Simple Regression

Code

Grace Davis

2023-03-08

Hide

```
library(tidyverse)
library(openintro)
library(lmtest)
library(kableExtra)
```

Exercise 1

The relationship appears to be curvilinear. The graph shows a positive relationship between stiffness and density –as one increases, so does the other. The relationship is not an exact match or straight line, so the graph and portrayed relationship is curvilinear.

Hide

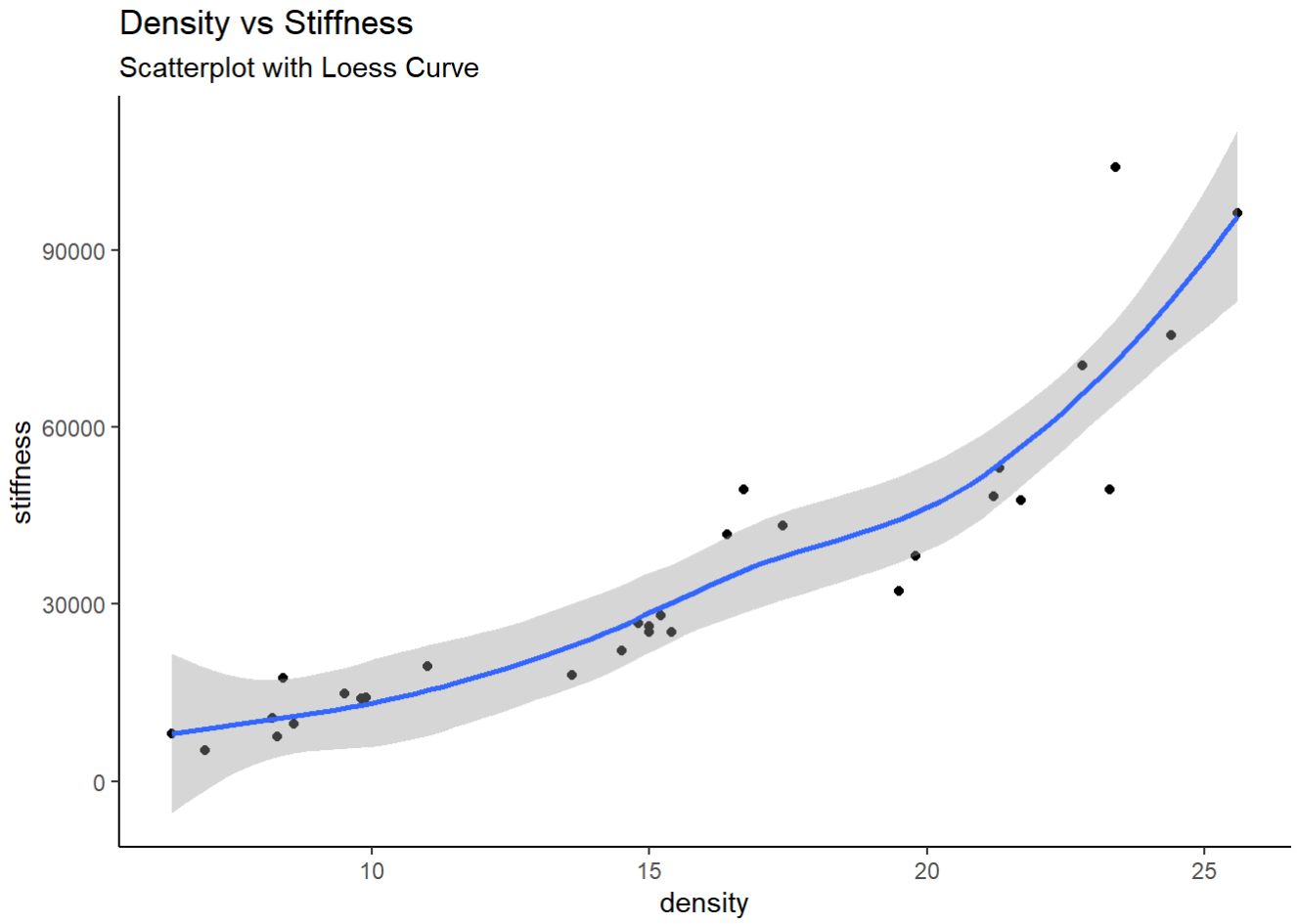
```
stiffness.data <- read.table(header=TRUE, file="particleboard.txt")
head(stiffness.data)
```

```
##   density stiffness
## 1     9.5     14814
## 2     9.8     14007
## 3     8.3      7573
## 4     8.6      9714
## 5     7.0      5304
## 6    17.4     43243
```

Hide

```
ggplot(stiffness.data, aes(x=density, y=stiffness)) +
  geom_point() +
  geom_smooth(method=loess) +
  theme_classic() +
  labs(title="Density vs Stiffness", subtitle="Scatterplot with Loess Curve")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Exercise 2

Based on the p value being less that 0.05, the line is statistically significant.

Hide

```
summary(
  sl.model <- lm(stiffness ~ density, data=stiffness.data)
)
```

```
##
## Call:
## lm(formula = stiffness ~ density, data = stiffness.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18326  -8520   1070   4220  38380
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25753.8     6126.8  -4.203 0.000243 ***
## density       3912.1      371.3  10.535 3.01e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11660 on 28 degrees of freedom
## Multiple R-squared:  0.7985, Adjusted R-squared:  0.7913
## F-statistic:   111 on 1 and 28 DF,  p-value: 3.009e-11
```

Exercise 3

The 95% interval prediction for the stiffness of a board with a density of 20 pound per cubic foot is (27957.1, 77020.45).

Hide

