MODELS IN R

Tobias Niedermaier

LINEAR REGRESSION

LINEAR MODEL

$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i + \ldots + \epsilon_i$,with $\epsilon \sim N(0, \sigma^2)$

R command: `lm(formula, data, ...)`

Examples of `formula`:

- `y~age+sex`
- `y~I(log(age)) + as.factor(sex)`
- see `?formula` for more details

Note: `data` must be a data.frame.

Example:

```
1  library(MASS)
2  data(cats)
3  with(cats, plot(Bwt, Hwt, xlab="Body Weight (kg)",
4          ylab="Heart Weight (g)",
5           main="Heart Weight vs. Body Weight of Cats"))
```

```
1  with(cats, cor.test(Bwt, Hwt))
```

```
    Pearson's product-moment correlation

data:  Bwt and Hwt
t = 16.119, df = 142, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7375682 0.8552122
sample estimates:
      cor
0.8041274
```

```r
lm(Hwt ~ Bwt,data=cats)
```

```
Call:
lm(formula = Hwt ~ Bwt, data = cats)

Coefficients:
(Intercept)          Bwt
    -0.3567       4.0341
```

FUNCTION SUMMARY()

- The output of `summary()` provides an overview on the model:

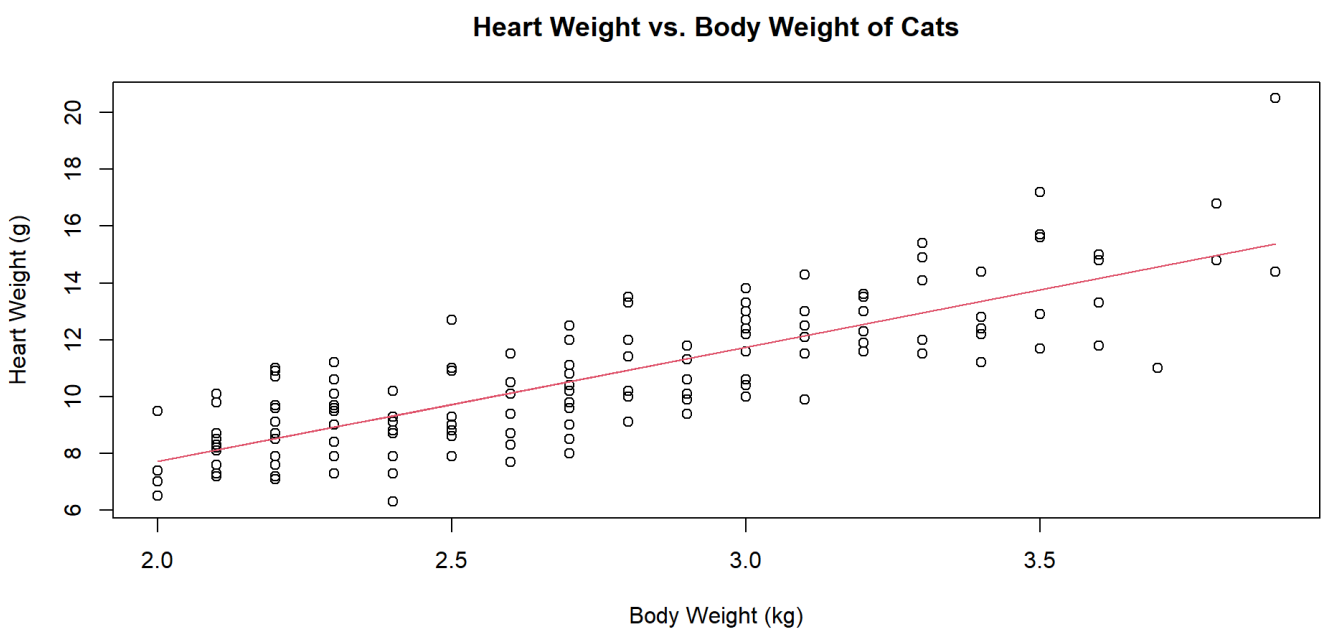
```
1  model <- lm(Hwt ~ Bwt,data=cats)
2  summary(model)
```

```
Call:
lm(formula = Hwt ~ Bwt, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5694 -0.9634 -0.0921  1.0426  5.1238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3567     0.6923  -0.515    0.607
Bwt           4.0341     0.2503  16.119   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.452 on 142 degrees of freedom
Multiple R-squared:  0.6466,    Adjusted R-squared:  0.6441
```

