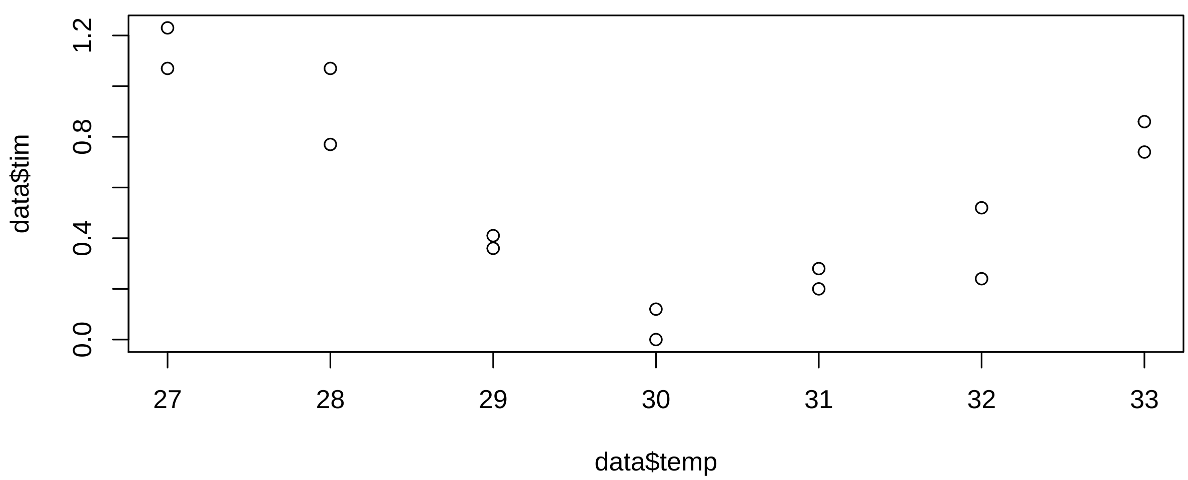
SD2 - uge 3, tirsdag

Anne Petersen

Opgave 3.6 i dokument fra Absalon

Vi starter med at loade data og plotte tid til udklækning mod temperatur:

```
setwd("C:/Users/zms499/Dropbox/Arbejde/STATforLIFE2/uge3")
data <- read.table("bananfluer.txt", header=T)
plot(data$temp, data$tim)
```



Vi ser, at punkterne ser ud til at følge en parabel (et andengradspolynomium). Altså virker det fornuftigt at modellere med en kvadratisk effekt af `temp`. Da parablen er konveks ("glad"), forventes en positiv β_2-værdi, hvis modellen skrives som i opgaveformuleringen, dvs.

$$Y_i = \beta_0 + \beta_1 \cdot \text{temp}_i + \beta_2 \cdot \text{temp}_i^2 + e_i$$

Bemærk, at denne model klart er lineær i den forstand, der beskrives i bogen. Dvs. at den er lineær i parametrene (β_0, β_1 og β_2), hvilket betyder, at disse parametre indgår i forskellige led (plusset sammen), og evt. ganges med andre ting. Bemærk desuden, at modellen ikke er lineær i `temp` (da denne variabel indgår kvadratisk), men at det heller ikke er sådan, vi karakteriserer en lineær model.

Vi fitter modellen fra ovenfor og betragter parameterestimaterne. Bemærk, at vi bruger `I()`, når vi anvender matematiske funktioner i modelformularudtrykket i `lm()` - ellers forstår R ikke, at den skal opfatte `temp^2` som at vi gerne vil benytte den matematiske funktion `^2` på `temp`.

```
model <- lm(tim ~ temp + I(temp^2), data)
summary(model)
```

```
##
## Call:
```

```
## lm(formula = tim ~ temp + I(temp^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19857 -0.06714 -0.00946  0.06518  0.34536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 84.43964   10.63784   7.938 7.04e-06 ***
## temp        -5.53482    0.71145  -7.780 8.51e-06 ***
## I(temp^2)    0.09089    0.01185   7.669 9.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1536 on 11 degrees of freedom
## Multiple R-squared:  0.8713, Adjusted R-squared:  0.8479
## F-statistic: 37.24 on 2 and 11 DF,  p-value: 1.267e-05
```

```r
#Alternativ metode, hvor vi gemmer temp2-variablen før vi bruger den:
data$temp2 <- data$temp^2
model <- lm(tim ~ temp + temp2, data)
```

