MASB11: BIOSTATISTISK GRUNDKURS
DATORLABORATION 4, 22 MAJ 2019
REGRESSION OCH FORTSÄTTNING PÅ MINIPROJEKT II

Syfte

Syftet med dagens laboration är att du ska

- bekanta dig med lite av de funktioner som finns i R vad det gäller korrelations- och regressionsanalys
- arbeta med Miniprojekt II.

Förberedelseuppgifter

Du måste ha arbetat ordentligt med de väsentligaste begreppen i kapitlet om regression i kursboken. Repetera vid behov begreppen +emphregresionslinje, residualer, konfidensintervall för förväntat värde samt *prediktionsintervall*.

Du skall ha gjort följande uppgift *innan* du kommer till laborationen.

Hemuppgift 1:

Gör uppgift 6.10 i arbetsmaterialet (ingår på övningen 16 maj).

1 Introduktion - regressionsanalys i R

1.1 Datamaterial: torskar

För 10 torskar har vi värden på variablerna Längd (cm) och Ålder (år).

Fisk	1	2	3	4	5	6	7	8	9	10
Längd (cm)	15	30	35	50	55	60	58	25	12	43
Ålder (år)	1	2	3	4	5	6	5	2	1	4

Kan vi påvisa något (linjärt) samband mellan längd och ålder?

Inläsning av data

Börja med att mata in data till R. Lägg in värdena i en dataframe, som du döper till `torskar`, med värdena i två kolumner: Längd och Ålder:

```
torskar <- data.frame(Längd=c(15,30,35,50,55,60,58,25,12,43),
                      Ålder=c(1,2,3,4,5,6,5,2,1,4))
```

Datamaterialet skall alltså innehålla 2 kolumner med 10 värden i varje kolumn.

Beskrivning av data

Gör en grafisk beskrivning av sambandet genom att rita ett spridningsdiagram med Ålder på x-axeln och Längd på y-axeln:

```
plot(torskar$Ålder, torskar$Längd)
```

Ser det ut som det finns ett linjärt samband?

Korrelationer

Vi börjar med att beskriva sambandet mellan variablerna med hjälp av korrelationskoefficienten. Beräkna korrelationskoefficienten (Pearson) och testa om den är skild från noll med

```
cor(torskar$Ålder, torskar$Längd)
cor.test(torskar$Ålder, torskar$Längd)
```

Uppgift 1.1:

Tyder resultaten på att det finns något linjärt samband mellan Längd och Ålder?

Enkel linjär regression

Vi skall nu undersöka hur sambandet mellan variablerna ser ut genom att anpassa en rät linje till data. Kommandot `lm(y ~ x)` anpassar en linjär modell för den beroende variabeln `y` som funktion av en eller flera förklarande variabler `x` (lm är förkortning av linear model). Sedan kan vi få ut olika egenskaper hos modellen och skattningarna med ytterligare kommandon:

```
modell <- lm(Längd ~ Ålder, data=torskar)
modell           # skattningarna av beta0 och beta1
summary(modell)  # mer information, t.ex. signifikanser för skattningarna
confint(modell)  # konfidensintervall för beta0 och beta1
```

Uppgift 1.2:

Gör analysen. Identifiera följande mått i utskriften: r — korrelationskoefficienten, R^2 — förklaringsgraden, s —- residualspridningen, de skattade koefficienterna med standardfel, t-test och konfidensintervall.

För att få den skattade regressionslinjen utritad i figuren ni skapade tidigare kan du använda kommandot

```
abline(modell)
```

Prognoser och konfidensintervall

Om man vill använda sin regressionsmodell för att göra prognoser så kan detta enkelt göras efter att man skattat modellen. I R kan man prediktera i samma datamaterial som man använde för att

skatta modellen. Då får man en prediktion för varje individ=rad. Man kan också mycket enkelt ange en helt annan uppsättning individer som man vill prediktera för istället. Det är praktiskt när man t.ex. vill rita ut konfidensintervall och prediktionsintervall snyggt.

```
# vi vill prediktera för en ålderssekvens i steg om halvår:
# 0.5, 1.0, 1.5, ..., 7.0, 7.5:
x0 <- data.frame(Ålder=seq(0.5,7.5,0.5))
mu0konf <- predict(modell, x0, interval=''confidence'')  # konfidensintervall
mu0pred <- predict(modell, x0, interval=''prediction'')  # prediktionsintervall
cbind(x0, mu0pred)
```

Uppgift 1.3:

Vad blir prognosen för längden för en sju år gammal torsk? Vad blir prognosintervallet?

För att få intervallen utritade i figur ritar vi linjer med våra prediktionsåldrar på x-axeln och tillhörande intervallgränser på y-axeln. Vi vill dessutom rita konfidensintervallet som streckade blå linjer och prediktionsintervallet som prickade röda:

```
lines(x0$Ålder, mu0konf[,''lwr''], col=''blue'', lty=2)  # undre (lower) gränsen
lines(x0$Ålder, mu0konf[,''upr''], col=''blue'', lty=2)  # övre (upper) gränsen
lines(x0$Ålder, mu0pred[,''lwr''], col=''red'', lty=3)
lines(x0$Ålder, mu0pred[,''upr''], col=''red'', lty=3)
```

Kontroll av förutsättningar

Vi skall nu kontrollera två av de antagande som finns i analysen. För det första antagandet om normalfördelning och för det andra antagandet om lika varianser. Residualerna beräkna med kommandot

```
res <- residuals(modell)
```

Undersök nu om residualerna är normalfördelade genom att göra en Q-Q-plot (`qqnorm`).

