STAT 210
Applied Statistics and Data Analysis
Problem list 8
(Due on week 9)

```
library(car); library(caret)
library(alr4)
library(MASS)
```

Exercise 1

In this exercise we will use the data set `iris`.

(i) Extract the data corresponding to species `setosa` to a separate data frame. Plot the numerical variables for this set in a matrix of plots.

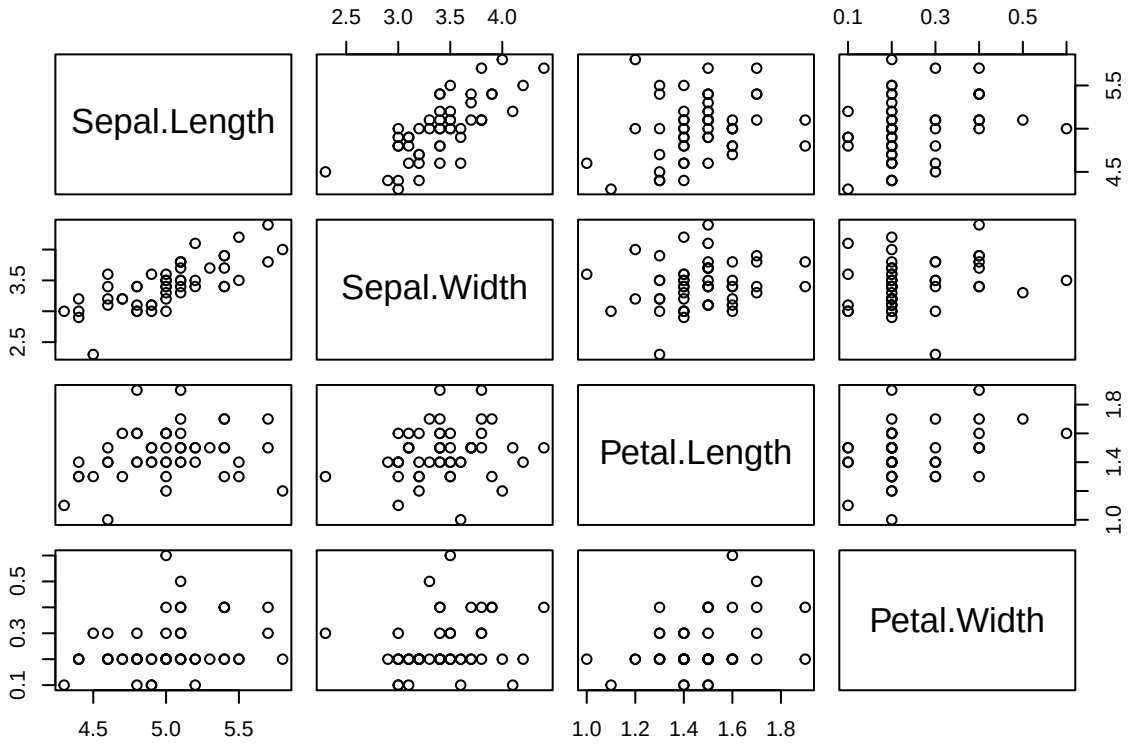
```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
iris.set <- subset(iris, Species == 'setosa')
str(iris.set)
```

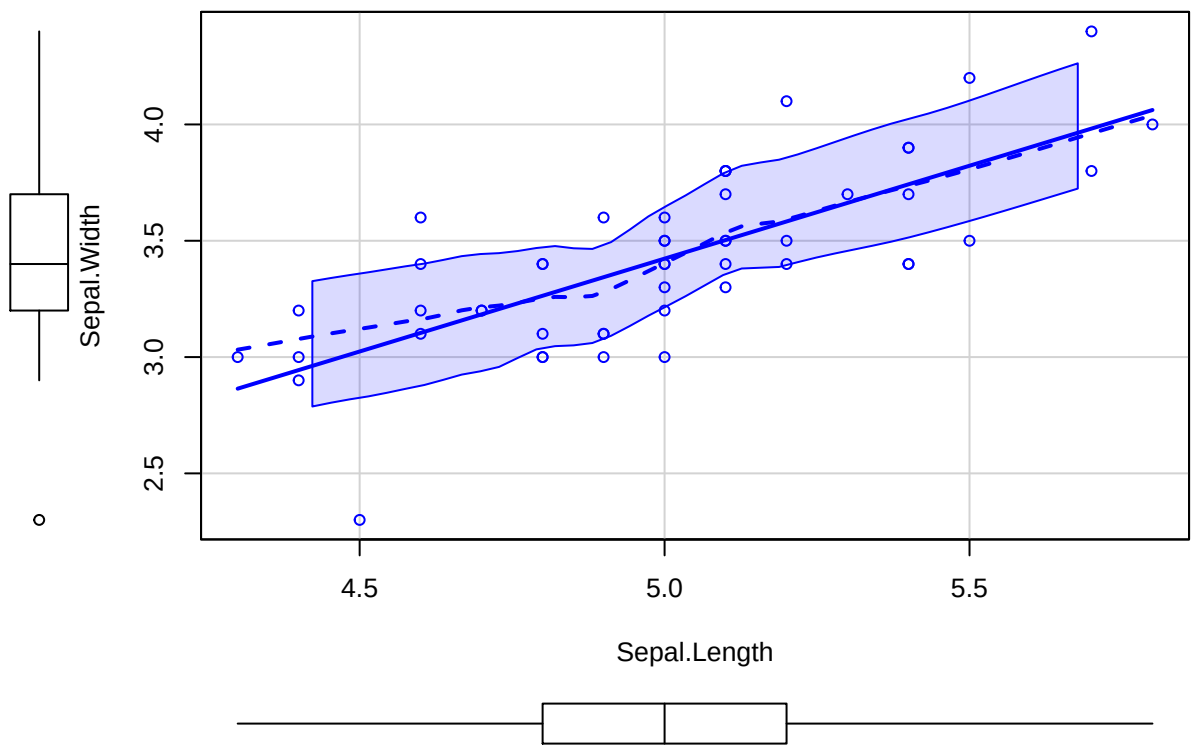
```
## 'data.frame':    50 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
plot(iris.set[,1:4])
```



(ii) Use the function `scatterplot` from the `car` package to plot `Sepal.Width` as a function of `Sepal.Length`. Comment on the graph.

```
scatterplot(Sepal.Width ~ Sepal.Length, data = iris.set)
```



This plot produces a local smoother curve (broken line) that can be compared with the regression line. Important discrepancies may indicate that the linear regression model may not be adequate. In this case the

agreement is good.

(iii) Fit a linear regression model for `Sepal.Width` as a function of `Sepal.Length`. Produce a table using `summary` and discuss the results.

```
modelA <- lm(Sepal.Width ~ Sepal.Length, data = iris.set)
summary(modelA)
```

```
##
## Call:
## lm(formula = Sepal.Width ~ Sepal.Length, data = iris.set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72394 -0.18273 -0.00306  0.15738  0.51709
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.5694     0.5217  -1.091    0.281
## Sepal.Length   0.7985     0.1040   7.681 6.71e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2565 on 48 degrees of freedom
## Multiple R-squared:  0.5514, Adjusted R-squared:  0.542
## F-statistic: 58.99 on 1 and 48 DF,  p-value: 6.71e-10
```

The slope estimator is significantly different from zero, accordint to the t-test, but the slope is not. This happens sometimes with physical models, since having one of the dimensions equal to zero implies that the other must also be equal to zero. Th eestimated standard deviation is 0.257 and the R^2 is 0.5514$.

(iv) Find the R^2 and verify that for simple linear regression, this coefficient is equal to the square of the correlation between the two variables.

The R^2 from the regression output is 0.5514.

```
with(iris.set, cor(Sepal.Width,Sepal.Length)^2)
```