```
stiffness20data <-data.frame(density=20)

predict(sl.model, newdata=stiffness20data, interval="prediction")
```

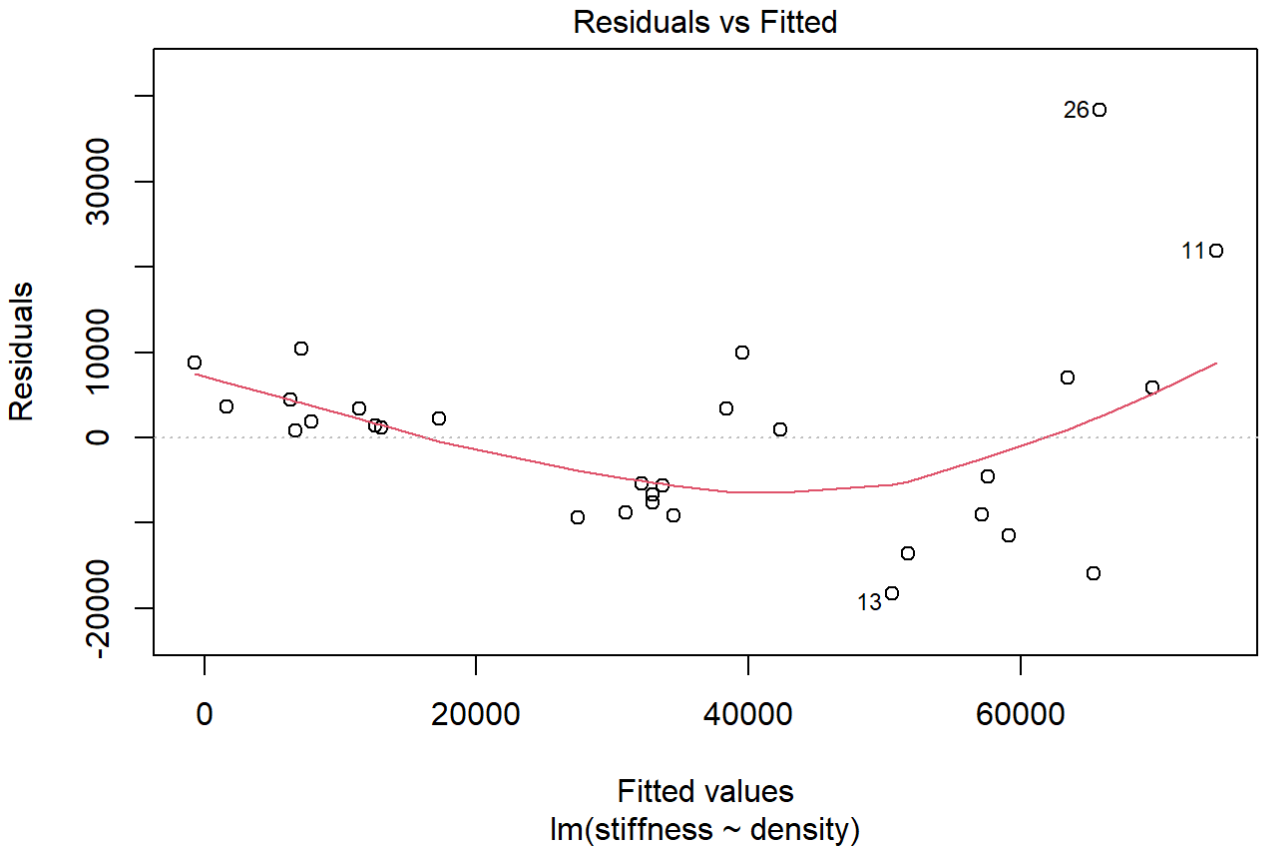
```
##        fit     lwr      upr
## 1 52488.77 27957.1 77020.45
```

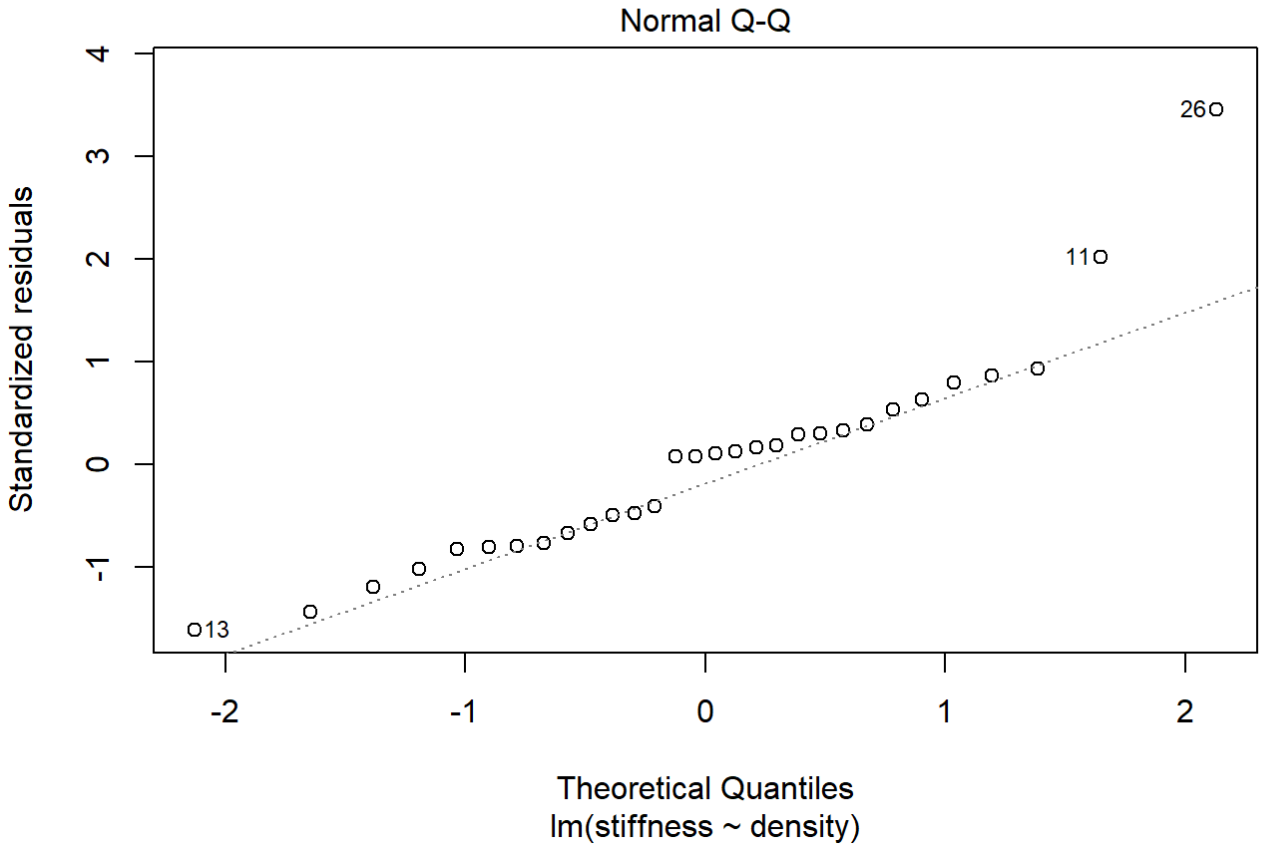
