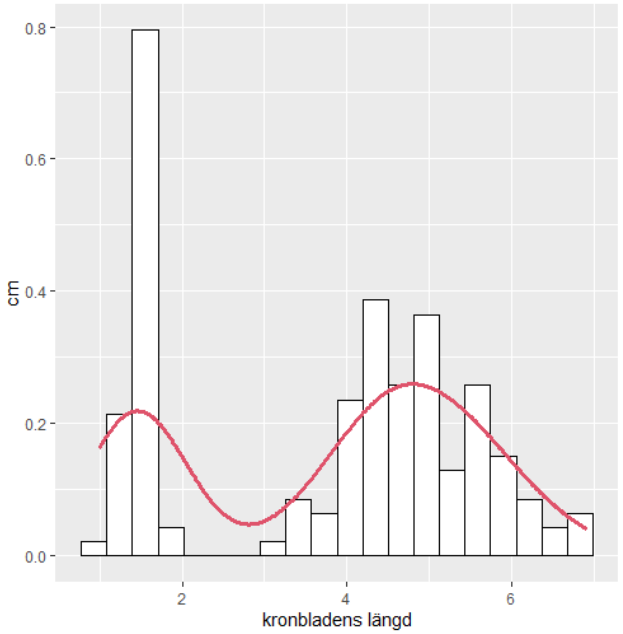


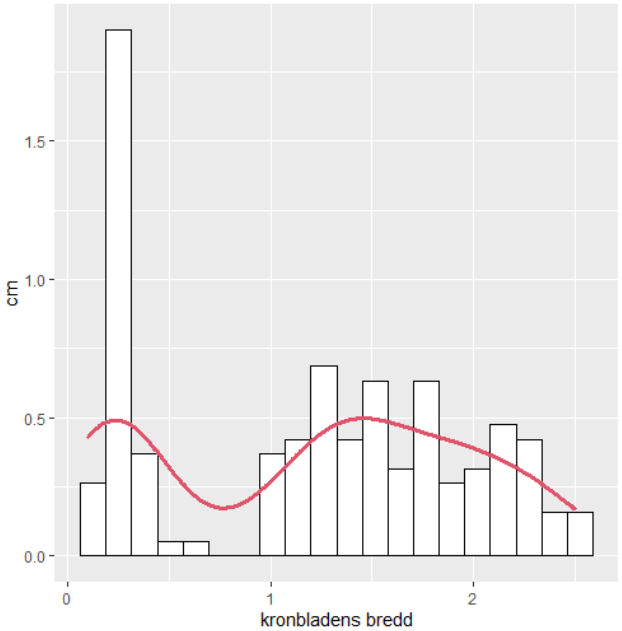
Histogram

Kan vi med hjälp av ett histogram se om det finns en korrelation mellan kronbladens längd och dess bredd?
Vi kan antyda att det finns ett återkommande trend hos de båda variablerna.



Ännu tydligare blir mönstret när vi jämför med linjen för utjämnade densitet-uppskattningar. Vi ser att båda histogrammen har bimodala linjer med en ojämn spridning men där båda linjernas kurvor liknar varandra.





Korrelationskoefficienten (r)

För att räkna ut korrelationen mellan de två variablerna gör vi en uträkning för korrelationskoefficienten. Det högsta och minsta möjliga siffran är -1 till +1, där -1 är en perfekt negativ korrelation och +1 är en perfekt positiv korrelation.

Formeln för r:

$$(r) = \frac{n * (\sum xy) - (\sum x) * (\sum y)}{\sqrt{[n * \sum x^2 - (\sum x)^2] * [n * \sum y^2 - (\sum y)^2]}}$$

Kronbladens längd räknas som x (Petal.Length) - det självständiga värdet
Kronbladens längd räknas som y (Petal.Width) - det osjälvständiga värdet

n (antal samplingar) = 150
∑xy 869.11
∑x 563.7
∑y 179.9
∑x² 2582.71
∑y² 302.33

$$0.9628654 = \frac{150 * (869.11) - (563.7) * (179.9)}{\sqrt{[150 * 2582.71 - (563.7)^2] * [150 * 302.33 - (179.9)^2]}}$$

Korrelationskoefficienten = 0.9628654

Eller räkna ut i r
cor(iris$Petal.Length, iris$Petal.Width)
[1] 0.9628654

Bestämningskoefficienten (CD)

Denna beräkning används för att se om linjen för minsta kvadrat-linje ska användas. För denna uträkning behöver vi veta korrelationskoefficienten (r). vilket räknades ut ovan och som är: 0.9628654

Formeln för CD:

$CD = r^2 * 100$ (100 för att det är i procent)

$92.71098 = 0.9628654^2 * 100$

≈ 93%

Bestämningskoefficienten bör ligga över en gräns på 50% för att det ska finnas ett samband mellan variablerna.

