SD2 - uge 3, torsdag

Anne Petersen

Opgave 3.A (6.2 i BMS)

Vi starter med at loade data kigge på de første 6 observationer:

```
setwd("C:/Users/zms499/Dropbox/Arbejde/STATforLIFE2/uge3")
data <- read.table("steers2.txt", header=T)
head(data, 6)
```

```
##   block treat   x    y
## 1     1     1 560 1330
## 2     1     2 440 1280
## 3     1     3 530 1290
## 4     1     4 690 1340
## 5     2     1 470 1320
## 6     2     2 440 1270
```

Bemærk at datasættet består af tre forklarende variable (`block`, `treat` og `x`) og en responsvariabel (`y`). De to første forklarende variable er faktorer, så vi starter med at gemme dem som sådanne i R:

```
data$block <- factor(data$block)
data$treat <- factor(data$treat)
```

Vi bliver bedt om at undersøge, om der er en effekt af `treat` på `y` (som måler vægten af nyrefedt). Vi opstiller derfor den mest generelle model vi kan, og ser og der er en signifikant effekt af `treat` i denne model. Bemærk, at det ikke er muligt at opstille en model med vekselvirkning mellem `treat` og `block`, da vi kun har én observation pr. kombination af de to faktorer. Vi opstiller i stedet en additiv model, hvor alle de tre forklarende variable indgår:

$$Y_i = \alpha(\text{block}_i) + \beta(\text{treat}_i) + \gamma \cdot x_i + e_i$$

for $i = 1...16$ og hvor det antages at e_i'erne er iid med $e_1 \sim N(0, \sigma^2)$. Y_i repræsenterer her vægten af den i'te studs (steer) nyrefedt, treat_i angiver dens behandling, block_i angiver blokken og x_i angiver dens vægt før behandlingen. Vi vil gerne teste hvorvidt der er effekt af `treat`, dvs. vi vil teste hypotesen

$$H: \ \beta(\text{treat}_1) = \beta(\text{treat}_2) = ... = \beta(\text{treat}_4)$$

Vi fitter derfor dels den fulde model, dels modellen, hvor `treat` er udeladt i R og tester de to modeller mod hinanden:

```
model_full <- lm(y ~ block + x + treat, data)
modelB <- lm(y ~ block + x, data)
anova(modelB, model_full)
```

