Opgave 2

2024-10-07

initialisering

```
# install.packages("ggplot2")
library("ggplot2") # Plots
# install.packages("isdals")
library(isdals) # Data
```

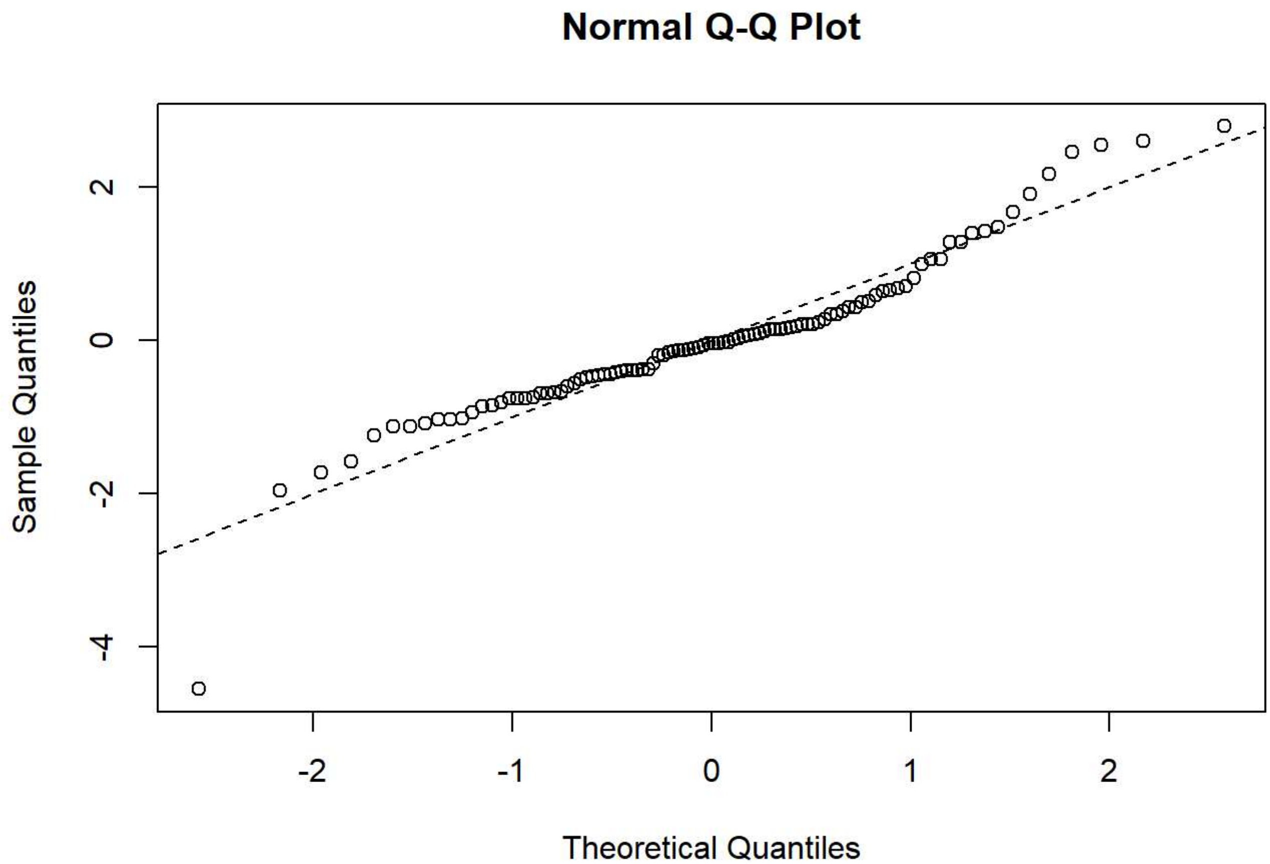
```
florida <- read.table("../data/florida.txt", header = 1)
florida <- data.frame(florida) # Convert to dataframe
```