Exercise 4

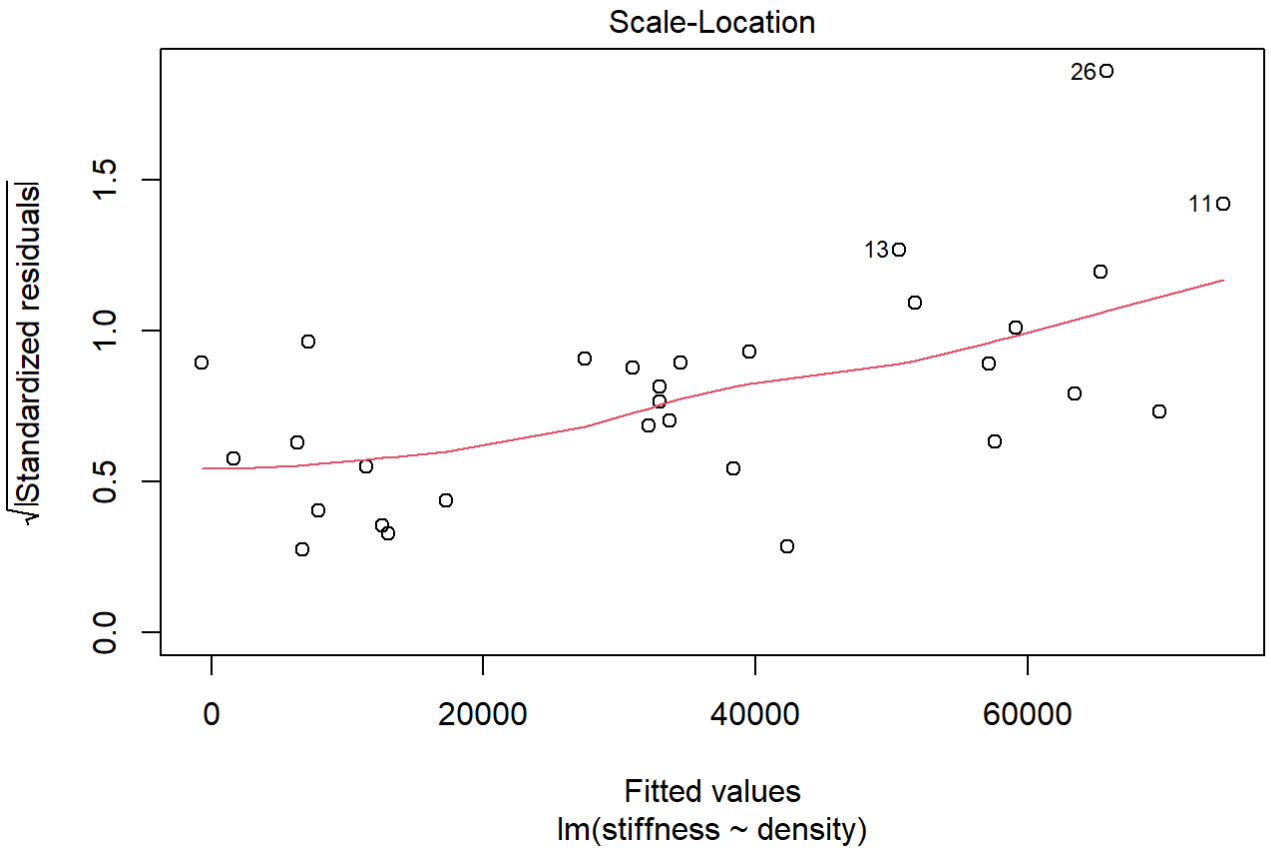
Based on the bptest on the straight line model, we can see a significant p-value which means the residual variance is not constant and there is not homoskedasicity in the model.

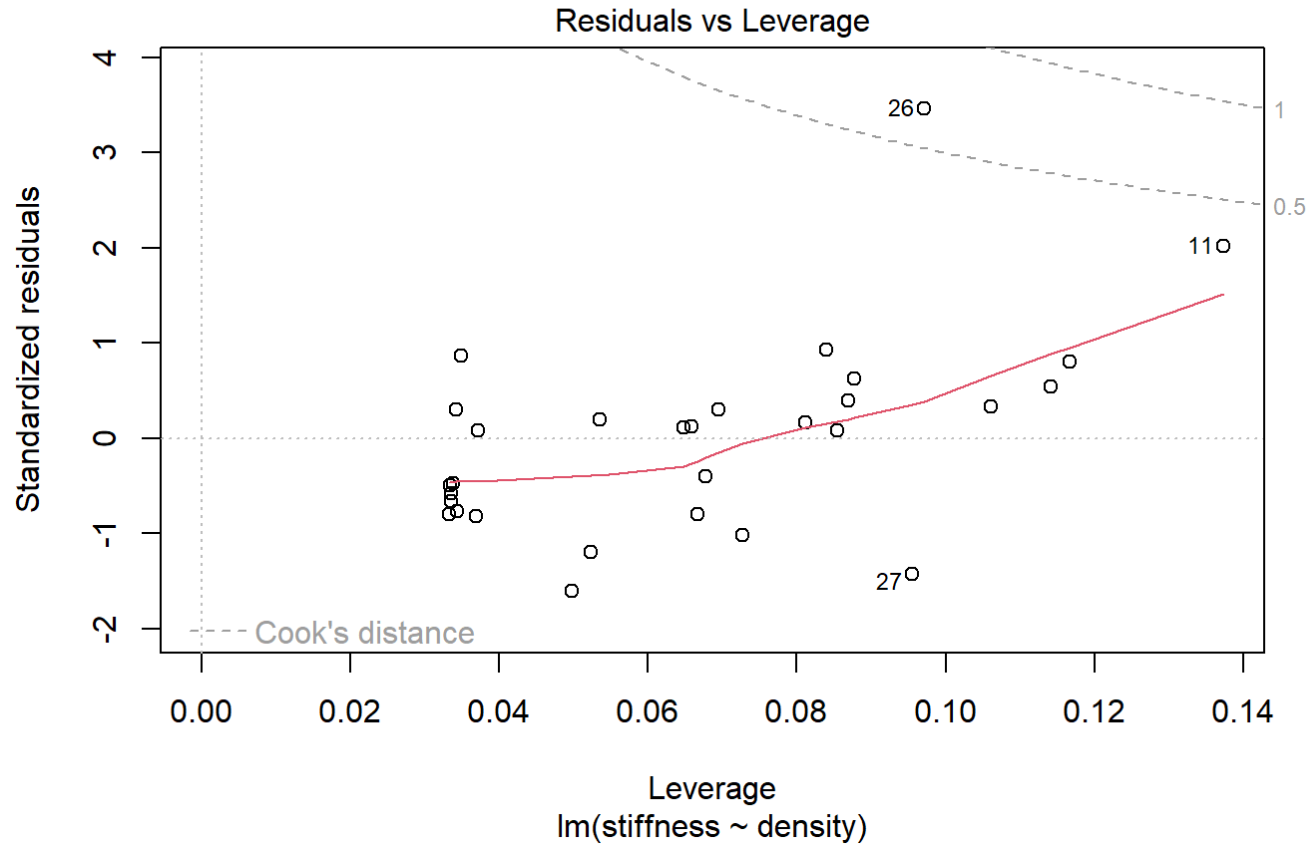
Hide