PLOT REGRESSION

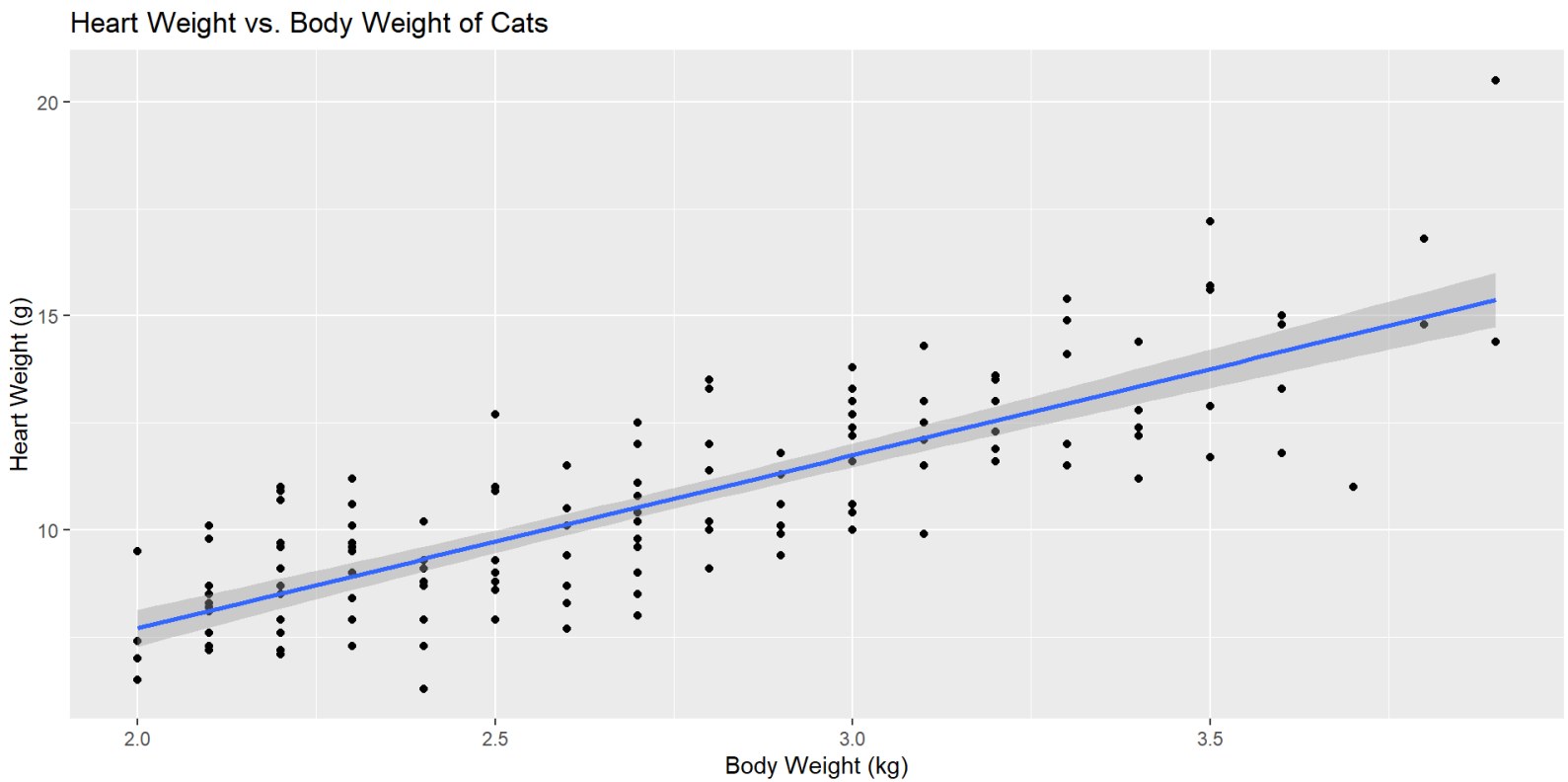
- Plot the regression line

```r
plot(cats$Bwt, cats$Hwt, xlab="Body Weight (kg)",
        ylab="Heart Weight (g)",
         main="Heart Weight vs. Body Weight of Cats")
yhat <- coef(model)[1] + coef(model)[2]*cats$Bwt

lines(x=cats$Bwt,y=yhat,col=2)
```