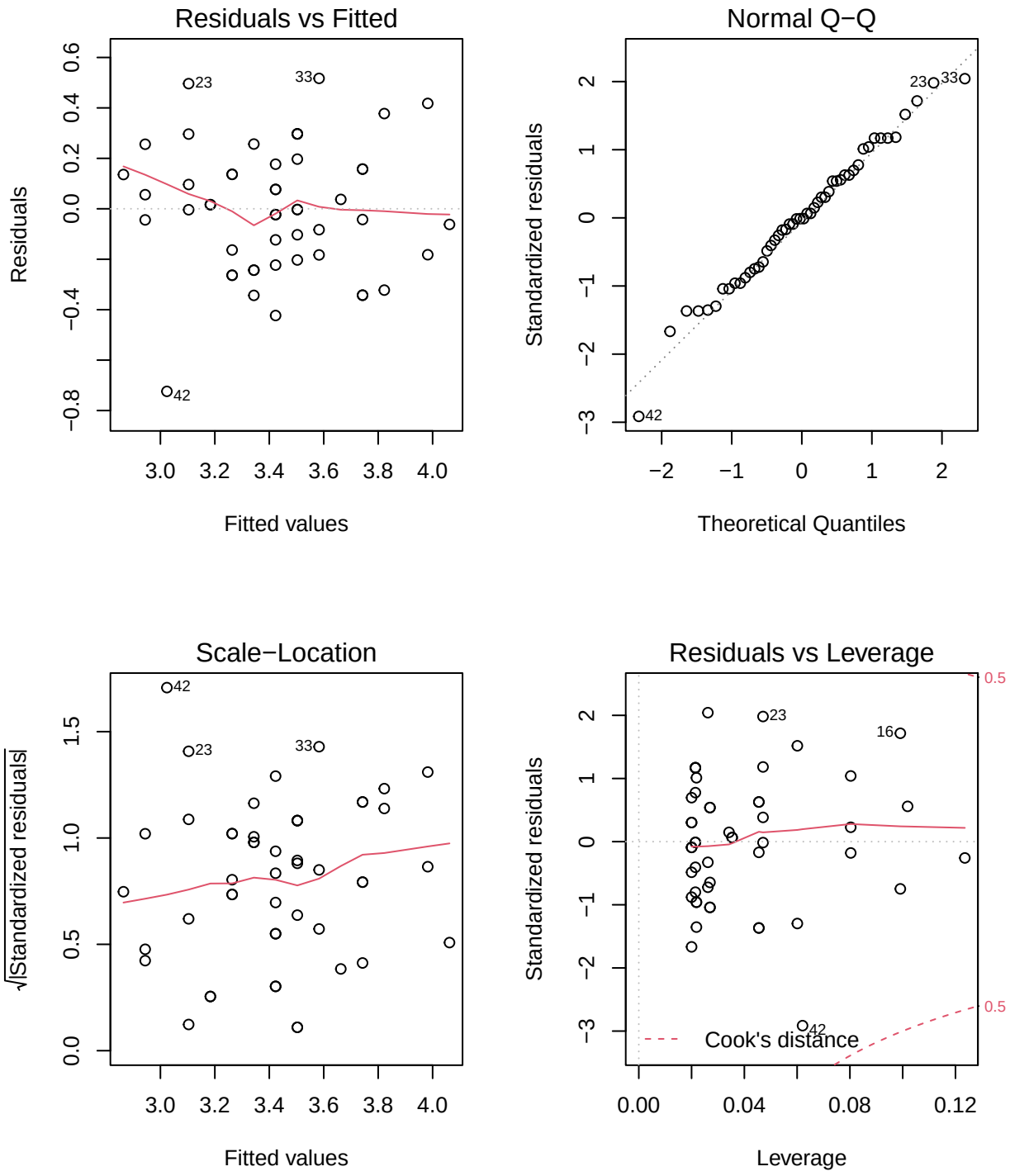
```
## [1] 0.5513756
```

(v) Write down the equation for the regression line and interpret the parameters.

$$\texttt{Sepal.Width} = -0.5694 + 0.7985 * \texttt{Sepal.Length}$$

(vi) Do the diagnostic plots for this model and comment.

```
par(mfrow=c(2,2))
plot(modelA)
```




```
par(mfrow=c(1,1))
```

In general, the plots look good. The quantile plot is partivularly good, so there are no doubts about normality. The only point that may raise cause for concern is the assumption of homoscedasticity, since the scale-location lot shows a small increasing tendency. We can check this with a test.

(vii) In this case, the diagnostic plots give sufficient information about the normality assumption. However, if we wanted to test this assumption, we could use the Shapiro-Wilk test. Do this test and comment on the result.

```
shapiro.test(rstandard(modelA))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(modelA)
## W = 0.98583, p-value = 0.8066
```

The p-value for this test is large, so we cannot reject the hypothesis of normality.

(viii) The assumption of uniform variance is not so clear from the plots, particularly from the Scale-Location graph. The test we used for analysis of variance does not work here, because we do not have grouped data. A test that can be used in this situation is the Score Test, proposed by Cook and Weisberg (1983) and described in Applied Linear Regression by S. Weisberg, Wiley. This test is available in the `car` package as `ncvTest`. Do this test and comment on the results.

```
ncvTest(modelA)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.07602347, Df = 1, p = 0.78276
```

The p-value for this test is large, so we cannot reject the hypothesis of homogeneous variance.

Exercise 2

For this exercise we will use the data set `birdsdiet.csv`, downloaded from https://wiki.qcbs.ca/r_workshop4. The set has the following variables

- `Family` Common name of family
- `MaxAbund` The highest observed abundance at any site in North America
- `AvgAbund` The average abundance across all sites where found in NA
- `Mass` The body size in grams
- `Diet` Type of food consumed. Categorical variable with five levels: `Plant; PlantInsect; Insect; InsectVert; Vertebrate`.
- `Passerine` Is it a songbird/ perching bird.
- `Aquatic` Is it a bird that primarily lives in/ on/ next to the water.

Load packages `MASS`, `car` and `alr4`.

(i) Load the dataset `birdsdiet.csv` and do an exploratory data analysis.

```
birds <- read.csv('birdsdiet.csv')
str(birds)
```

```
## 'data.frame':    54 obs. of  7 variables:
##  $ Family   : chr  "Hawks&Eagles&Kites" "Long-tailed tits" "Larks" "Kingfishers" ...
##  $ MaxAbund : num  2.99 37.8 241.4 4.4 4.53 ...
##  $ AvgAbund : num  0.674 4.04 23.105 0.595 2.963 ...
##  $ Mass     : num  716 5.3 35.8 119.4 315.5 ...
##  $ Diet     : chr  "Vertebrate" "Insect" "PlantInsect" "Vertebrate" ...
##  $ Passerine: int  0 1 1 0 0 0 0 0 0 0 ...
##  $ Aquatic  : int  0 0 0 0 1 1 1 0 1 1 ...
```

To get numerical descriptions of the data there many tools in R. Here we use `describe` and `describeBy` in the `psych` package, that give a very complete table of parameters.

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
describe(birds)
```

```
##           vars  n   mean     sd median trimmed   mad min     max   range  skew
## Family*      1 54  27.09  15.31  27.50   27.11 19.27 1.0   53.00   52.00 -0.02
## MaxAbund     2 54  44.91  73.47  24.15   27.95 28.50 0.2  413.60  413.40  3.17
## AvgAbund     3 54   5.69   8.25   3.11    3.87  2.97 0.2   47.60   47.40  3.24
## Mass         4 54 468.48 945.41  59.18  248.19 70.56 5.3 5296.23 5290.93  3.15
## Diet*        5 54   3.00   1.67   4.00    3.00  1.48 1.0    5.00    4.00 -0.19
## Passerine    6 54   0.46   0.50   0.00    0.45  0.00 0.0    1.00    1.00  0.14
## Aquatic      7 54   0.28   0.45   0.00    0.23  0.00 0.0    1.00    1.00  0.96
##           kurtosis     se
## Family*      -1.24   2.08
## MaxAbund     11.12  10.00
## AvgAbund     11.82   1.12
## Mass         11.40 128.65
## Diet*        -1.74   0.23
## Passerine    -2.02   0.07
## Aquatic      -1.09   0.06
```

```
describeBy(birds[,1:3],group = birds$Aquatic)
```

```
##
##  Descriptive statistics by group
## group: 0
##          vars  n  mean    sd median trimmed   mad min   max range  skew
## Family*     1 39 19.56 10.97  20.00   19.58 13.34 1.0  38.0  37.0 -0.02
## MaxAbund    2 39 47.60 77.33  25.11   31.56 27.83 0.2 413.6 413.4  3.17
## AvgAbund    3 39  5.58  8.13   3.70    4.06  3.50 0.2  47.6  47.4  3.75
##          kurtosis    se
## Family*     -1.25  1.76
## MaxAbund    11.06 12.38
## AvgAbund    15.89  1.30
## ------------------------------------------------------------
## group: 1
##          vars  n  mean    sd median trimmed   mad min    max  range skew
## Family*     1 15  8.00  4.47   8.00    8.00  5.93 1.0  15.00  14.00 0.00
## MaxAbund    2 15 37.90 64.26  20.00   24.05 22.60 0.2 255.65 255.45 2.54
## AvgAbund    3 15  5.96  8.84   2.74    4.41  2.14 0.2  31.80  31.60 1.89
##          kurtosis    se
## Family*     -1.44  1.15
## MaxAbund     5.83 16.59
## AvgAbund     2.44  2.28
```