Minsta kvadrat-linjen

För att vara helt säkra på att vi kan använda oss av uträkningen ovan kontrollräknar vi med hjälp av minsta kvadrat-linjen.
För att räkna ut minsta kvadrat-linje behöver vi följande uppgifter:

$\sum xy$ 869.11
$\sum x$ 563.7
$\sum y$ 179.9
$\sum x^2$ 2582.71
$\bar{x}$ 3.758
$\bar{y}$ 1.199333

Formel för minsta kvadrat-linje:

$$b = \frac{n * \sum xy - (\sum x) * (\sum y)}{n * \sum x^2 - (\sum x)^2}$$

När vi vet b (lutningen) kan vi räkna ut a (riktningen) med hjälp av att multiplicera medelvärdet av x och subtrahera med medelvärdet av y.

$a = \bar{y} - b * \bar{x}$

Vi räknar ur uppskattningen av y ($\hat{y}$). Var vi tror att y bör vara i förhållande till x. Vi multiplicerar in ett värde från x (kronbladens längd) med lutningen (b) och adderar sedan med riktningen (a).

$\hat{y} = b * x + a$

Tillsist räknar vi ut vad resten av summan av y och det uppskattade y:et ($\hat{y}$). Om resultatet blir litet är linjen bra.

$y - \hat{y}$

Uträkningen av minsta kvadrat för kronbladens bredd och längd

Räkna ut b:

$$0.4157554 = \frac{150 * 869.11 - (563.7) * (179.9)}{150 * 2582.71 - (563.7)^2}$$

b = 0.4157554

Räkna ut a:

$-0.3630758 = 1.199333 - 0.4157554 * 3.758$

a = -0.3630758

Räkna ut den uppskattade summan av y

$0.9451334 = 0.4157554 * 5.6 + 0.3630758$

$\hat{y}$ = 2.691306

Räkna ut rest av y jämfört med uppskattningen av y

$-0.891306 = 1.8 - 2.691306$

I det här fallet är skillnaden mellan y och det uppskattade y:et ca -0.89 cm vilket är en liten rest, som antyder att linjen är bra.

Med R blev uträkningen:

```
plot(y = iris$Petal.Width, x = iris$Petal.Length, pch =20, main="Minsta kvadrat-linje",
xlab = "Kronbladens Längd",ylab = "Kronbladens Bredd")

coef (iris.model)
abline (a = coef (iris.model) [1],
        b = coef (iris.model) [2],
        col = "red")
```

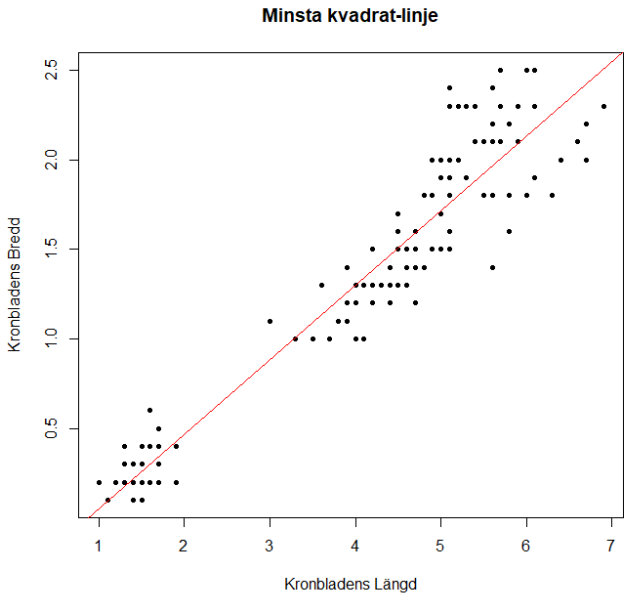
```
(Intercept) Petal.Length
 -0.3630755    0.4157554
```

Med funktionen coef skriver den ut koefficienten för vald plot.
Den adderade linjen har den uträknade koefficient som grund för a och b.

[1] ger riktning och [2] ger lutningen.

(Intercept) -0.3630755 (motsvarar “a” som är riktningen på linjen)

Petal.Length 0.4157554 (motsvarar “b” som är lutningen)



Slutsats

Svaret på frågan om det finns någon korrelation mellan kronbladens längd och kronbladens bredd är att det finns en stark korrelation mellan kronbladens längd och bredd. Det, eftersom den minsta kvadratens linje var mycket liten. Desto större rest, desto större kvadrater (längre från linjen), därmed en illa passande linje. I detta fall är kvadraten liten och är nära linjen, därmed en bra passande linje.
Den uträknade bestämningskorrelationen (CD) är 93 %. Uträkningen tyder på en stark koppling mellan kronbladens längd och dess bredd. Då 100% är minsta möjliga avstånd från ett samband och 50% är en gräns för då en korrelation existerar, är alltså 93% en stark korrelation.
Bladets längd är i proportion till dess bredd.