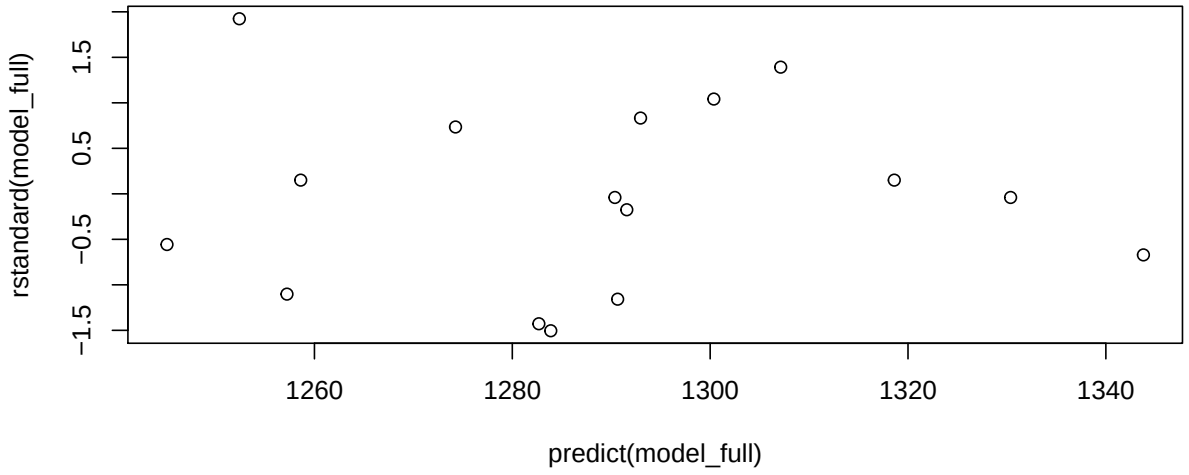
```
## Analysis of Variance Table
##
## Model 1: y ~ block + x
## Model 2: y ~ block + x + treat
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     11 3622.4
```

```
## 2      8 1216.7  3    2405.8 5.2729 0.02676 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi finder $p = 0.02676 < 0.05$ og altså afviser vi H og modellen kan ikke reduceres. Det konkluderes dermed, at der er en signifikant effekt af hormonbehandlingen.

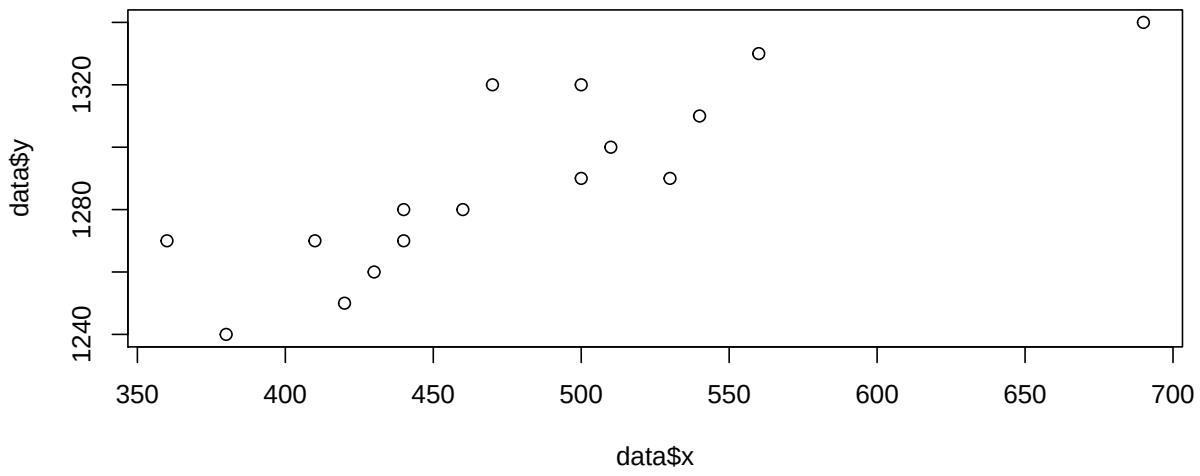
Vi vil nu se, om vi kan identifere en enkel observation, som har uforholdsmæssig stor indflydelse på modellen, dvs. en outlier. Vi starter med at betragte et residualplot for at se om der her er nogen observationer, som er påfaldende:

```
plot(predict(model_full), rstandard(model_full))
```



Vi ser ikke umiddelbart nogen outliers her. Vi prøve i stedet at plotte variablene `x` og `y` mod hinanden:

```
plot(data$x,data$y)
```



Vi ser her, at der er en x-værdi, som er meget langt fra de øvrige x-værdier. Vi kan identificere denne vha. `boxplot()`-funktionen:

```
outlier <- boxplot(data$x, plot=F)$out
#bemærk: plot=F fordi vi ikke behøver kigge på boxplottet
```

Vi fjerner nu denne outlier fra datasættet og prøve at køre analysen igen for at se, om det påvirker konklusionerne:

```
data2 <- subset(data, x!=outlier)
model_full2 <- lm(y ~ factor(block) + factor(treat) + x, data=data2)
model_B2 <- lm(y ~ factor(block) + x, data=data2)
anova(model_B2, model_full2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ factor(block) + x
## Model 2: y ~ factor(block) + factor(treat) + x
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     10 3400.5
## 2      7 1148.2  3    2252.4 4.5773 0.04468 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi får nu $p = 0.04468 < 0.05$ og altså hælder vi mod samme overordnede konklusion som før, dvs. at der er en effekt af `treat`. Men bemærk, at p-værdien nu er lige på grænsen til den modsatte konklusion.

Vi vender os nu i stedet for mod at teste om der er en signifikant effekt af `block`. Vi modellerer på datasættet inkl. outlieren. Med notation fra ovenfra vil vi altså teste hypotesen

$$H_2 : \alpha(1) = \alpha(2) = \alpha(3) = \alpha(4)$$

Vi fitter den nye model uden `block` i R og tester den mod den fulde model:

```
model_C <- lm(y ~ treat + x, data)
anova(model_C, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ treat + x
## Model 2: y ~ block + x + treat
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     11 1387.2
## 2      8 1216.7  3    170.52 0.3737 0.7744
```

Vi finder $p = 0.7744$ og accepterer derfor hypotesen. Vi konkluderer derfor, at der ikke er signifikant effekt af `block` og går videre med den reducerede `model_C`. Informationen om at eksperimentet var udført som et blokdesign gav altså ikke yderligere forklaringskraft.