```
with(birds, table(Diet, Aquatic))
```

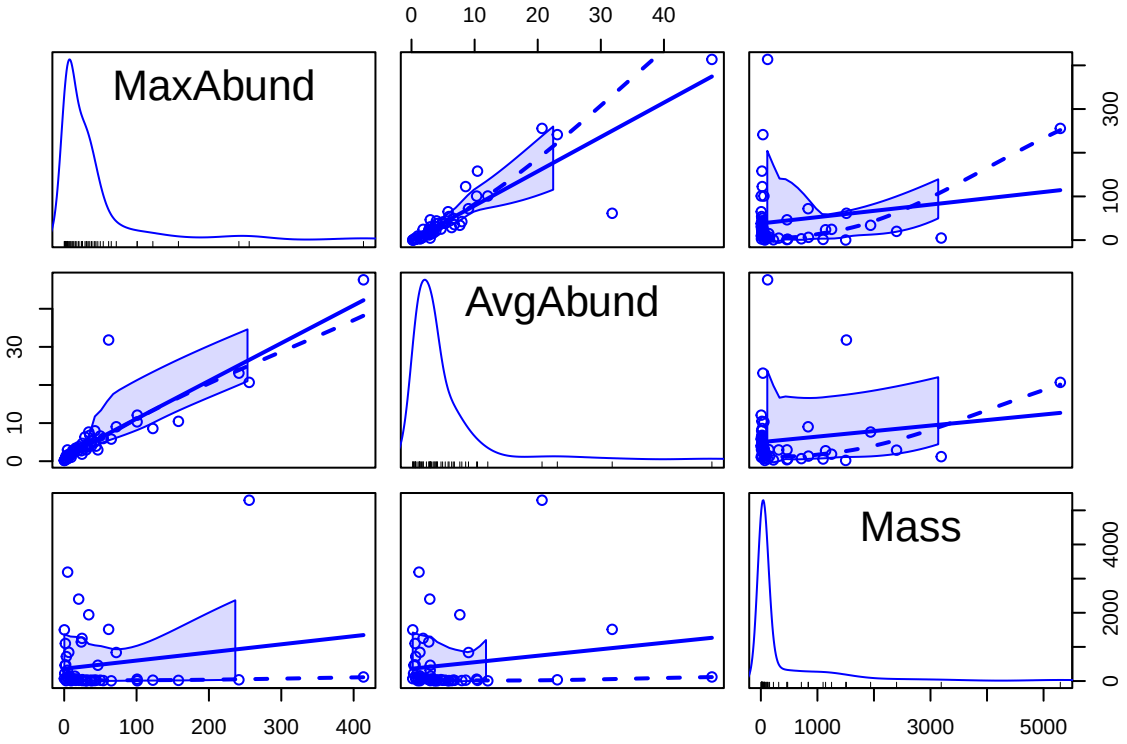
```
##               Aquatic
## Diet            0  1
##   Insect       14  6
##   InsectVert    1  1
##   Plant         2  0
##   PlantInsect  17  1
##   Vertebrate    5  7
```

```
with(birds, table(Diet, Passerine))
```

```
##               Passerine
## Diet            0  1
##   Insect       11  9
##   InsectVert    1  1
##   Plant         1  1
##   PlantInsect   4 14
##   Vertebrate   12  0
```

(ii) Focus on numerical variables and graph a scatterplot matrix.

```
scatterplotMatrix(birds[,2:4])
```



(iii) Do a simple regression of `MaxAbund` on `Mass`. Look at the summary for this model and interpret the results.

```
mod1 <- lm(MaxAbund ~ Mass, data = birds)
summary(mod1)
```

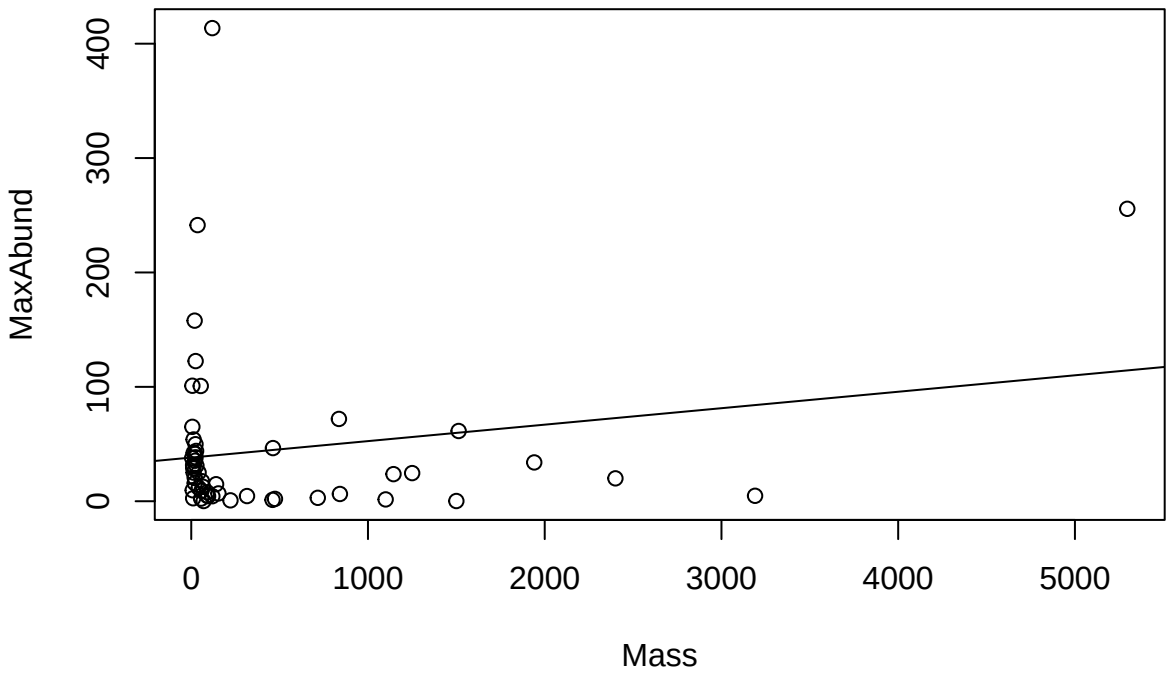
```
##
## Call:
## lm(formula = MaxAbund ~ Mass, data = birds)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79.30 -35.39 -22.06   2.62 373.72
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.16646   11.09065   3.441  0.00115 **
## Mass         0.01439    0.01059   1.358  0.18021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.89 on 52 degrees of freedom
## Multiple R-squared:  0.03427,    Adjusted R-squared:  0.0157
## F-statistic: 1.845 on 1 and 52 DF,  p-value: 0.1802
```

The residual table is highly assymetric. The slope cefficient is not significantly different from zero according to the test t-test. The R^2 value is very low. This does not seem to be a very good model.

(iv) Graph a scatterplot of the variables with the regression line. Comment.

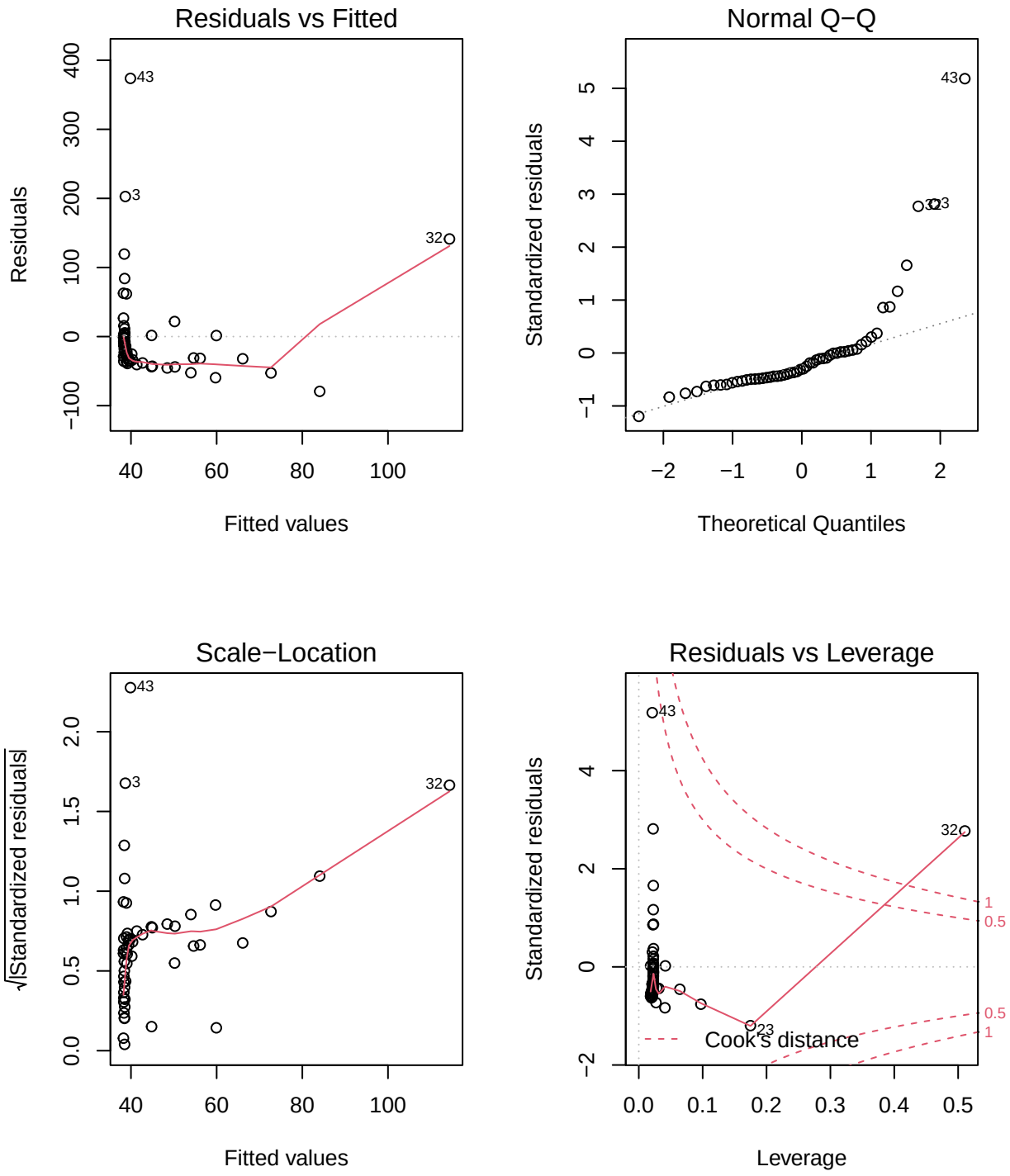
```
plot(MaxAbund ~ Mass, data = birds)
abline(mod1)
```



The lot shows that, indeed, this model is not useful, as it does not reflect the structure of the data.

(v) Do the diagnostic plots (standard and `residualPlots()`). Comment.