REGRESSION LINE IN GGPLOT

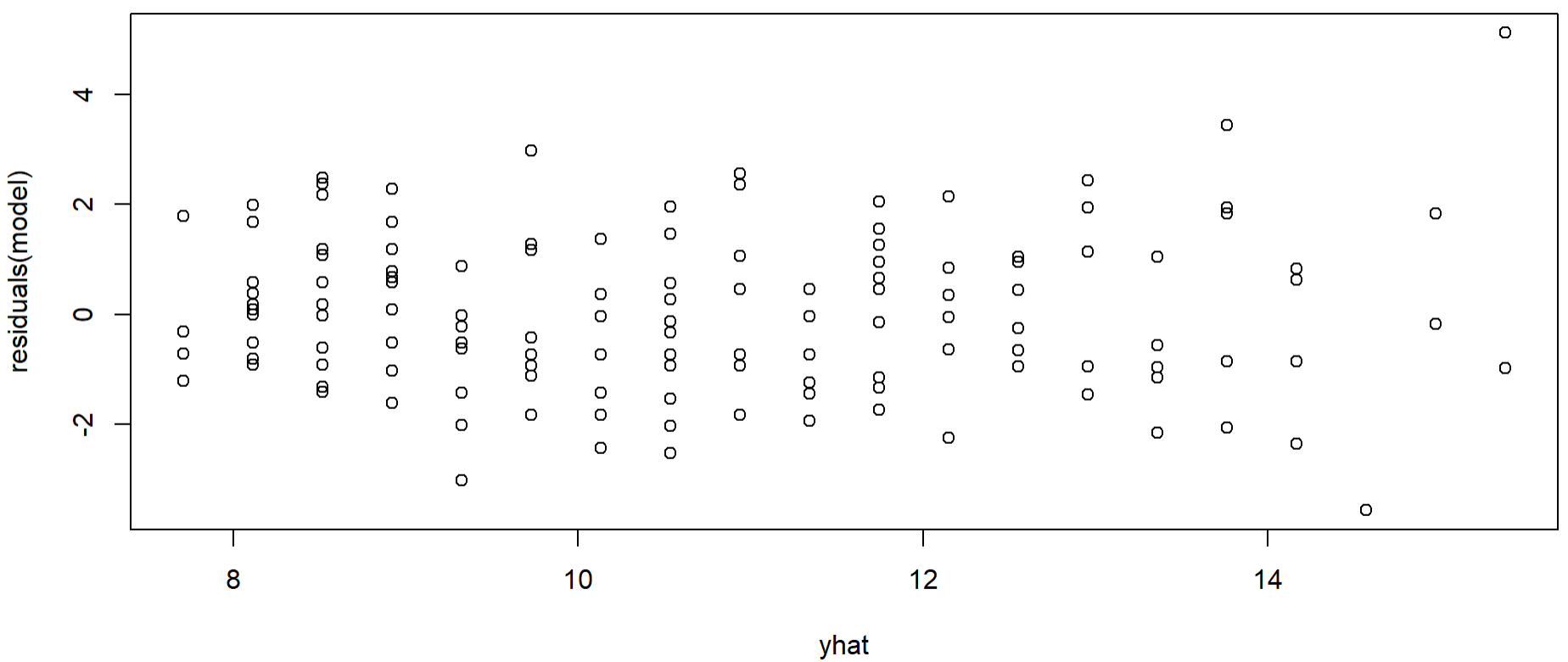
```
1  library(ggplot2)
2  ggplot(cats, aes(Bwt, Hwt))+
3    geom_point()+
4    geom_smooth(method="lm")+
5    xlab("Body Weight (kg)")+
6    ylab("Heart Weight (g)")+
7    ggtitle("Heart Weight vs. Body Weight of Cats")
```