Vi bliver til sidst bedt om at antage, at dyrenes vægte inden forsøget ikke er kendte, dvs. at vi ikke har observeret `x`. Vi ønsker igen at undersøge, om der er effekt af hhv. `treat` og `block` i denne nye model. Vi fitter modeller uden `x` i R og gennemfører tests svarende til dem fra ovenfor:

```
model_D <- lm(y ~ treat + block, data) #"fuld" model
model_E <- lm(y ~ block, data) #model uden treat
model_F <- lm(y ~ treat, data) #model uden block
model_G <- lm(y ~ 1, data) #tom model

anova(model_E, model_D) #Test: Effekt af treat?
```

```
## Analysis of Variance Table
##
## Model 1: y ~ block
## Model 2: y ~ treat + block
##   Res.Df  RSS Df Sum of Sq      F Pr(>F)
## 1     12 7100
## 2      9 4225  3      2875 2.0414 0.1786
```

```
anova(model_F, model_D) #Test: Effekt af block?
```

```
## Analysis of Variance Table
##
## Model 1: y ~ treat
## Model 2: y ~ treat + block
##   Res.Df  RSS Df Sum of Sq      F  Pr(>F)
## 1     12 9900
## 2      9 4225  3      5675 4.0296 0.04517 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi ser, at der er en signifikant effekt af `block` i den additive model ($p = 0.0452$), men ingen signifikant effekt af `treat` ($p = 0.1786$). Altså reducerer vi til den simplere model `model_E` og tester om der i denne model er en effekt af `block`:

```r
anova(model_G, model_E)
```

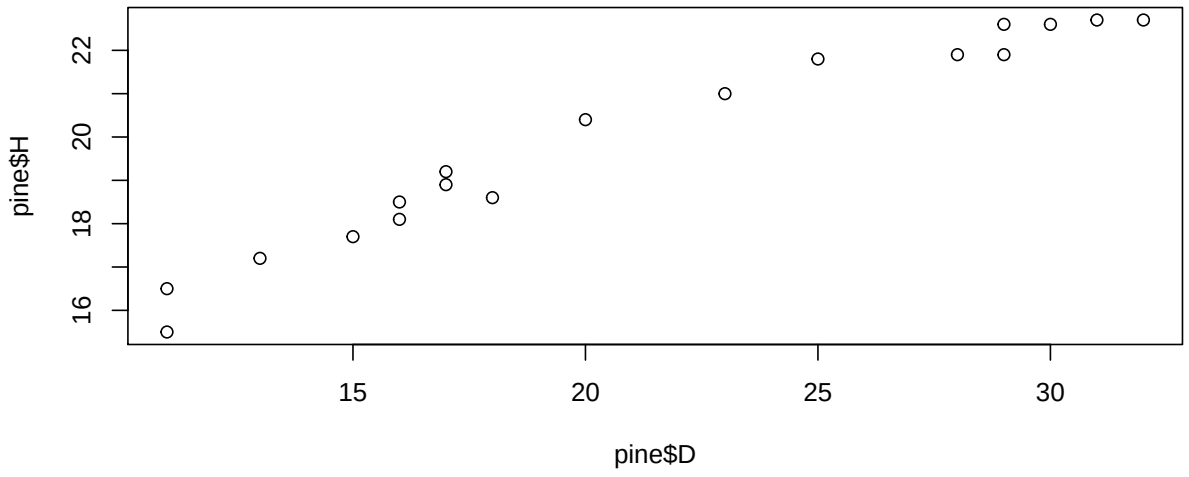
```
## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ block
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     15 12775
## 2     12  7100  3      5675 3.1972 0.06241 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

og vi finder $p = 0.0624$ og konkluderer derfor, at der heller ikke er en signifikant effekt af `block`. Bemærk dog at p-værdien er tæt på 0.05, så her er det en vurderingssag, om man vil udelade `block` og gå videre med `model_G` eller ej. Uanset hvad, så ser vi, at vi når en fuldstændigt anderledes konklusion, når ikke vi tager højde for studenes startvægt. Moralen er altså, at det kan gøre en enormt stor forskel, om man får opstillet en model med passende forklarende variable fra start af.