```
par(mfrow = c(2,2))
plot(mod1)
```



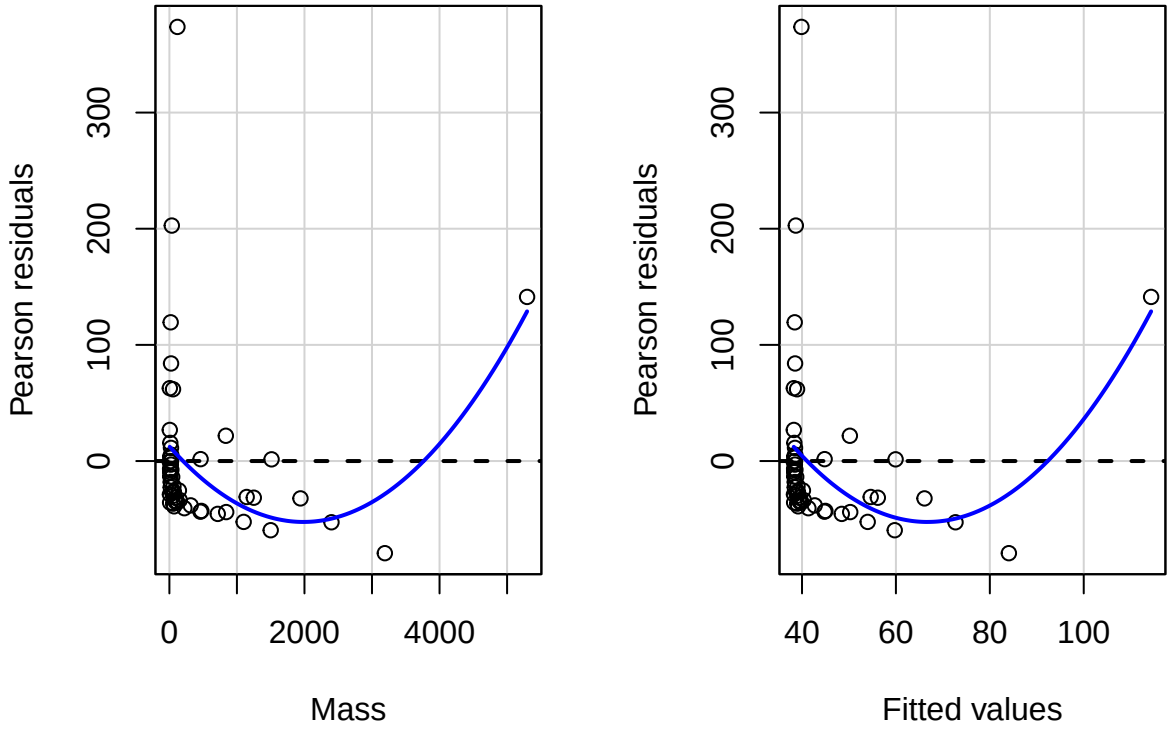

```
par(mfrow=c(1,1))
```

In this case all the diagnostic plots have issues. In residuals against fitted values, the majority of the residuals are negative, the red line is far from 0 and is not horizontal, and the residuals are not homogeneously spread in the plot. The quantile plot has some very large values on the right tail. The scale-location plot shows an increasing pattern for the dispersion of the data and the residuals vs leverage plot has one point with a very large value for leverage and high residual. This would not be an acceptable model.

Below we give another version of teh residual plot using the function `residualPlots` in the `car` package. The quadratic line corresponds to fitting a quadratic terms and the p-values tha appear below correspond to

a test on the significance of this term. The graphs show, again, that the fit is not good.

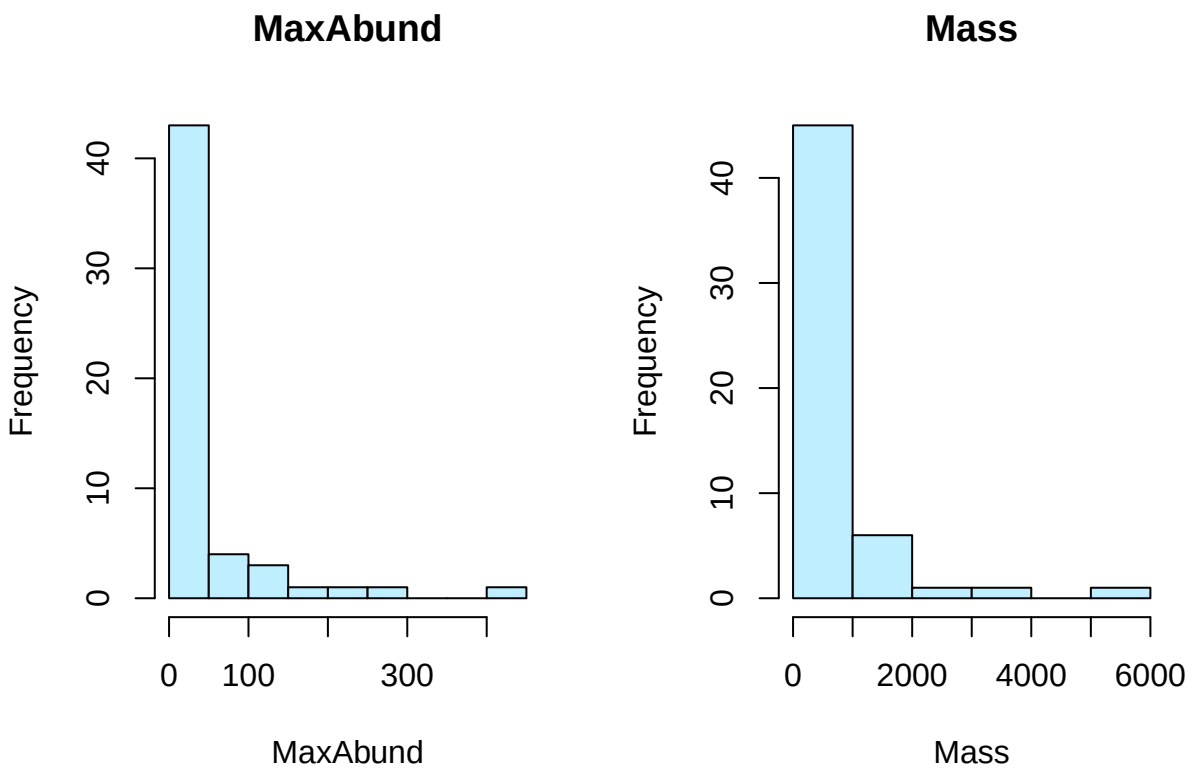
```
residualPlots(mod1)
```




```
##             Test stat Pr(>|Test stat|)
## Mass           2.9137         0.005291 **
## Tukey test     2.9137         0.003572 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(vi) Do histograms of response and regressor. Comment

```
par(mfrow = c(1,2))
hist(birds$MaxAbund, col = 'lightblue1',
     xlab='MaxAbund', main = 'MaxAbund')
hist(birds$Mass, col = 'lightblue1',
     xlab = 'Mass', main = 'Mass')
```




```
par(mfrow=c(1,1))
```

These histograms show that neither variable looks normal. The regressor variables does not have to follow a normal distribution, but we expect the output variable to be Gaussian. The histogram on the left clearly shows that this is not the case.

(vii) Perform a Shapiro-Wilk test for normality for the two variables. Interpret the result.

```
shapiro.test(birds$MaxAbund)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  birds$MaxAbund
## W = 0.5831, p-value = 3.872e-11
```