Eksamen, Januar 2019 - Opgave 2

Opgave 1

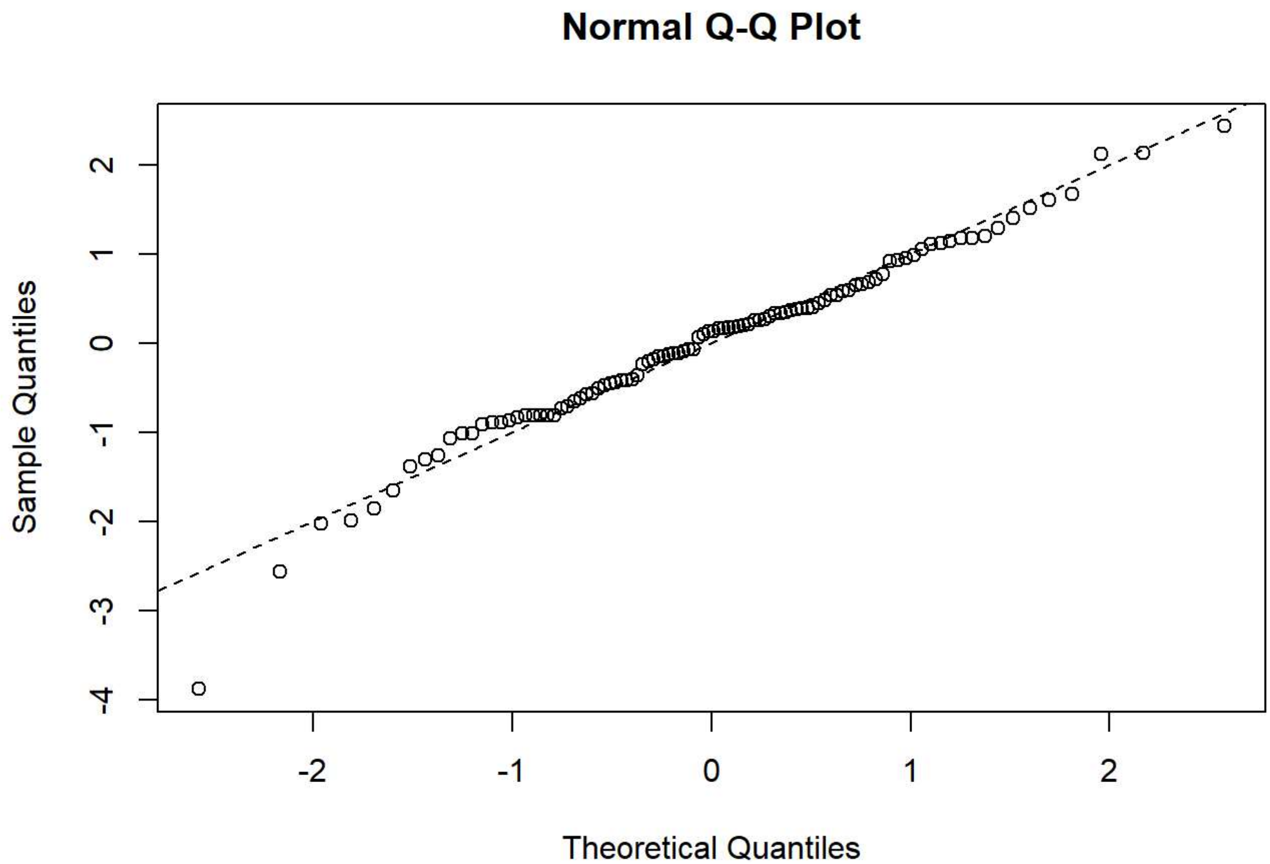
Det er mest oplagt at anvede pris som responsvariabel, da det ikke er en målbar størrelse før huset er solgt. Og vi anvender size som forklarende variabel da den er målbare. Det ville mest brugbart at udvikle en model, som kan give et estimat for prisen givet størrelsen på huset.

```
linreg1 <- lm(Price ~ Size, data=florida)

qqnorm(rstandard(linreg1))
abline(0,1, lty=2)
```