Opgave 3.B

Vi starter med at loade data og betragte det:

```r
pine <- read.table("pine.txt", header=T)
plot(pine$D, pine$H)
```



I kompendiet foreslås det at betragte følgende sammenhæng mellem højde (H) og diameter (D):

$$H = \alpha \cdot D^{\beta}$$

Vi vil gerne opstille en lineær, statistisk model ud fra denne sammenhæng, men da sammenhængen ikke er lineær, bliver vi nødt til at omskrive den. Ved at anvende den naturlige logaritme på begge sider af lighedstegnet fås

$$\begin{aligned}\log(H) &= \log(\alpha \cdot D^{\beta}) \\ &= \log(\alpha) + \log(D^{\beta}) \\ &= \log(\alpha) + \beta \cdot \log(D) \\ &= a + \beta \cdot \log(D)\end{aligned}$$

hvor $a = log(\alpha)$. Vi har nu en sammenhæng, som er lineær i parametrene og kan opskrive den tilsvarende statistiske model:

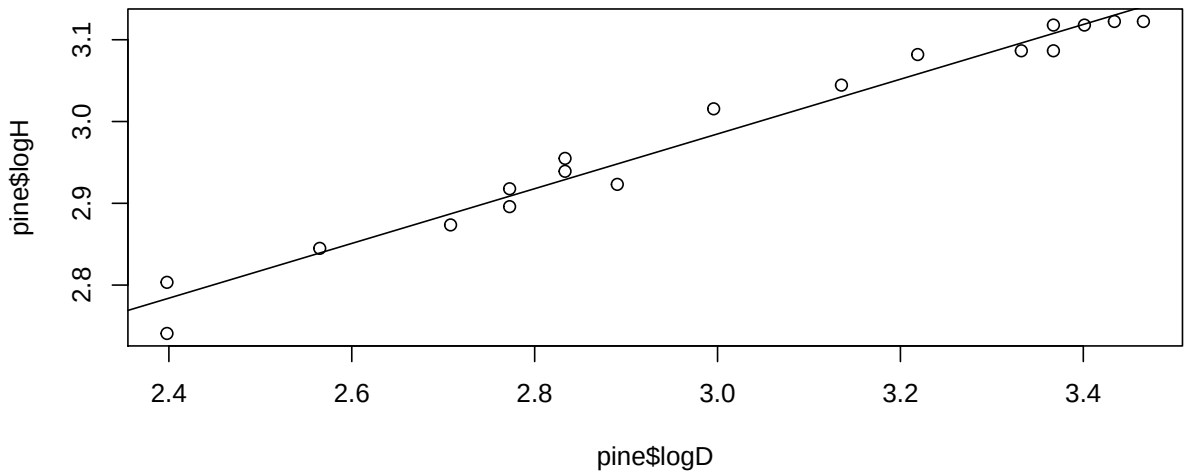
$$\log(H_i) = a + \beta \cdot \log(D_i) + e_i$$

for $i = 1...18$ og hvor det antages at e_i'erne er iid med $e_1 \sim N(0, \sigma^2)$. Vi vil gerne fitte denne model i R og derfor gemmer vi to nye variable i datasættet, nemlig `log(D)` og `log(H)`, og fitter modellen:

```
pine$logD <- log(pine$D)
pine$logH <- log(pine$H)
model1 <- lm(logH ~ logD, pine)
```