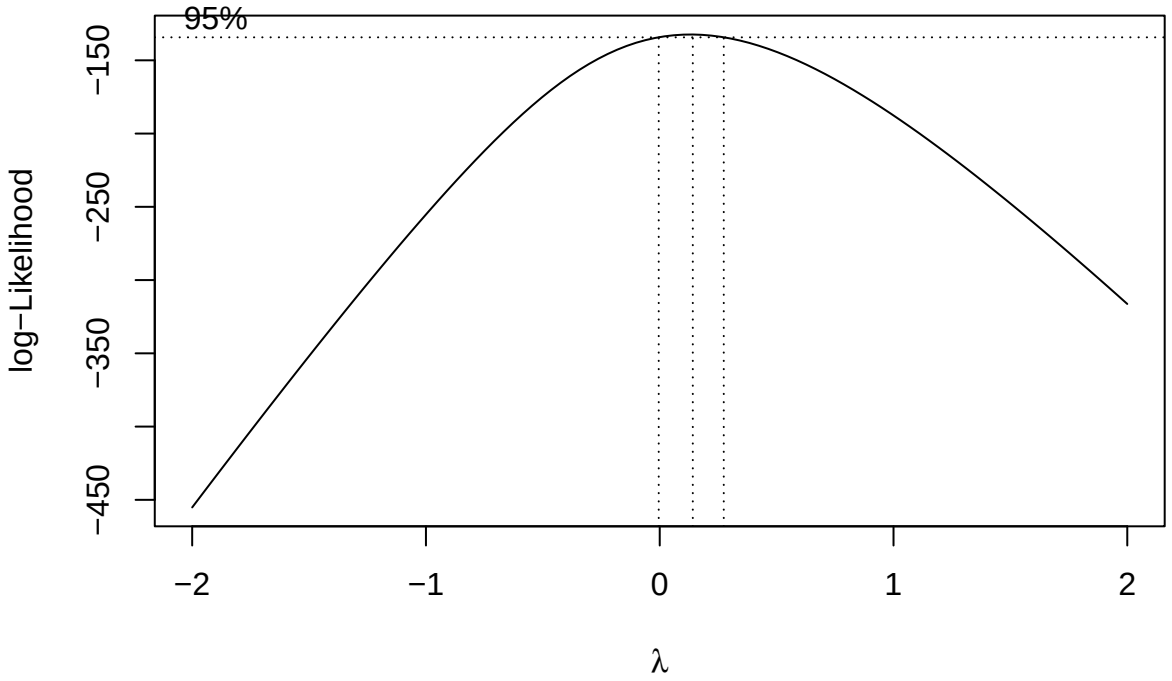
```
shapiro.test(birds$Mass)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  birds$Mass
## W = 0.54667, p-value = 1.155e-11
```

This confirms the comment we made above.

(viii) Use `boxcox` to select a transformation.

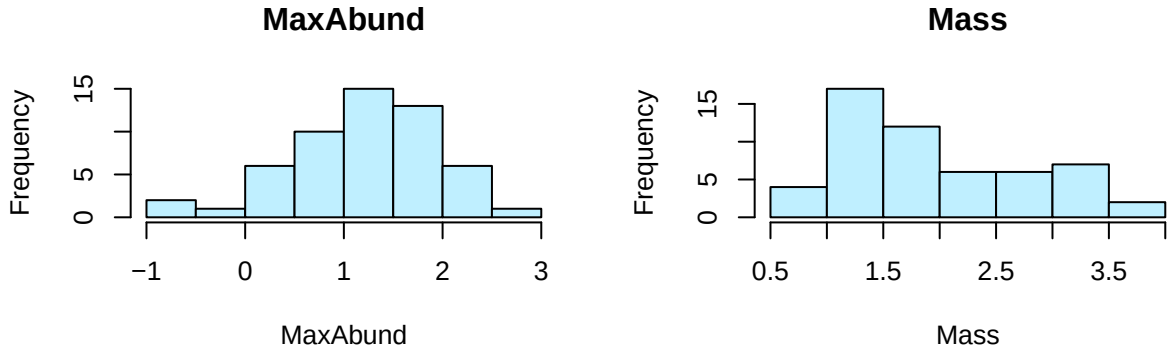
```
boxcox(mod1)
```



The Box-Cox transformations are a family of transformation that try to correct non-normality on the data. The graph presents the loglikelihood for the parameter λ. The plot indicates the maximum likelihood value and a confidence interval at approximately 95%. Usually, if this confidence interval includes zero, one should try a logarithmic transformation, which is the transformation corresponding to $\lambda = 0$.

(ix) Transform the data using `log10`, plot histograms and do Shapiro-Wilk again.

```
birds$logMaxAbund <- log10(birds$MaxAbund)
birds$logMass <- log10(birds$Mass)
attach(birds)
par(mfrow=c(2,2))
hist(logMaxAbund, col = 'lightblue1',
     xlab = 'MaxAbund', main = 'MaxAbund')
hist(logMass, col = 'lightblue1',
     xlab = 'Mass', main = 'Mass')
par(mfrow=c(1,1))
```




```
shapiro.test(logMaxAbund)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  logMaxAbund
## W = 0.96708, p-value = 0.1431
```

```
shapiro.test(logMass)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  logMass
## W = 0.93716, p-value = 0.007107
```

After the transformation, the histogram `MaxAbund` variable looks closer to a Gaussian distribution, and the Shapiro Wilk test confirms this. The regressor does not have a Gaussian distribution, but as we argued before, this is not expected, and is not a condition for the model.

(x) Fit a new model with this data.

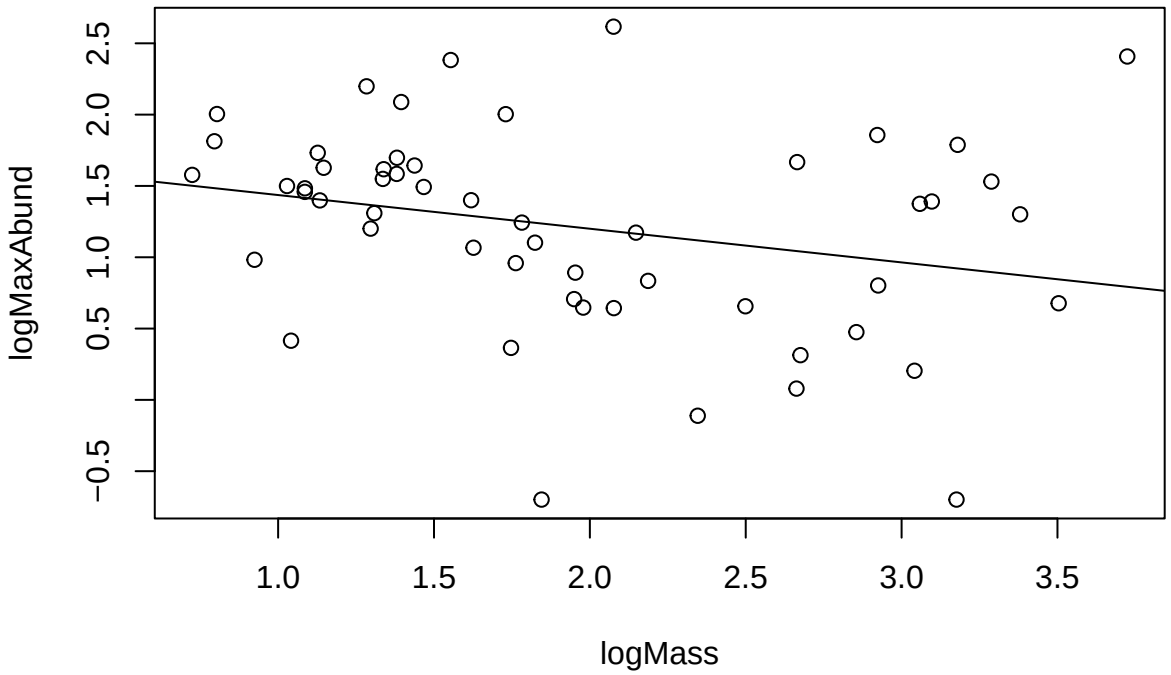
```
model2 <- lm(logMaxAbund ~ logMass, data = birds)
summary(model2)
```

```
##
## Call:
## lm(formula = logMaxAbund ~ logMass, data = birds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93562 -0.39982  0.05487  0.40625  1.61469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6724     0.2472   6.767 1.17e-08 ***
## logMass      -0.2361     0.1170  -2.019   0.0487 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6959 on 52 degrees of freedom
## Multiple R-squared:  0.07267,    Adjusted R-squared:  0.05484
## F-statistic: 4.075 on 1 and 52 DF,  p-value: 0.04869
```