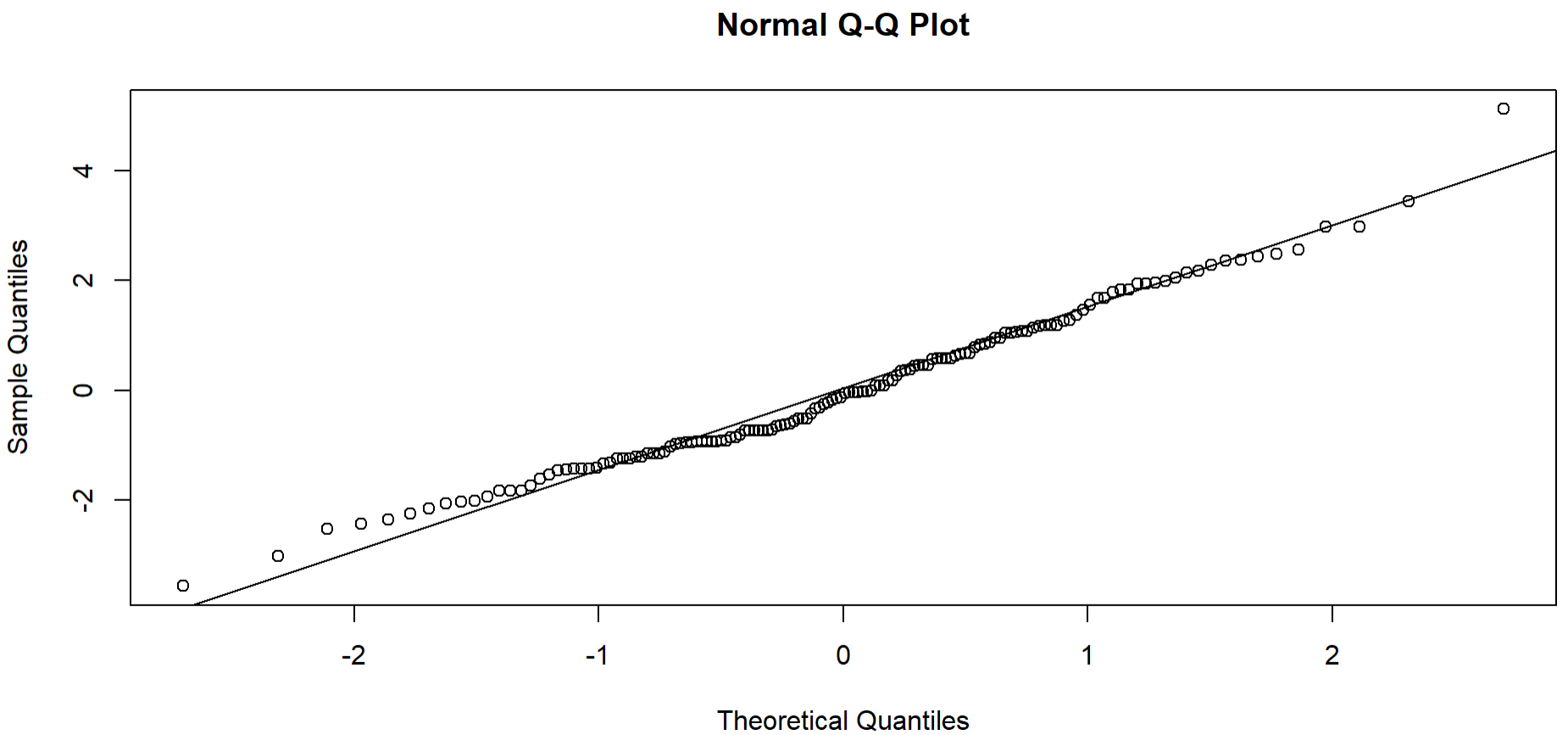
GRAPHICAL DIAGNOSIS

Check residuals directly on plots or use the function `plot()` function on regression object for diagnosis graphics:

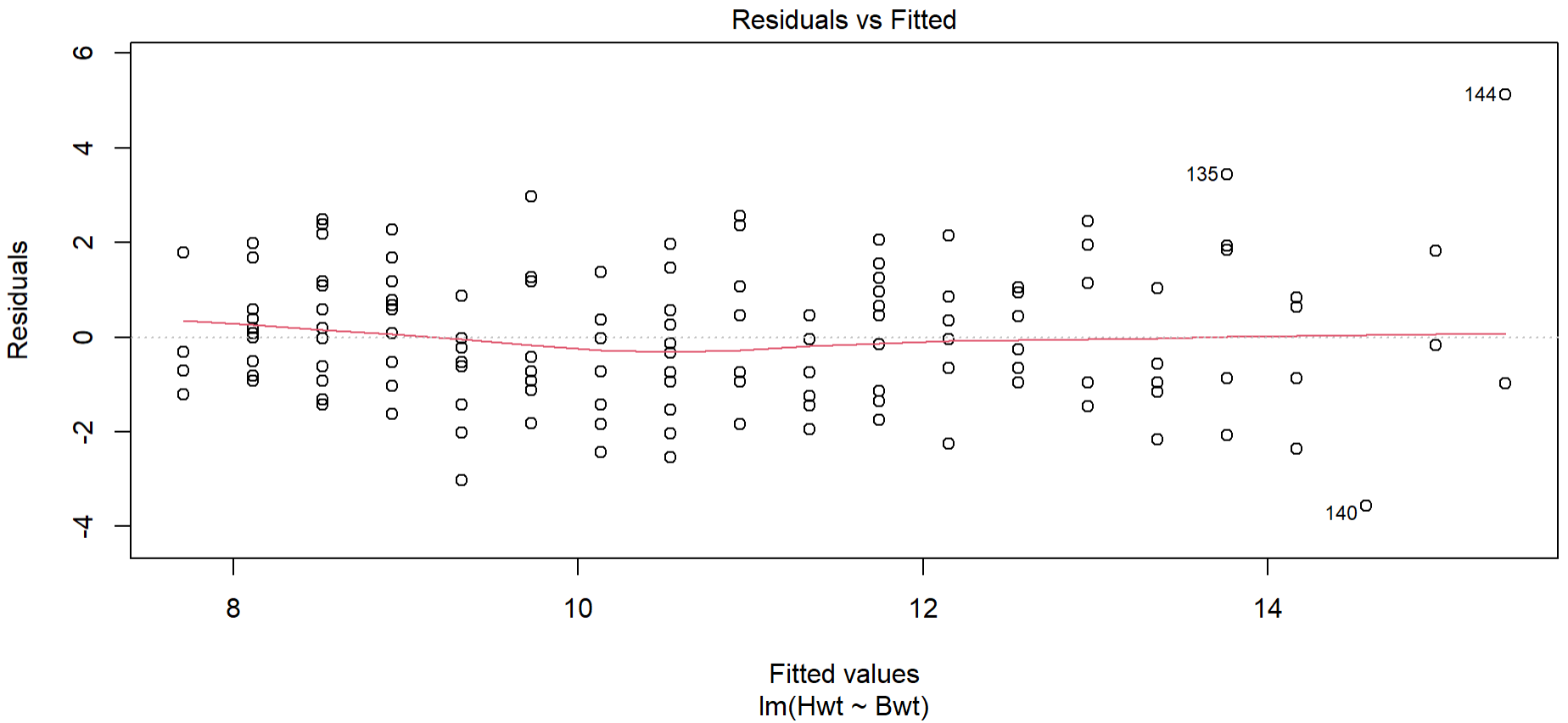
```
1  plot(yhat,residuals(model))
```

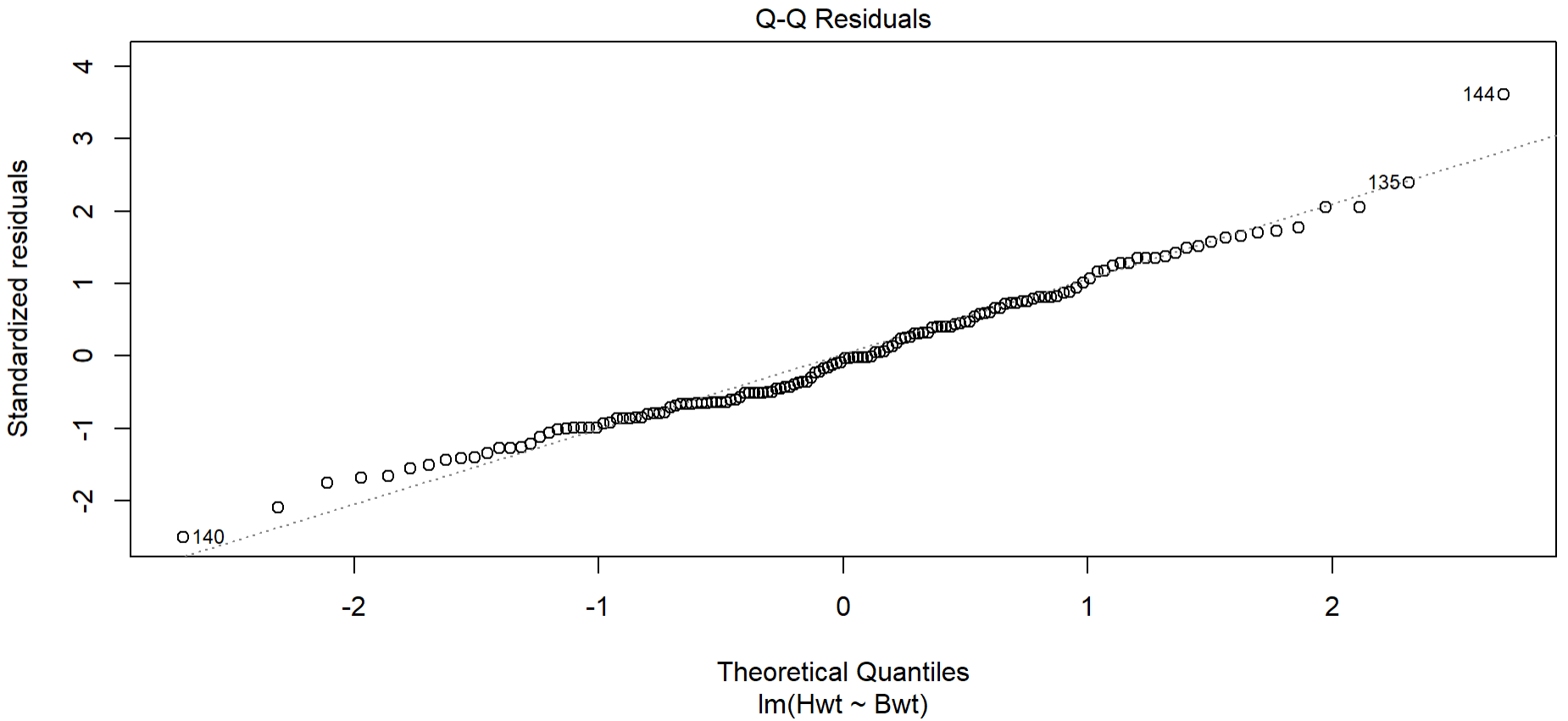


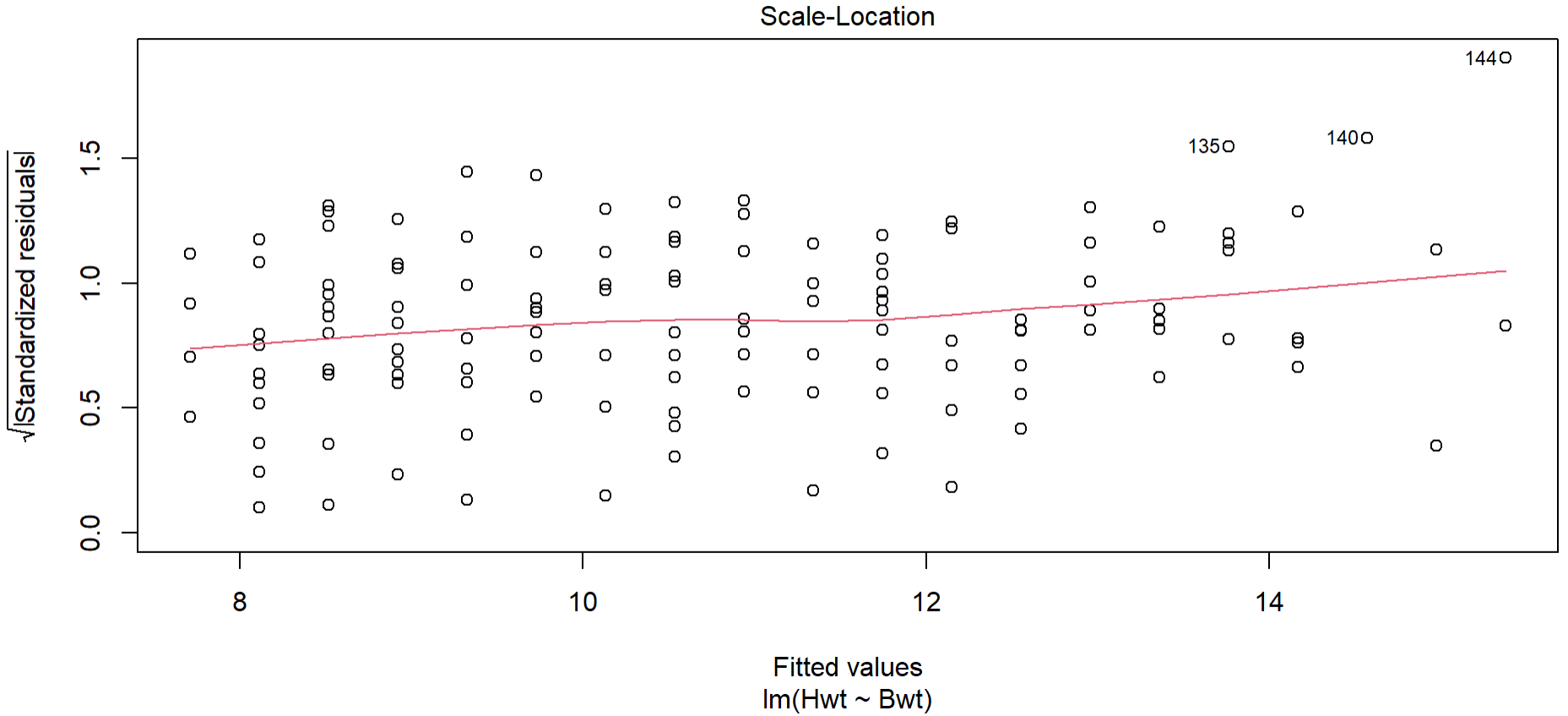
```r
qqnorm(residuals(model))
qqline(residuals(model))
```

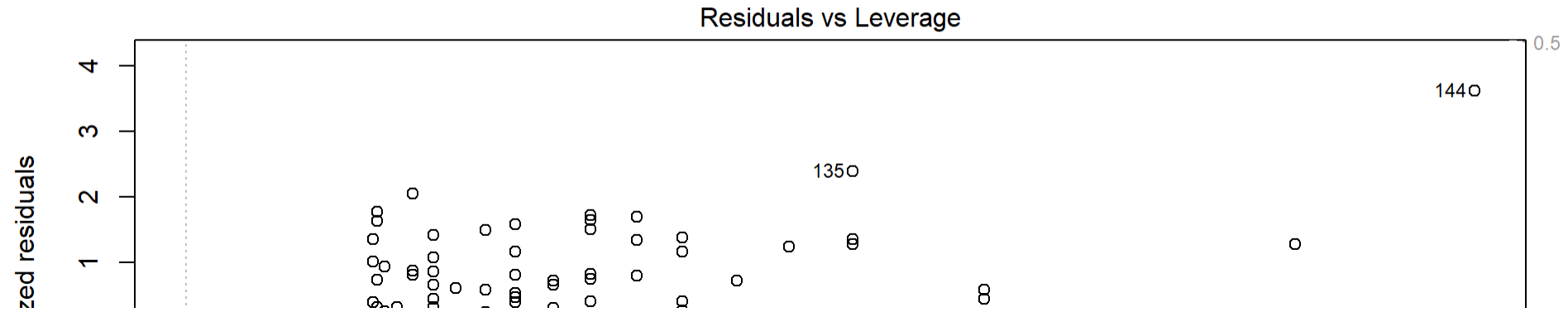



```r
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Hwt ~ Bwt)
144
135
140
6
4
2
0
-2
-4
8
10
12
14


Q-Q Residuals
Standardized residuals
4
3
2
1
0
-1
-2
144
135
140
-2
-1
0
1
2
Theoretical Quantiles
lm(Hwt ~ Bwt)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
8
10
12
14
144
135
140
Fitted values
lm(Hwt ~ Bwt)


Residuals vs Leverage
zed residuals
1
2
3
4
0.5
144
135

MODEL SELECTION

- In R there are pre-built functions which perform automatic model selection, for example, `stepAIC()` in the package MASS