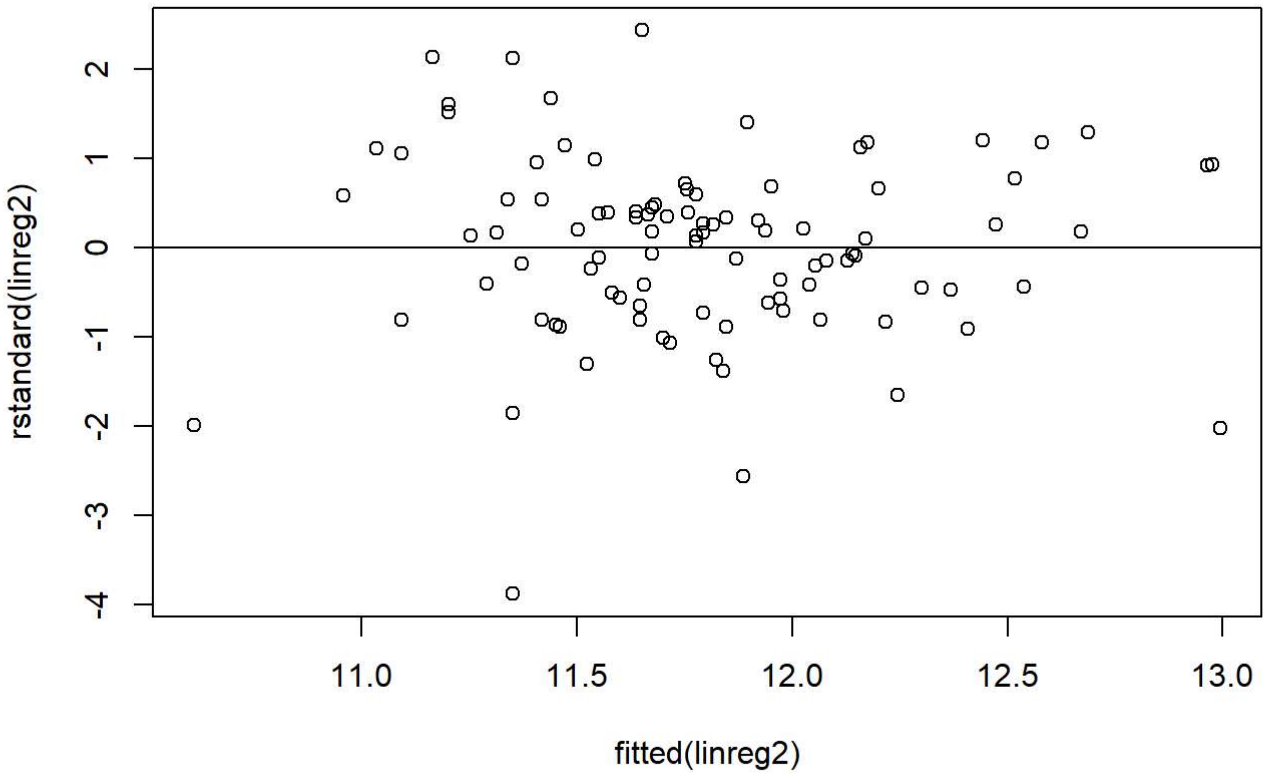
Vi ser en systematisk avigelse fra i QQ-plottet, for de laveste og højeste kvartiler ses in konkav krumning omkring den ligefrem propotionelle linje og i midten ser vi en konveks krumning.

```
linreg2 <- lm(log(Price) ~ log(Size), data=florida)


qqnorm(rstandard(linreg2))
abline(0,1, lty=2)
```



Efter logtransformationen er der ikke nogen systematisk afvigelse fra linjen i QQ-plottet, hvilket betyder at data ligger i de kvartiler, som vi forventer ved en normalfordeling.

```
plot(fitted(linreg2), rstandard(linreg2))
abline(a = 0, b = 0)
```



Residualerne ligger fordelt omkring linjen, som vi forventer det, med ca. 95% af datapunkterne indenfor 2 og -2 på y-aksen og ca. samme mængde punkter over og under linjen.

Opgave 2

Data:
Par $(x_1, y_1), \ldots, (x_n, y_n)$ bestående af kvantitative kontinuerte data, både for vores responsvariabel log(price) og den forklarende variabel log(Size).