```
plot(sl.model)
```


Residuals vs Fitted
Residuals
30000
10000
0
-20000
26
11
13
0
20000
40000
60000
Fitted values
lm(stiffness ~ density)


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
26
11
13
-2
-1
0
1
2
Theoretical Quantiles
lm(stiffness ~ density)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
20000
40000
60000
26
11
13
Fitted values
lm(stiffness ~ density)


Residuals vs Leverage
Standardized residuals
-2
-1
0
1
2
3
4
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
26
11
27
1
0.5
Cook's distance
Leverage
lm(stiffness ~ density)

Hide

```
bptest(sl.model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  sl.model
## BP = 5.8882, df = 1, p-value = 0.01524
```

Exercise 5

The plot for log model is a better fit and more linear than the previous model. Additionally, the p-value is not significant, therefore it is a better fit.

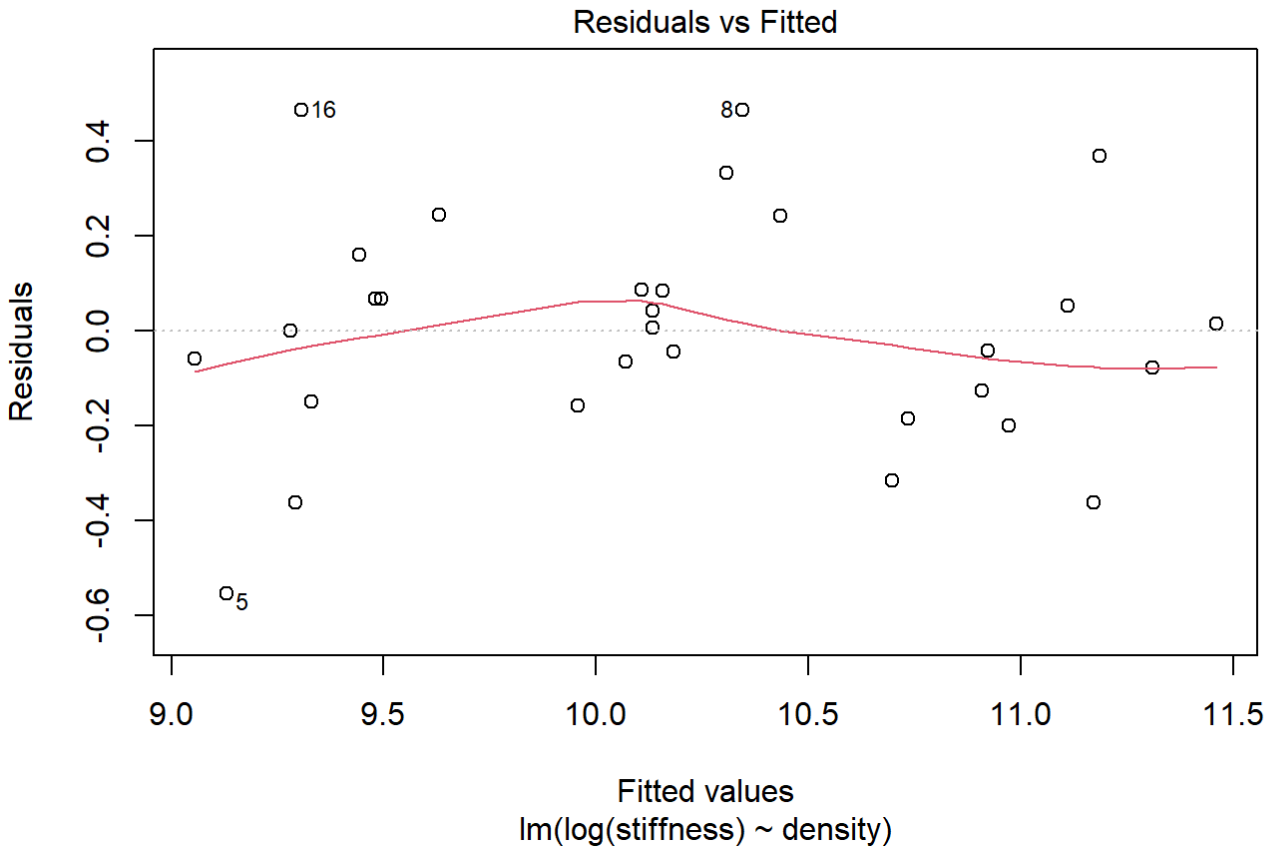
Hide

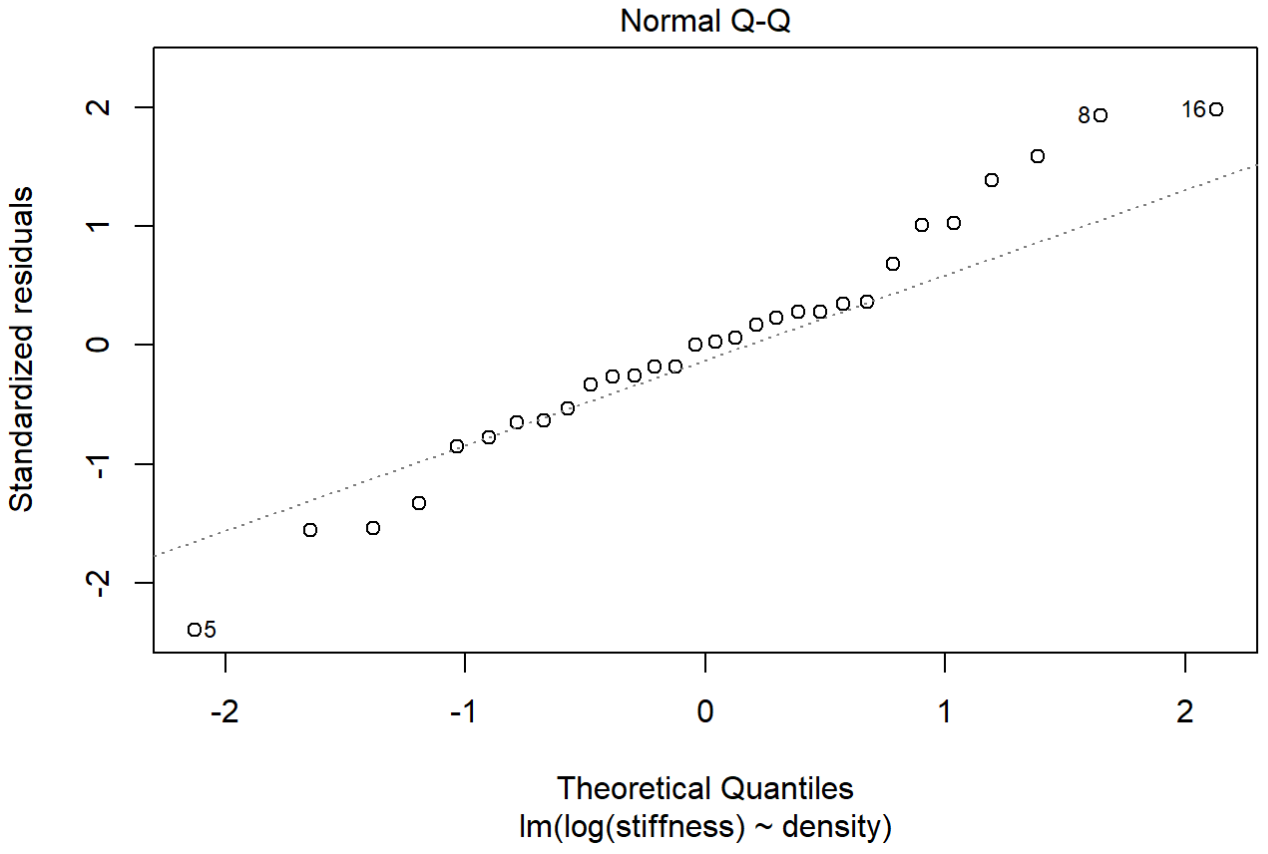
```
summary(
  log.model <- lm(log(stiffness) ~ density, data=stiffness.data)
)
```

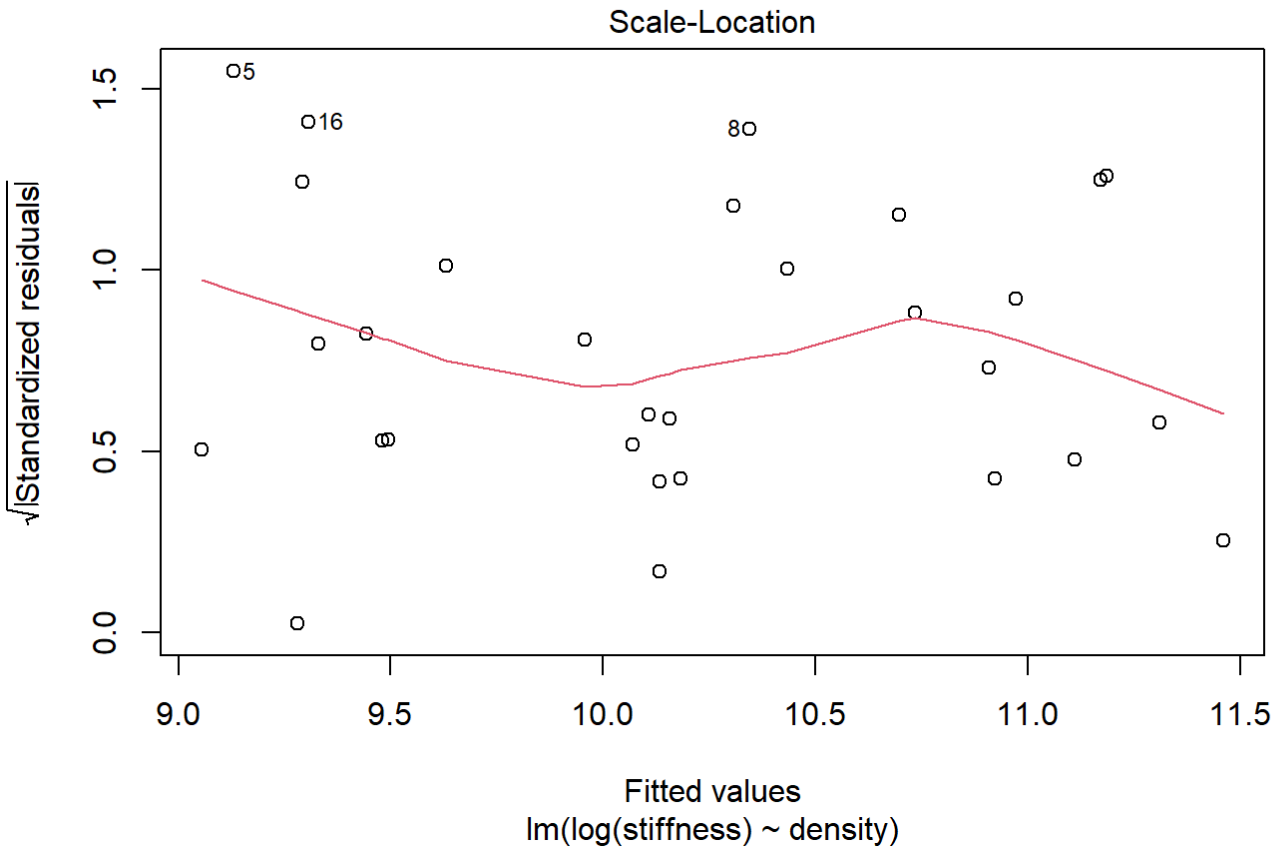
```