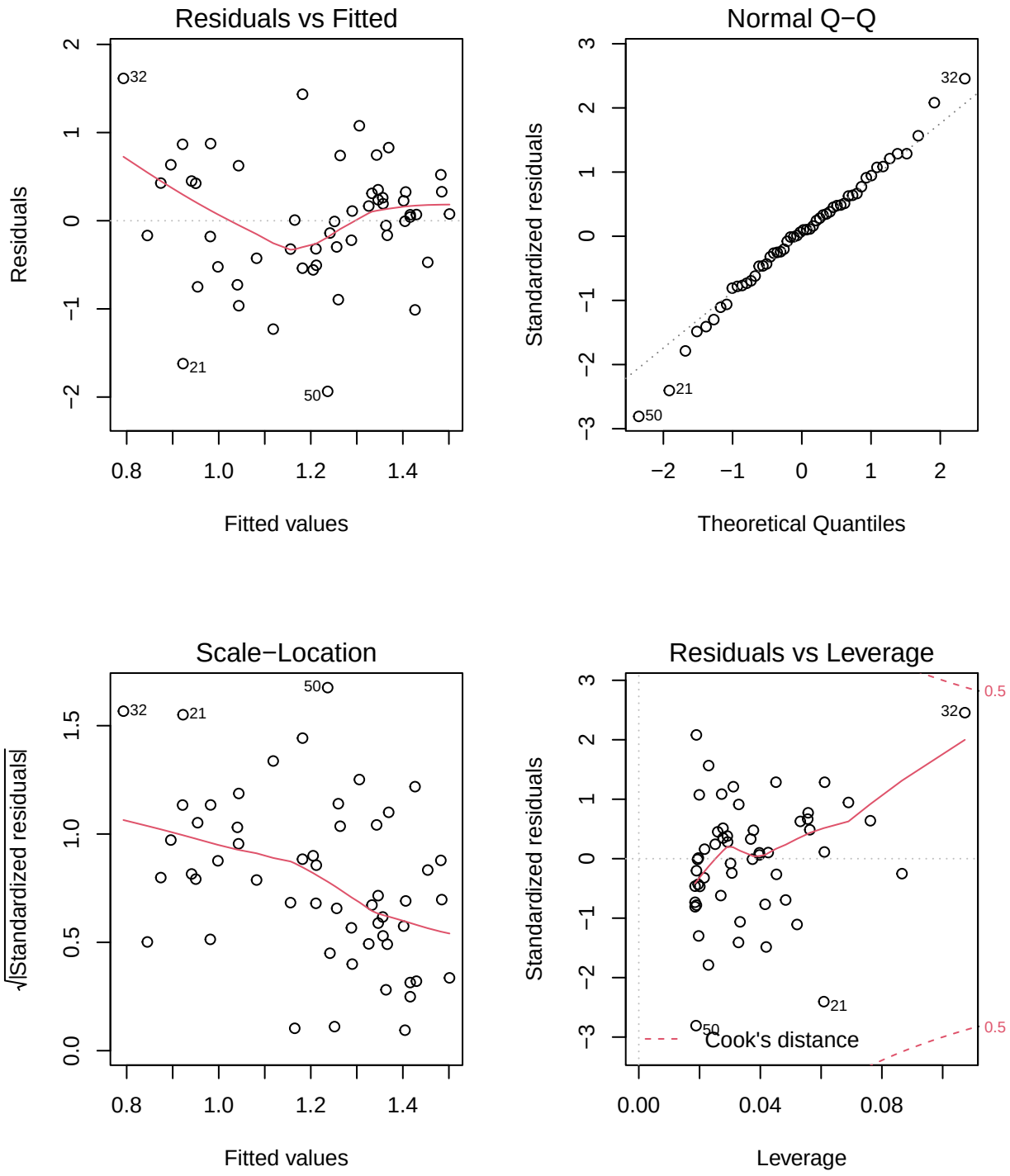
The model has improved somewhat, although the slope is only marginally significant and the R^2 is still very low. The residual table is much better.

(xi) Plot the model on a scatterplot of the transformed data. Plot the diagnostic graphs.

```
plot(logMaxAbund ~ logMass, data = birds)
abline(model2)
```




```
par(mfrow=c(2,2))
plot(model2)
```




```
par(mfrow=c(1,1))
```

The diagnostic plots have improved, particularly the quantile plot, but still say that there are problems with this model. The residuals vs fitted values plot does not show a homogeneous cloud of points, and the red line oscillates, the scale-location plot shows a decreasing tendency, and the residuals vs leverage graph, although much better than before still shows some large standardized residuals.

(xii) Can you write down the equation for your model?

In logarithmic terms

$$logMaxAbund = 1.6724 - 0.2361 logMass$$

and in the original variables

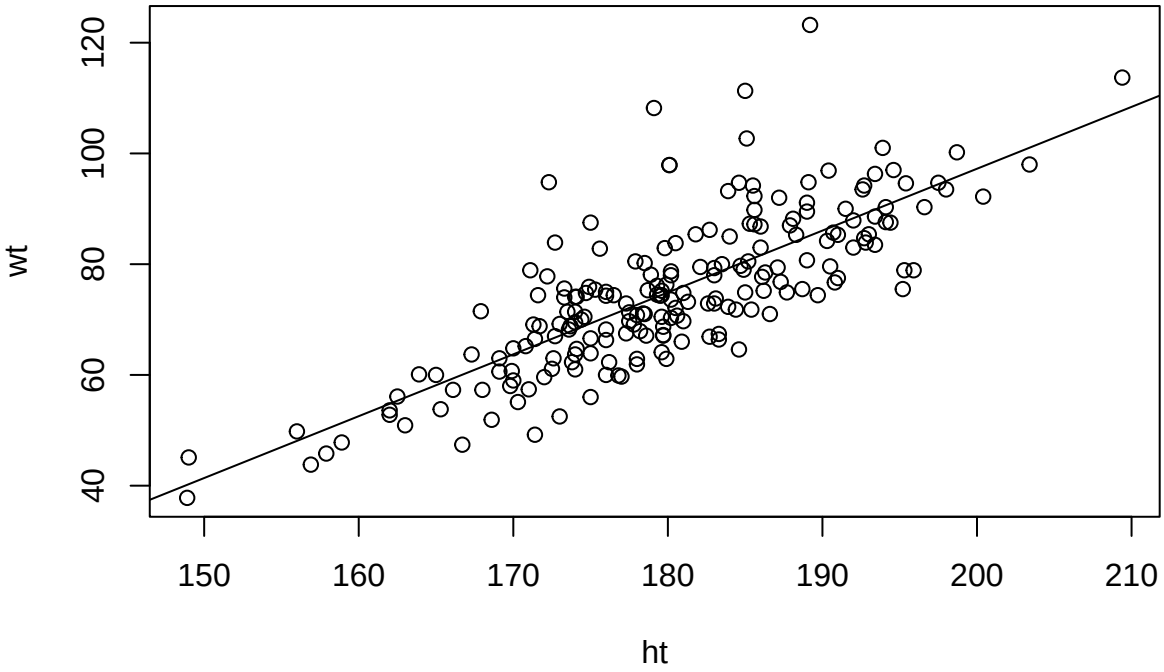
$$MaxAbund = 47.033 * Mass^{0.2361}$$

Exercise 3

For this exercise we will use the data set `ais` in the package `DAAG`, that has information about Australian athletes. Look up the help for this set and get familiar with the variables in it. To load the data set you need to load the library and run the command `data(ais)`.

(i) Plot a scatterplot of `wt` as a function of `ht`. Add the corresponding regression line.

```
library(car)
library(DAAG)
data(ais)
plot(wt ~ ht, data = ais)
abline(lm(wt ~ ht, data = ais))
```



(ii) Fit a regression to this data. What are the estimated values for the intercept and slope? Write down the regression model in this case and interpret the meaning of the coefficients.

```
wt.lm <- lm(wt ~ ht, data = ais)
summary(wt.lm)
```

```
##
## Call:
## lm(formula = wt ~ ht, data = ais)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.372  -5.296  -1.196   4.378  38.031
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -126.19049   11.39566  -11.07   <2e-16 ***
## ht             1.11712    0.06318   17.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.72 on 200 degrees of freedom
## Multiple R-squared:  0.6099, Adjusted R-squared:  0.6079
## F-statistic: 312.6 on 1 and 200 DF,  p-value: < 2.2e-16
```

$$wt = -126.19 + 1.117 \times ht$$

For every increase of 1 com in height, weight increases by 1.117 kg. The weight for height = 0 is -126.19 (which does not make any sense)

(iii) What are the results of the t-tests in this example?

Both p values are very small, so we reject the null hypotheses of parameters equal to zero.

(iv) What would be the predicted value according to this model for the weight corresponding to a height of 195 cm?

```
predict.lm(wt.lm, data.frame(ht = 195), interval = 'p')
```

```
##        fit      lwr     upr
## 1 91.64864 74.31233 108.985
```

The predicted value is 91.65

(v) Describe the sampling distribution for the estimated parameters in the previous regression.

The estimated parameters are $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$, which have a normal distribution:

$$\hat{\boldsymbol{\beta}} = N\big((\beta_0, \beta_1)', \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\big)$$

The matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is obtained in R with

```
(invXtX <- summary(wt.lm)$cov.unscaled)
```

```
##              (Intercept)            ht
## (Intercept)  1.707985851 -9.455846e-03
## ht          -0.009455846  5.250216e-05
```

The variance is unknown and is estimated by the mean square. The standard deviation is

```
summary(wt.lm)$sigma
```

```
## [1] 8.71962
```

and the estimated variance is

```
summary(wt.lm)$sigma^2
```

```
## [1] 76.03176
```

The estimated covariance matrix for $\hat{\boldsymbol{\beta}}$ can be obtained with

```
vcov(wt.lm)
```

```
##             (Intercept)           ht
## (Intercept) 129.8611778 -0.718944673
## ht           -0.7189447  0.003991832
```

or multiplying $\hat{\sigma}^2$ times $(\mathbf{X}'\mathbf{X})^{-1}$

```
(summary(wt.lm)$sigma^2)*invXtX
```

```
##             (Intercept)           ht
## (Intercept) 129.8611778 -0.718944673
## ht           -0.7189447  0.003991832
```

(vi) Give a confidence interval at a confidence level of 98% for the parameters of the regression.

```
confint(wt.lm,level = 0.98)
```

```
##                        1 %       99 %
## (Intercept) -152.9148859 -99.466093
## ht             0.9689558   1.265292
```