```
stepAIC(object, scope, scale = 0, direction = c("both", "backward", "forward"), trace = 1, keep = N
```

- By default, it uses AIC as stopping criterion (argument `k=2`)
- setting `k=log(n)` we can use BIC. Example:

```
library(MASS)
model <- lm(Hwt ~ Bwt+Sex,data=cats)
model.new<-stepAIC(model)
```

```
Start:  AIC=111.39
Hwt ~ Bwt + Sex

       Df Sum of Sq    RSS    AIC
- Sex   1      0.15 299.53 109.47
<none>              299.38 111.39
- Bwt   1    405.88 705.26 232.78

Step:  AIC=109.47
Hwt ~ Bwt
```

```
       Df Sum of Sq    RSS    AIC
<none>              299.53 109.47
- Bwt   1    548.09 847.63 257.26
```

PREDICTION

- The function `predict()` computes the predicted value of the response for a new observation.

```
predict(object, newdata, se.fit = FALSE, interval = c("none", "confidence", "prediction"), level =
```

In particular:

- `object` must be replaced by the model fit on the data
- `newdata` must be a data.frame with the new observation(s)
- `se.fit` allows the computation of the standard error
- setting `interval="prediction"`, we can compute the prediction interval, with `level` (default `level = 0.95`);

Example:

```
model <- lm(Hwt ~ Bwt,data=cats)
y.hat <- predict(model, newdata = cats, interval="prediction")
```

EXERCISES 5 TASK 1

LOGISTIC REGRESSION

LOGISTIC MODEL

$$logit(\pi) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots, \text{with } \pi = Pr(Y = 1|X)$$

- is a special case of generalized linear model

```
glm(formula, family = gaussian, data, ...)
```

- to fit a logistic model, the argument `family` must be set equal to binomial.

Example:

```
load(file="myNhanes.RData")
glm(cancer_ever~workpollut,family=binomial,data=mydata)
```

```
Call:  glm(formula = cancer_ever ~ workpollut, family = binomial, data = mydata)

Coefficients:
   (Intercept)  workpollutTRUE
       -2.6074          0.1048

Degrees of Freedom: 4192 Total (i.e. Null);  4191 Residual
  (807 Beobachtungen als fehlend gelöscht)
Null Deviance:      2176
```

LIKE FOR THE LINEAR MODEL...

- we can have an overview on the model through the function `summary()`:

```
1  summary(glm(cancer_ever~workpollut,family=binomial, data=mydata))
```

```
Call:
glm(formula = cancer_ever ~ workpollut, family = binomial, data = mydata)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.60744    0.08758 -29.774   <2e-16 ***
workpollutTRUE  0.10480    0.11961   0.876    0.381
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2175.8  on 4192  degrees of freedom
Residual deviance: 2175.0  on 4191  degrees of freedom
  (807 Beobachtungen als fehlend gelöscht)
AIC: 2179
```

MODEL SELECTION

- we can perform model selection (e.g., based on AIC) through the function `stepAIC()`
- we can use the function predict() to predict

```
1  stepAIC(glm(mydata$cancer_ever~mydata$workpollut,family=binomial))
2
3  predict(glm(cancer_ever~workpollut,family=binomial, data=mydata),newdata=mydata)
```

EXERCISES 5 TASK 2