##
## Call:
## lm(formula = log(stiffness) ~ density, data = stiffness.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55391 -0.14339  0.00350  0.08607  0.46454
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.25310    0.12836   64.30  < 2e-16 ***
## density      0.12529    0.00778   16.11 1.09e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2444 on 28 degrees of freedom
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.8991
## F-statistic: 259.4 on 1 and 28 DF,  p-value: 1.09e-15
```

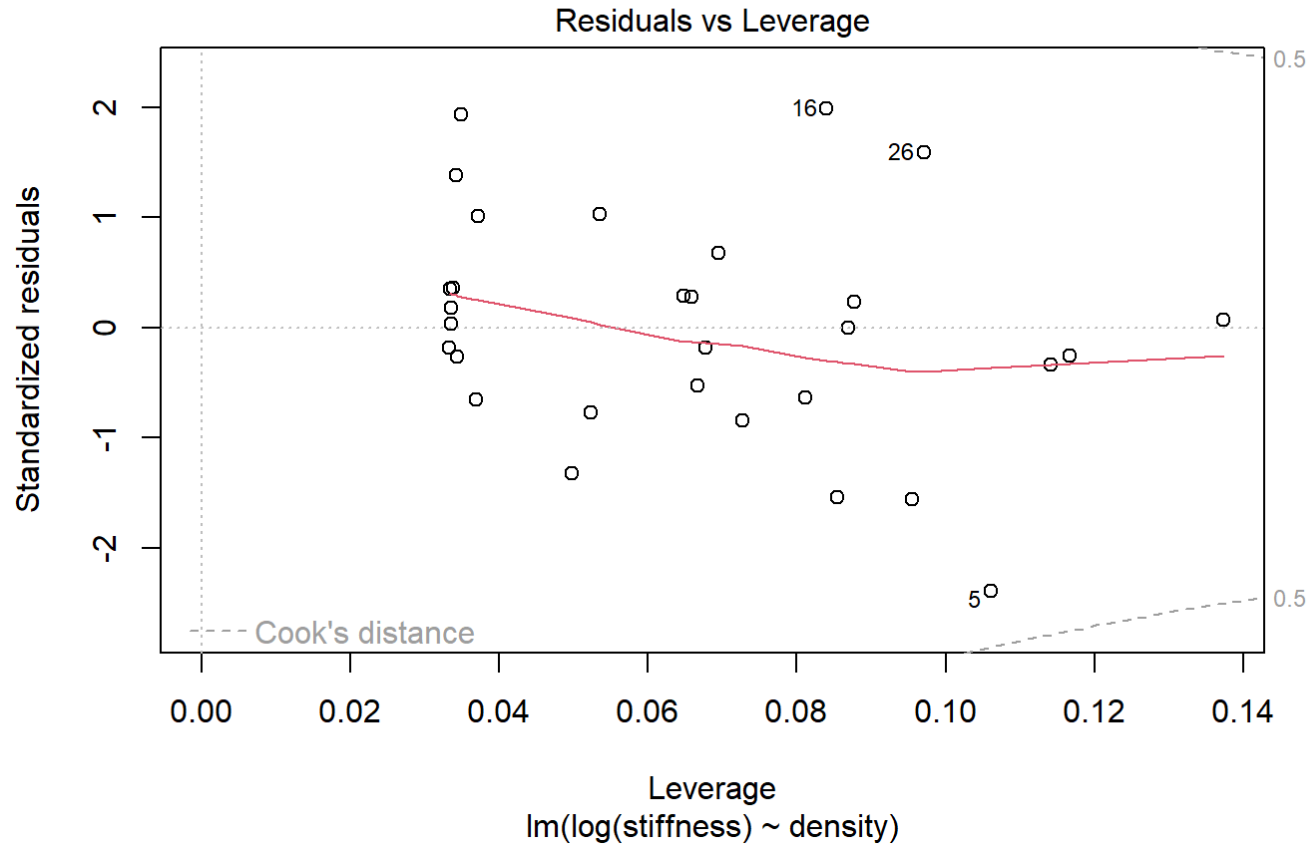
Hide