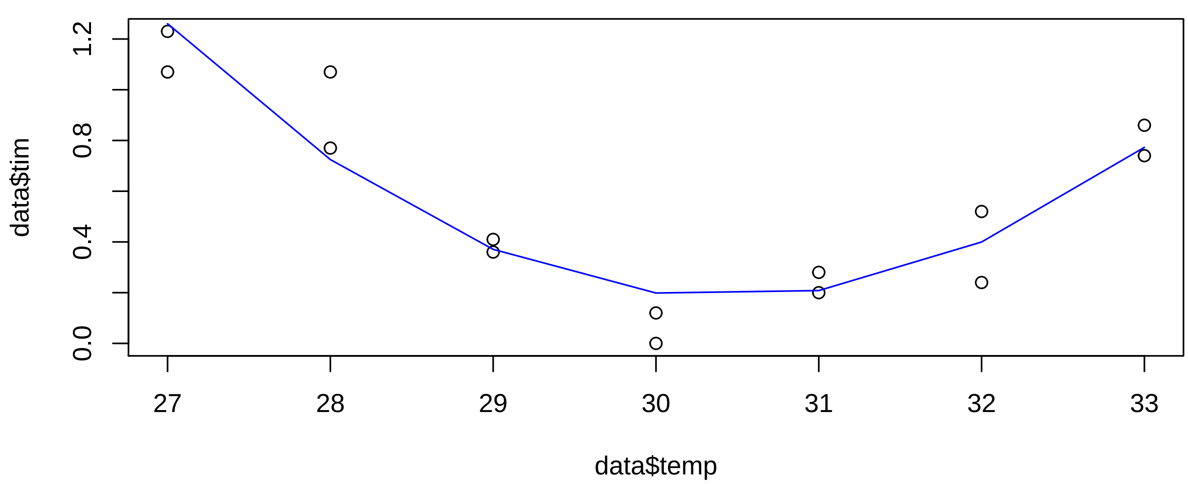
Vi ser at

$$\hat{\beta}_0 = 84.44,\ \hat{\beta}_1 = -5.53,\ \hat{\beta}_2 = 0.09$$

og at $s = 0.1536$.

Vi plotter modellens estimerede kurve oveni plottet fra ovenfor:

```r
plot(data$temp, data$tim)
points(data$temp, fitted(model), type="l", col="blue")
```



```
#Bemærk: fitted(model) og predict(model) giver samme resultat
```

En lille kommentar til syntaksen, dvs. måden koden hænger sammen på, i kommandoen `points()`: Først angives x-værdierne (`temp`), dernæst angives modellens forudsigelser af y-værdierne (`fitted(model)`), til sidst angives det, at vi gerne vil have tegnet en linje (`type="l"`) og at den skal være blå (`col="blue"`).

Vi vil nu beregne et estimat for temperaturen hvor udklækningstiden er mindst i følge modellen. Vi skal altså finde minimum for det andengradspolynomium som modellen definerer. Vi starter med at pille parameterestimaterne ud af modellen og gemme dem som variable:

```
beta0 <- summary(model)$coefficients[1,1]
beta1 <- summary(model)$coefficients[2,1]
beta2 <- summary(model)$coefficients[3,1]
```

Eftersom $\hat{\beta}_2 > 0$, ved vi, at det globale minimum for funktionen $f(x) = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2$ findes ved $f'(x) = 0$ - det er det eneste sted, vi kan finde en vandret tangent. Altså differentieres funktionen:

$$f'(x) = \beta_1 + 2 \cdot x \cdot \beta_2$$

og vi finder funktionens rod ved at sætte den lig nul og løse ligningen for x:

$$f'(x) = 0 \;\Leftrightarrow\; x = -\frac{\beta_1}{2 \cdot \beta_2}$$

Altså findes minimum ved

$$x_0 = -\frac{\beta_1}{2 \cdot \beta_2}$$

og dermed er temperaturen for den mindst mulige udklækningstid i følge modellen 30.44695:

```
-beta1/(2*beta2)
```

```
## [1] 30.44695
```

Bemærk at dette resultat stemmer fint overens med det, vi ser på plottet ovenfor.

Vi fitter nu en ensidet variansanalysemodel hvor temperatur bruges som faktor:

```
fact_model <- lm(tim ~ factor(temp), data)
```

Denne model ses klart at være mere generel end modellen ovenfor: Her siger vi blot, at der er en eller anden sammenhæng mellem tid og temperatur - ovenfor specificerede vi, at denne sammenhæng *skulle* være kvadratisk. Hvis vi sætter $\beta_2 = 0$ og restringerer β_1 til særlige værdier, fås netop den nye model ud af den gamle. Altså er den kvadratiske model indeholdt i variansanalysemodellen. Vi kan derfor teste de to modeller mod hinanden vha. en F-test:

```
anova(model,fact_model)
```

```
## Analysis of Variance Table
##
## Model 1: tim ~ temp + temp2
## Model 2: tim ~ factor(temp)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     11 0.25961
## 2      7 0.11585  4   0.14376 2.1717 0.1744
```

og vi finder $p = 0.1744 > 0.05$. Altså må (bør) vi gennemføre modelreduktion til den simplere, kvadratiske regressionsmodel.

Til sidst bliver vi bedt om at teste hvorvidt det er tilstrækkeligt at `temp` indgår lineært i modellen, selvom plottet ikke ligefrem peger i den retning. Vi fitter derfor en model uden det kvadratiske led og tester denne model mod `model`, som har et kvadratisk led:

```r
lin_model <- lm(tim ~ temp, data)
anova(lin_model, model)
```

```
## Analysis of Variance Table
##
## Model 1: tim ~ temp
## Model 2: tim ~ temp + temp2
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     12 1.64755
## 2     11 0.25961  1    1.3879 58.807 9.744e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bemærk, at vi her tester hypotesen

$$H : \beta_2 = 0$$

og at vi finder $p = 9.7 \cdot 10^{-6}$, og dermed forkaster hypotesen. Vi konkluderer altså, at der er en signifikant effekt af det kvadratiske led (som forventet).