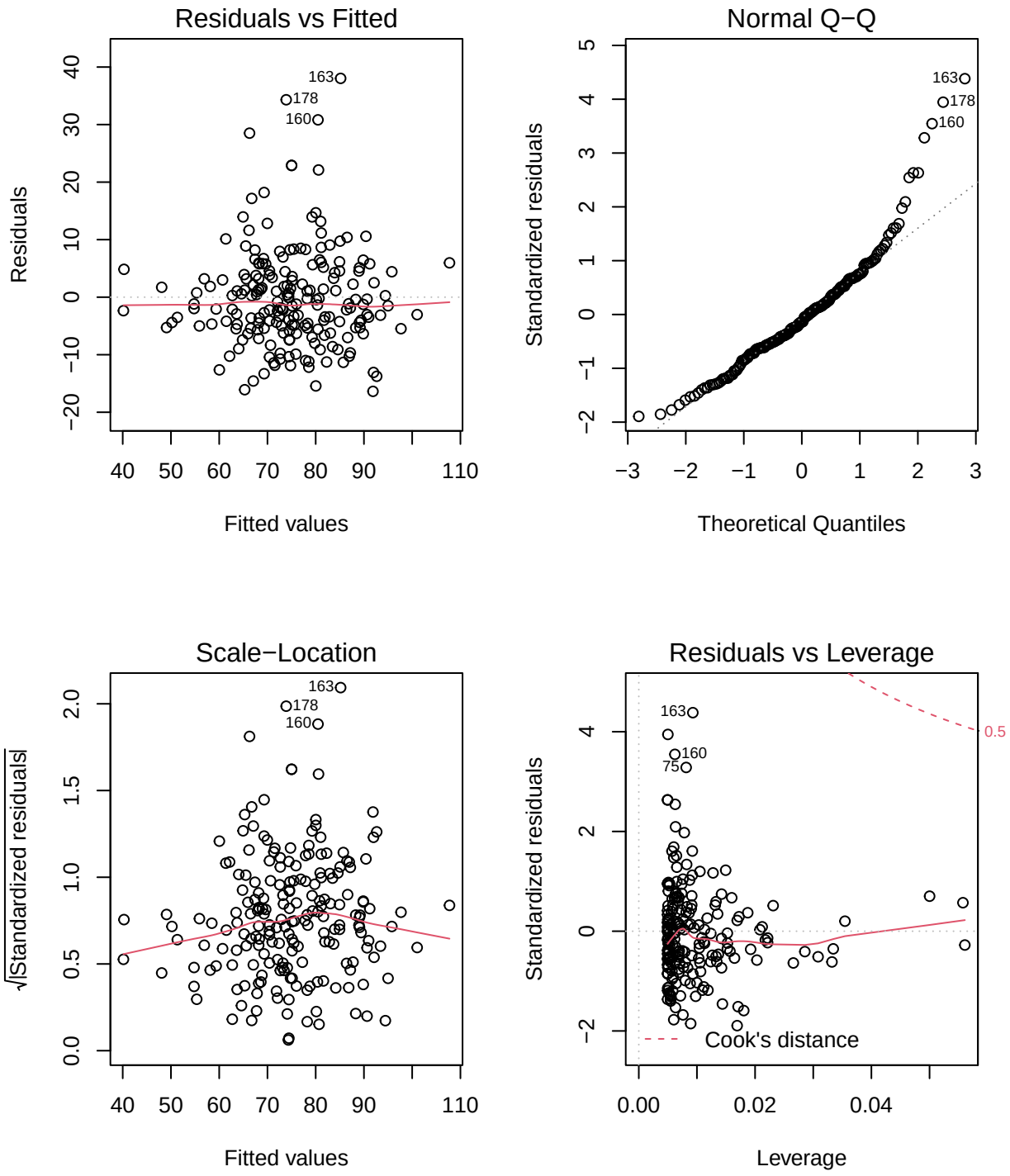
(vii) Find the value for the R^2 and comment on its meaning. $R^2 = 0.6099$. This is the proportion of expleined variability. It is equal to the square of the correlation between `wt` and `ht`:

```
cor(ais$wt,ais$ht)^2
```

```
## [1] 0.6098549
```

(viii) Draw diagnostic plots and discuss the results.

```
par(mfrow=c(2,2))
plot(wt.lm)
```




```
par(mfrow=c(1,1))
```

In the first plot we see that there is a lack of symmetry in the residuals near the center of the fitted values, with larger positive values. The normal quantile plot is not good since the fit in the right tail is off the line. In the scale-location graph there is a curved pattern in the red line that indicates variations in the dispersion of the residuals, which may indicate lack of homogeneity in the variance. Finally, the last graph shows some very large residuals with small leverage and some large leverage points with small residuals. In general these plots are not very good and make the model suspect.

(ix) Produce an anova table for the regression and interpret the test. How does this compare to the results

of the summary table of the regression?

```
anova(wt.lm)
```

```
## Analysis of Variance Table
##
## Response: wt
##            Df Sum Sq Mean Sq F value    Pr(>F)
## ht          1  23770   23770  312.63 < 2.2e-16 ***
## Residuals 200  15206      76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

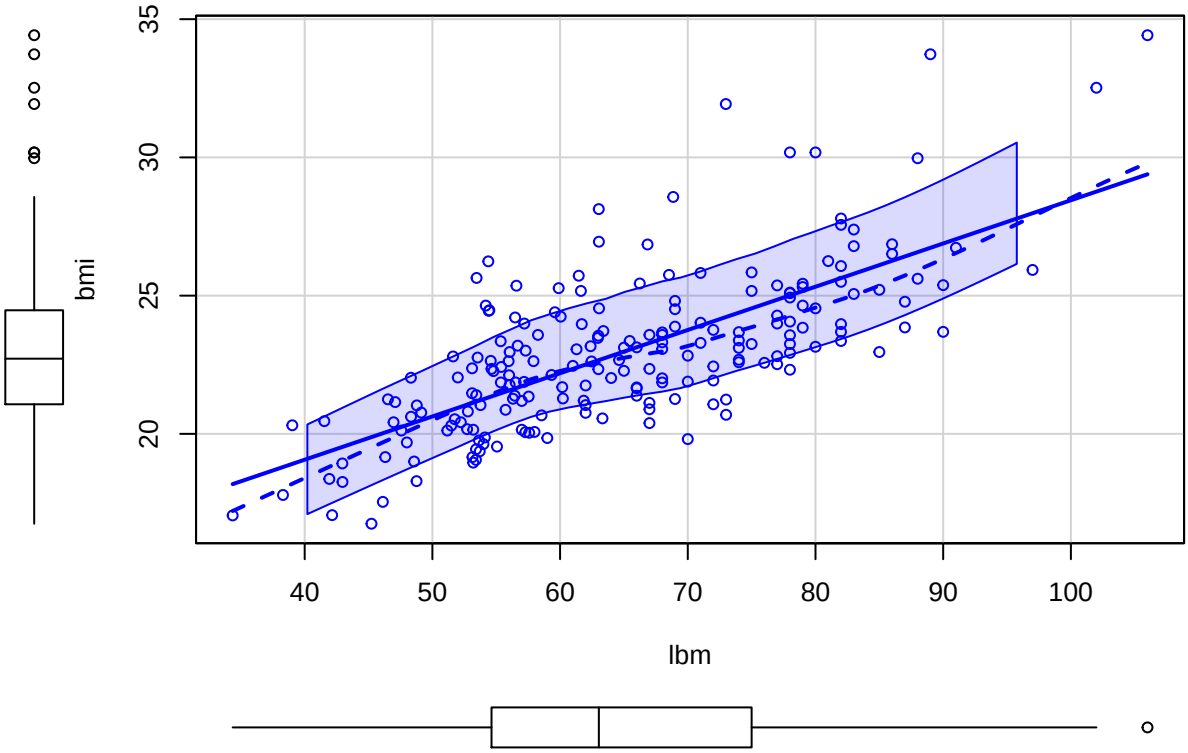
The test has a very small p-value, so the regression is significant. The test statistic and p-value appear at the bottom of the summary table printed previously.

Exercise 4

For this exercise we will use the data set `ais` in the package `DAAG`, that has information about Australian athletes. Look up the help for this set and get familiar with the variables in it. To load the data set you need to load the library and run the command `data(ais)`.

(i) Draw a scatterplot of `bmi` against `lbm`. For this, use the function `scatterplot` in the `car` package. This function draws the points and also a simple regression line for the two variables. Moreover, it also plots a broken line that represents a local smoother function for the points as well as confidence bands for the smoother. The function also graphs boxplots for both variables on the corresponding axes. How would you interpret the differences between the regression line and the local smoother function that you see on the graph?

```
library(car)
library(DAAG)
data(ais)
scatterplot(bmi ~ lbm, data = ais)
```



The difference indicates whether the local behavior of the cloud of points is captured in the linear regression model. In his case there seem to be no important discrepancies.

(ii) Use the function `lm` to fit a regression line to this data. Write down explicitly the model that you get and interpret the meaning of the coefficients.