```
plot(log.model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(log(stiffness) ~ density)
16
8
5
-0.6
-0.4
-0.2
0.0
0.2
0.4
9.0
9.5
10.0
10.5
11.0
11.5


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(stiffness) ~ density)
8
16
5
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
5
16
8
0.0
0.5
1.0
1.5
9.0
9.5
10.0
10.5
11.0
11.5
Fitted values
lm(log(stiffness) ~ density)


Residuals vs Leverage
Standardized residuals
16
26
5
Cook's distance
0.5
0.5
-2
-1
0
1
2
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Leverage
lm(log(stiffness) ~ density)

Hide

```
bptest(log.model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  log.model
## BP = 0.81536, df = 1, p-value = 0.3665
```

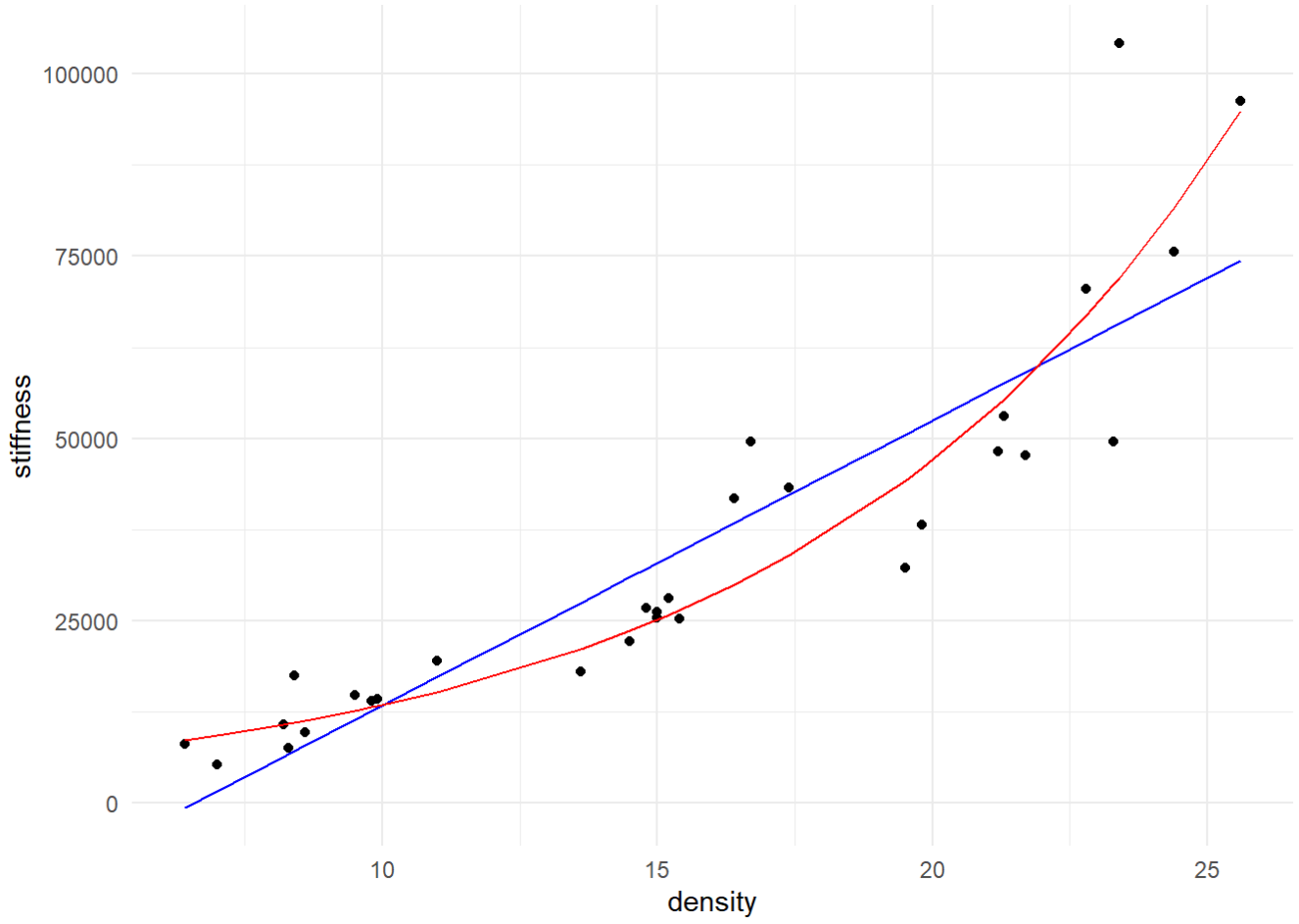
Exercise 6

Converted the fitted values from previous model back to original measurement scale. Plotted the fitted values from the simple linear model and this one as curves superimposed on the original data