Statistisk model:

$$y_i = \mu_i + \epsilon$$

hvor $\epsilon_i \sim N(0, \sigma^2)$

Vi antager $y_1, \ldots, y_n$ er uafhængihed, og at alle y_i er normaltfordelt med middelværdi $\mu_i = \alpha + \beta x_i$ (ret linje) med spredning σ, som er vores ukendte populationsparametre

De ukendte populationsparametre
Skæringen α, hældningen β og spredningen σ er alle ukendte parametre, som vi kan estimere.

Vi vedhæfter os at når vi projektetere μ tilbage til den ikke logtransformeret data, så er det ikke en middelværdi, men en median.

```
summary(linreg2)
```

```
##
## Call:
## lm(formula = log(Price) ~ log(Size), data = florida)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33000 -0.21601  0.04722  0.19906  0.83721
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.81449    0.70414   3.997 0.000124 ***
## log(Size)    1.22549    0.09599  12.767  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3461 on 98 degrees of freedom
## Multiple R-squared:  0.6245, Adjusted R-squared:  0.6207
## F-statistic:   163 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
confint(linreg2)
```

```
##                2.5 %   97.5 %
## (Intercept) 1.417138 4.211843
## log(Size)   1.035005 1.415971
```

Estimatet for hældningsparameterener $\hat{\beta} = 1.23$. \ Vi får et konfidensinterval på [1.035, 1.42]

Opgave 3

I stedet for at predict de to værdier og trække dem fra hindande kan vi nøjes med at kigge på hældningsparameteren gange differencen

$$\hat{\beta} \cdot (log(2000) - log(1000))$$

```
1.22549 * (log(2000) - log(1000))
```

```
## [1] 0.8494449
```

Vi skal dog kende de to e^{μ} værdien for at sige noget om faktoren

```
small <- exp(2.81449 + 1.22549 * log(1000))
big <- exp(2.81449 + 1.22549 * log(2000))


big
```

```
## [1] 185223.8
```

```
small
```

```
## [1] 79211.36
```

```
cat("\nFaktor: \n")
```

```
##
## Faktor:
```

```
big / small
```

```
## [1] 2.338349
```

Hvis jeg har forstået opgaven rigtig så er det store hus lidt mere end dobbelt så dyrt, hvilket giver mening da vores hældningsparameter er større end en. Hvis den havde været 1 med en skæring i 0, ville faktoren være præcis 2.

Opgave 4

```
# R log transformere automatisk
pred <- predict(linreg2, newdata = data.frame(Size = 3000), interval = "p")
pred
```

```
##       fit      lwr      upr
## 1 12.6262 11.92397 13.32842
```

```
exp(pred[1])
```

```
## [1] 304429.8
```

```
exp(pred[2])
```

```
## [1] 150839
```

```
exp(pred[3])
```

```
## [1] 614413.3
```

Da 215.000 ligger i prediktionsintervallet fra 150.839 til 614.413, anses det, udfra data, ikke som en usædvanlig pris

Opgave 5

```
multipel <- lm(log(Price) ~ log(Size) + Baths, data=florida)
```

$$\texttt{Log(Price)} = \alpha + \beta_1 \cdot \texttt{log(Size)} + \beta_2 \cdot \texttt{Baths}$$

$$H_0 : \beta_2 = 0$$

- *fuld model*: Multipel lineær regressions model med 2 forklarende variable log(Size) og Baths
- *nulmodel*: En undermodel hvor med 1 forklarende variabel log(Size)

```
fullModel <- lm(log(Price) ~ log(Size) + Baths, data=florida)
nulModel <- lm(log(Price) ~ log(Size), data=florida)

anova(nulModel, fullModel)
```

```
## Analysis of Variance Table
##
## Model 1: log(Price) ~ log(Size)
## Model 2: log(Price) ~ log(Size) + Baths
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     98 11.737
## 2     97 11.645  1  0.092386 0.7696 0.3825
```

Vi forkaster nulhyposten om at Baths ikke har nogen signifikat effekt på et signifikatnivuea på 95% udfra en p-værdi på 0.3825