```
model1 <- lm(bmi ~ lbm, data = ais)
summary(model1)
```

```
##
## Call:
## lm(formula = bmi ~ lbm, data = ais)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9477 -1.4628 -0.2844  0.9964  7.7061
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.80834    0.71798   17.84   <2e-16 ***
## lbm          0.15642    0.01085   14.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.011 on 200 degrees of freedom
## Multiple R-squared:  0.5096, Adjusted R-squared:  0.5071
## F-statistic: 207.8 on 1 and 200 DF,  p-value: < 2.2e-16
```

The model is

$$bmi = 12.81 + 0.156 \times lbm$$

For the interpretation see the previous exercise.

(iii) Use the function `summary` on the output of the regression. Interpret the t-tests in the table. Are the parameters different from zero?

Both parameters are significantly different from zero.

(iv) Describe the sampling distribution for the estimated parameters in this regression.

The estimated parameters are $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$, which have a normal distribution:

$$\hat{\boldsymbol{\beta}} = N\big((\beta_0, \beta_1)', \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\big)$$

The matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is obtained in `R` with

```
(invXtX <- summary(model1)$cov.unscaled)
```

```
##              (Intercept)           lbm
## (Intercept)  0.127518515 -1.889333e-03
## lbm         -0.001889333  2.912324e-05
```

The variance is unknown and is estimated by the mean square. The standard deviation is

```
summary(model1)$sigma
```

```
## [1] 2.010592
```

and the estimated variance is

```
summary(model1)$sigma^2
```

```
## [1] 4.042481
```

The estimated covariance matrix for $\hat{\boldsymbol{\beta}}$ can be obtained with

```
vcov(model1)
```

```
##              (Intercept)           lbm
## (Intercept)  0.515491201 -0.0076375915
## lbm         -0.007637591  0.0001177301
```

or multiplying $\hat{\sigma}^2$ times $(\mathbf{X}'\mathbf{X})^{-1}$

```
(summary(model1)$sigma^2)*invXtX
```

```
##              (Intercept)           lbm
## (Intercept)  0.515491201 -0.0076375915
## lbm         -0.007637591  0.0001177301
```

(v) Give confidence intervals at a confidence level of 98% for the parameters of the regression.

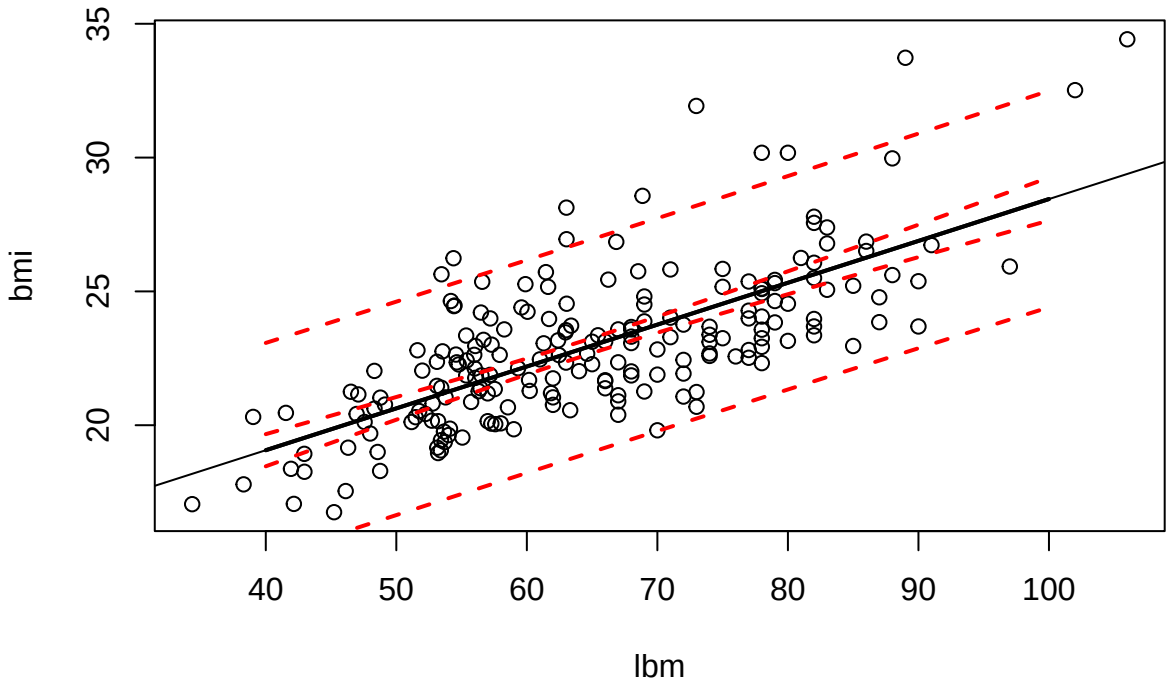
```
confint(model1, level = 0.98)
```

```
##                    1 %       99 %
## (Intercept) 11.1245833 14.4920930
## lbm          0.1309745  0.1818657
```

(vi) Draw a scatterplot of the data and include the regression line with confidence bands for the mean and predicted values.

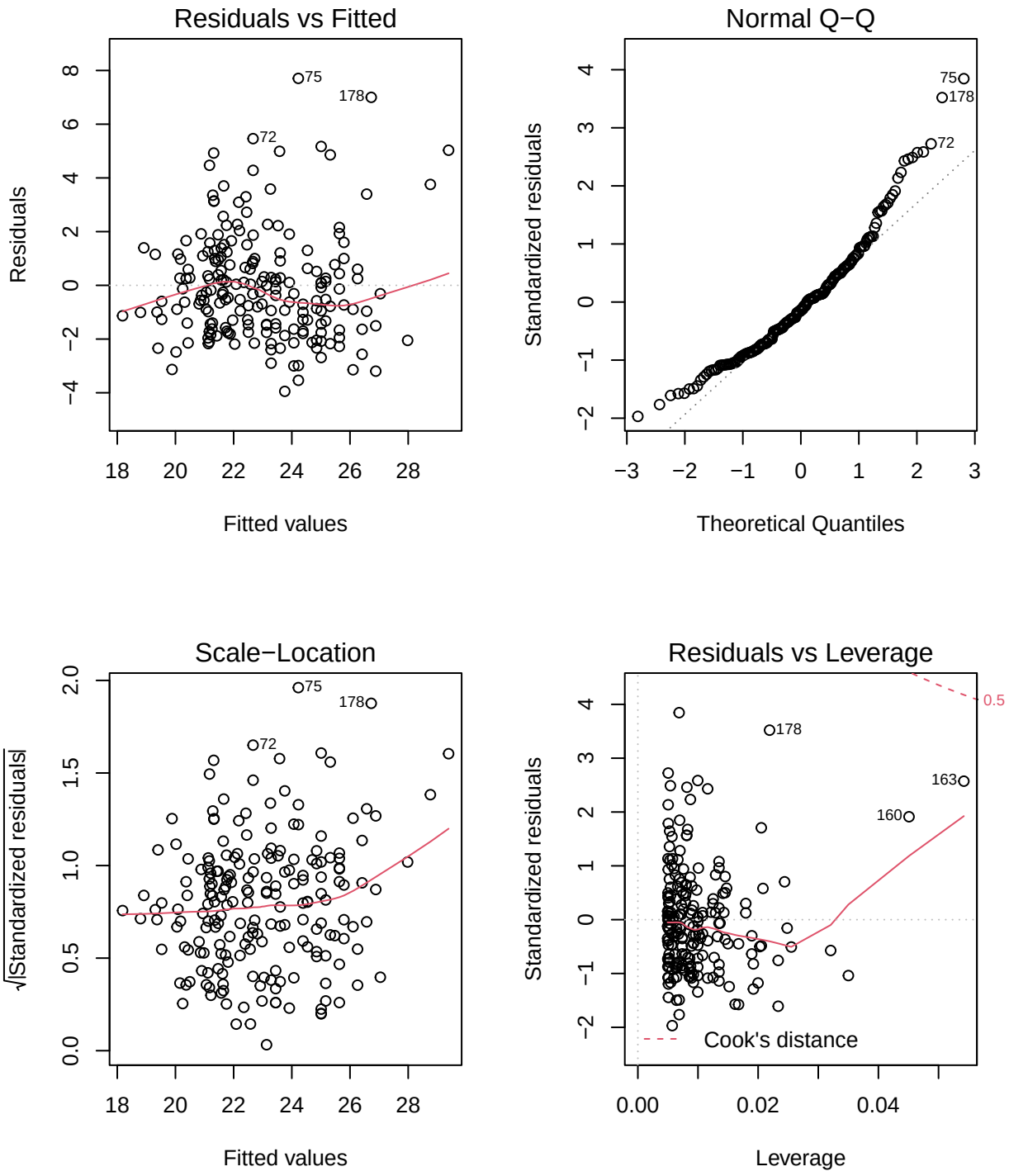
```
plot(bmi ~ lbm, data = ais)
abline(model1)
new.lbm <- data.frame(lbm=seq(40,100,
                              length.out = 15))
pc <- predict(model1,new.lbm, int='c')
matlines(new.lbm$lbm, pc, lty=c(1,2,2),lwd=rep(2,3),
```

```
        col=c('black','red','red'))
pp <- predict(model1,new.lbm, int='p')
matlines(new.lbm$lbm, pp, lty=c(1,2,2),lwd=rep(2,3),
        col=c('black','red','red'))
```



(vii) Draw diagnostic plots and discuss the results.

```
par(mfrow=c(2,2))
plot(model1)
```




```
par(mfrow=c(1,1))
```

The comments are similar as those for the previous exercise.

(viii) Produce an anova table for the regression and interpret the test. How does this compare to the results of the summary table of the regression?

```
anova(model1)
```

```
## Analysis of Variance Table
```

```
##
## Response: bmi
##            Df Sum Sq Mean Sq F value    Pr(>F)
## lbm         1 840.13  840.13  207.82 < 2.2e-16 ***
## Residuals 200 808.50    4.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For comments, see previous exercise.