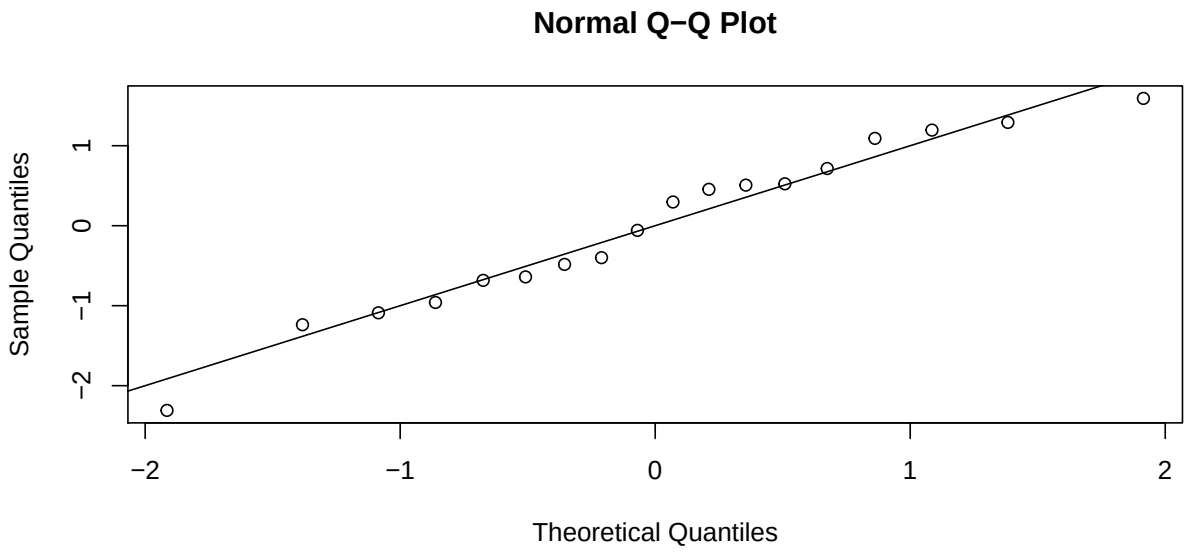
Vi vil gerne kontrollere modellens antagelser. Vi starter med at undersøge, om middelværdistrukturen er rimelig, dvs. om der er rimeligt at antage en lineær sammenhæng mellem `logD` og `logH`:

```
plot(pine$logD, pine$logH)
abline(model1) #tilføjer regressionslinjen fra model1
```



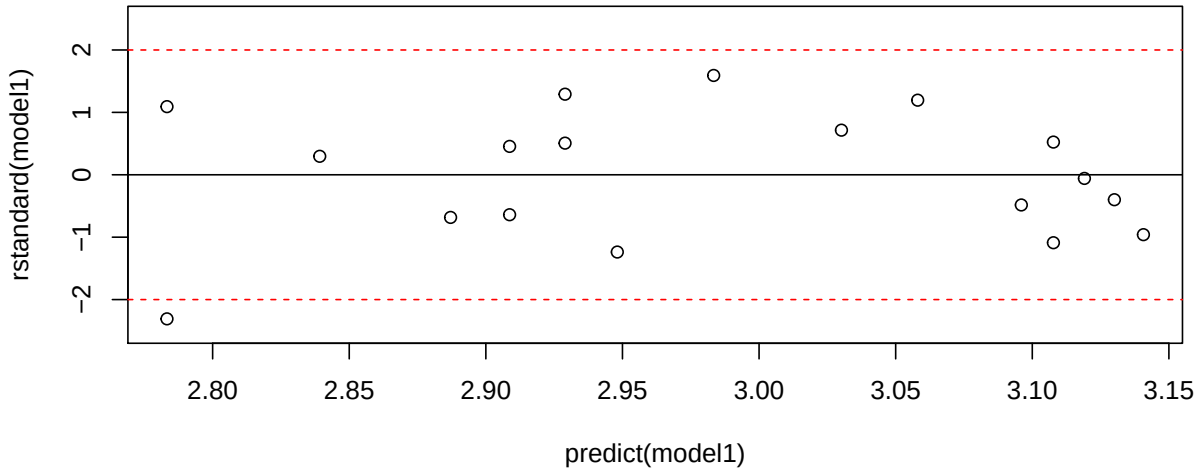
Det ser ud til at punkterne ligger pænt omkring den rette linje. Dette tyder på at antagelsen om at der er en lineær sammenhæng mellem `logD` og `logH` er rimelig. Vi undersøger nu om antagelsen om normalfordeling af residualerne er fornuftig:

```
qqnorm(rstandard(model1))
abline(0,1)
```



Vi ser, at punkterne ligger pænt om den rette linje, hvilket tyder på at normalfordelingsantagelsen også er rimelig. Vi undersøger antagelsen om varianshomogenitet:

```
plot(predict(model1),rstandard(model1), ylim=c(-2.5, 2.5))
      abline(0,0)
      abline(-2,0, col="red", lty="dashed")
      abline(2,0, col="red", lty="dashed")
```



Vi ser ingen systematiske placeringerne af punkterne langs x-aksen. Dette tyder på at det er rimeligt at antage varianshomogenitet. Bemærk dog, at vi arbejder med meget få observationer, så det er svært at konkludere noget endegyldigt.