Antagandet om lika varianser (konstant spridning kring linjen) kan vi undersöka genom att plotta residualerna mot Ålder.

```
plot(torskar$Ålder, res)
```

Uppgift 1.4:

Verkar antagandena om normalfördelning och konstant varians uppfyllda?

1.2 Datamaterial: Bradfordmetoden

I laborationen ”Proteinbestämning enligt Bradford-metoden” i kursen cellbiologi undersöktes absorbansen hos prov med olika spädningar av Bovint Serum Albumin (BSA)-standard. Prov med 0–10 μg protein spädes till 100 μl med vatten och två prover förberedes per koncentration.

Data för en laborationsgrupp finns i filen `Labbdata.RData` som du hittar på kursens hemsida.

Modell: Enligt Lambert-Beers lag gäller att absorbansen (A) kan beskrivas som en linjär funktion av koncentrationen (c): $A = k \cdot c$ där konstanten k beror på ämnets molära absorptionskoefficient vid en viss våglängd samt kyvettens längd. Vid mätningar får man naturligtvis räkna med en viss slumpmässig variation, en rimlig modell är att absorbansen vid mätning nr i, A_i, beskrivs linjärt av koncentrationen ci plus ett slumpmässigt fel:

$$A_i = \beta_0 + \beta_1 \cdot c_i + \varepsilon_i$$

där ε_i är oberoende och normalfördelad slumpfel med väntevärde 0 och standardavvikelse σ. Här motsvaras konstanten β_1 av den tidigare k medan β_0 är absorbansen i den lösning som BSA:n är löst.

Uppgift 1.5:

Undersök på labbdata om den linjära regressionsmodellen ovan är rimlig att anpassa.

Uppgift 1.6:

Hur mycket ökar absorbansen då man ökar koncentrationen en enhet? Ange ett 95 % konfidensintervall för denna storhet.

Uppgift 1.7:

Vad är genomsnittlig absorbans för prov med koncentration 50 (mg/l). Ange ett 95 % konfidensintervall för denna storhet. Skapa först en dataframe där värdet för koncentrationen läggs in:

```
x50 <- data.frame(koncentration=c(50))
mu50konf <- predict(modell2,x50,interval="confidence")
```

Uppgift 1.8:

Vi har ett prov med koncentration 50 (mg/l). Ange ett 95 % prediktionsintervall för absorbansen i just detta prov.

Huvudsyftet med mätningarna var att erhålla en standardkurva för hur absorbansen påverkas av koncentrationen. Anta att vi på ett prov med okänd koncentration c_0 uppmätte absorbansen 0.43. En skattning av c_0 kan vi få fram genom att lösa ut x ur sambandet $0.43 = \beta_0 + \beta_1 \cdot x$ så här (om den anpassade modellen sparats i variabeln `modell2`):

```
beta0 <- modell2$coefficients[1]
beta1 <- modell2$coefficients[2]
c0 <- (0.43 - beta0) / beta1
c0
```

Uppgift 1.9:

Vad blev den skattade koncentrationen?

Fortsätt med att göra klart Miniprojekt II, se laboration 3

Svar

1.1 Ja! $r = 0.9828, t = 15.0709$; P-värde=0.000 – Man kan förkasta hypotesen om inget samband

1.2 $R^2 = 0.966; r = 0.9828; s = 3.443; \beta_0 = 5.993(2.4044); \beta_1 = 9.790(0.6496); t = 15.071; p = 3.72 \cdot 10^{-7}$

1.3 Prognos vid åldern 7 år = 74.52; prediktionsintervall är $(64.52,\ 84.53)$

1.4 NF: Njä; Konstant varians: Ej helt lätt att bedöma (få värden)

1.6 0.0008 intervall: $(0.00063,\ 0.0011)$

1.7 konfidensintervall: $(0.425,\ 0.441)$

1.8 prediktionsintervall: $(0.407,\ 0.459)$

1.9 c_0 skattas till 47 mg/l

Sammanfattning R

```
cor(x, y)                        Korrelationskoefficient
cor.test(x, y)                   Test för korrelationskoefficient
lm(y ~ x)                        Regression av y som funktion av x
lm(y ~ x, data=dataframen som innehåller x och y) ... i ett visst datamaterial
summary(modell)                  Skattningar, signifikanser, etc,
confint(modell)                  Konfidensintervall för parametrarna
predict(modell, x0)              Prediction av förväntat värde när x=x0
predict(modell, x0, interval=''confidence'') ... med konfidensintervall
predict(modell, x0, interval=''prediction'') ... med prediktionsintervall
residuals(modell)                Residualer
```