Vi vil nu teste om sammenhængen mellem `logD` og `logH` er lineær, eller om vi fx. kan vise at den er kvadratisk. Vi opstiller altså en kvadratisk model og tester de to modeller mod hinanden:

```
model2 <- lm(logH ~ logD + I(logD^2), pine)
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: logH ~ logD
## Model 2: logH ~ logD + I(logD^2)
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     16 0.0069160
## 2     15 0.0058516  1 0.0010644 2.7286 0.1193
```

Vi finder at $p = 0.1193 > 0.05$, og altså skal vi reducere til den lineære model. Der er altså ikke nogen signifikant kvadratisk effekt af `logD`. Vi kan også teste, om der egentlig er en sammenhæng mellem de to variable ved at teste `model1` mod en model, som ingen effekt af `logD` har:

```
anova(model1, lm(logH ~ 1, pine))
```

```
## Analysis of Variance Table
##
## Model 1: logH ~ logD
## Model 2: logH ~ 1
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     16 0.006916
## 2     17 0.251400 -1  -0.24448 565.61 6.522e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

og vi finder her en meget lav p-værdi ($p = 6.5 \cdot 10^{-14}$), og altså konkluderer vi, at der helt bestemt er en signfikant sammenhæng mellem `logD` og `logH`.

Vi vil nu teste hypotesen

$$H : E(H_i) = \alpha \cdot D_i$$

som svarer til

$$E(log(H_i)) = \alpha + \log(D_i)$$

dvs. hypotesen

$$H : \beta = 1$$

Vi kan teste denne hypotese vha. en t-test eller (fuldstændigt ækvivalent) ved at undersøge, om konfidensintervallet for $\hat{\beta}$ indeholder værdien 1. Vi bestemmer dette konfidensinterval:

```
confint(model1)
```

```
##                 2.5 %   97.5 %
## (Intercept) 1.8908342 2.070666
## logD        0.3048514 0.364517
```

og vi ser at 1 ikke er indeholdt i konfidensintervallet. Altså afviser vi H og konkluderer, at β er signifikant forskellig fra 1. Det betyder altså, at undersøgelsen viser at alle træer ikke har samme "facon", uanset deres diameter. Bemærk dog, at vi ikke ved ovenstående metode får nogen p-værdi - og det kan man sommetider være interesseret i, så man har et mål for hvor sikre vi er på at hypotesen skal afvises. Hvis vi omskriver den oprindelige model, kan vi gøre det let at teste hypotesen direkte. Bemærk, at R gerne tester hypoteser om at

parametre er lig 0. Vi skal altså have skrevet modellen, så H svarer til at en parameter er lig 0. Bemærk at hvis vi opstiller en ny model

$$\log(H_i) - \log(D_i) = a + b \cdot \log(D_i) + e_i$$

kan vi ved omskrivning opnå

$$\log(H_i) = a + \log(D_i) \cdot (1 + b) + e_i$$

og hvis vi tester den sædvanlige hypotese, $b = 0$, i denne model, får vi p-værdien til testen for $\beta = 1$ i `model1`. Vi fitter modellen og betragter `summary()`-outputtet for at aflæse p-værdien:

```
model3 <- lm(logH - logD ~ logD, pine)
summary(model3)
```

```
##
## Call:
## lm(formula = logH - logD ~ logD, data = pine)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.042448 -0.013340  0.002292  0.013334  0.032160
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.98075    0.04242   46.70   <2e-16 ***
## logD        -0.66532    0.01407  -47.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02079 on 16 degrees of freedom
## Multiple R-squared:  0.9929, Adjusted R-squared:  0.9924
## F-statistic:  2235 on 1 and 16 DF,  p-value: < 2.2e-16
```

Vi ser at b er signifikant forskellig fra 0 ($p < 2 \cdot 10^{-16}$) og dermed at β er signifikant forskellig fra 1.