Hide

```
stiffness.data %>%
  mutate(., ols.fits = sl.model$fitted.values,
         log.fits=exp(log.model$fitted.values)) -> fits.data

ggplot(data=fits.data, aes(x=density, y=stiffness)) +
  geom_point() +
  geom_line(aes(x=density, y=ols.fits), color="blue") +
  geom_line(aes(x=density, y=log.fits), color="red") +
  theme_minimal()
```



Exercise 7

The predicted stiffness for a 20ln/cft particleboard using the log transformed model is (28140.11, 78655.1). This interval is only slightly larger than that from the simpler model.

Hide

```
exp(predict(log.model, newdata=stiffness20data, interval="prediction"))
```

```
##        fit      lwr     upr
## 1 47046.39 28140.11 78655.1
```

Exercise 8

Yes, this model does give a better fit because of the reduction of randomness and outliers.

Hide

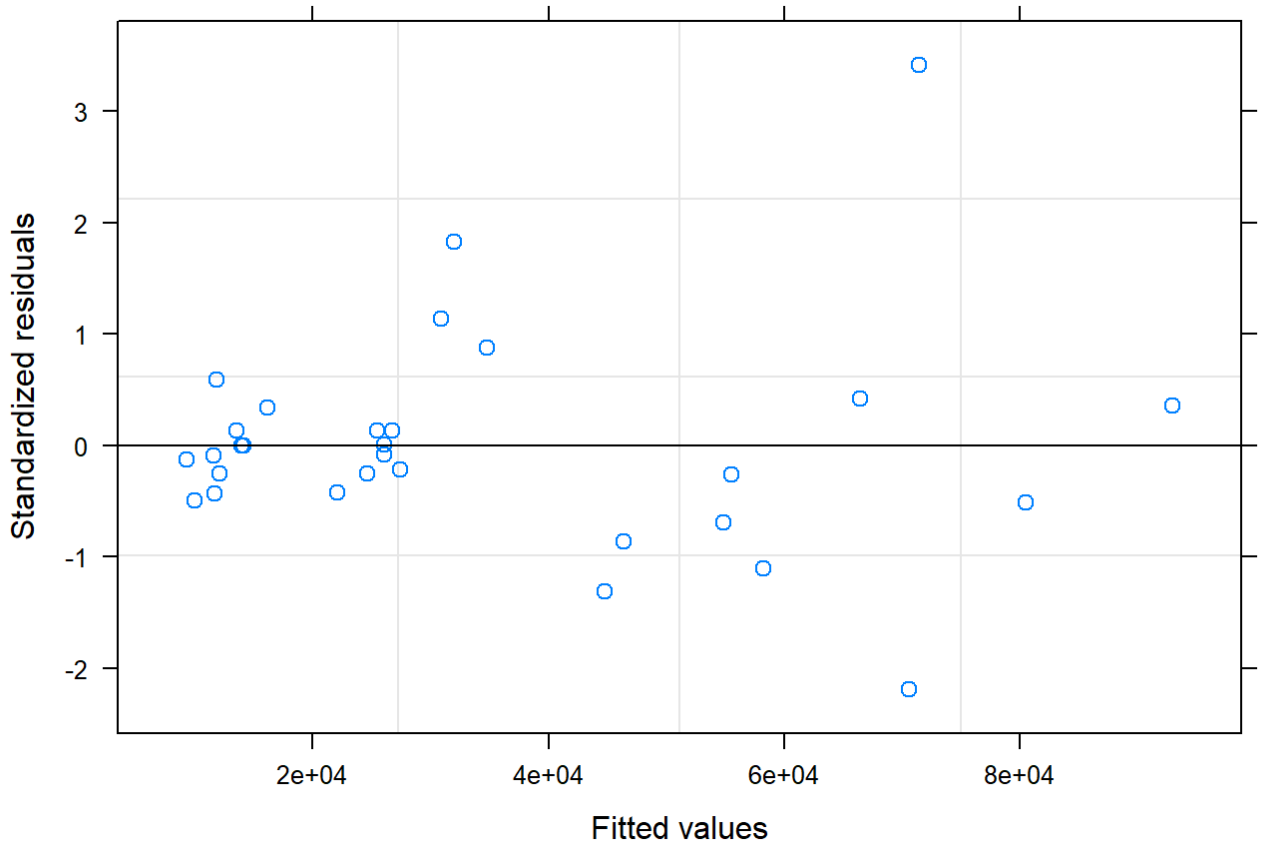
```
A_start <- exp(log.model$coefficients[1])
B_start <- log.model$coefficients[2]

summary(
  nls.model <- nls(stiffness ~ A*exp(B*density),
                   start=list(A=A_start, B=B_start),
                   data=stiffness.data)
)
```

```
##
## Formula: stiffness ~ A * exp(B * density)
##
## Parameters:
##    Estimate Std. Error t value Pr(>|t|)
## A 4336.7420  1036.6955   4.183 0.000257 ***
## B    0.1197     0.0109  10.985 1.16e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9589 on 28 degrees of freedom
##
## Number of iterations to convergence: 2
## Achieved convergence tolerance: 6.662e-06
```

Hide

```
plot(nls.model)
```



Hide

```
predict(nls.model, newdata=stiffness20data, interval="prediction")
```

```
## [1] 47525.3
```

Exercise 9

I would not choose the straight line model, instead I would choose the log or nls model. They are both generally similar in the fit, number of outliers, and constant residual values.

Hide

```
pseudoRsq <- function(Y, Yhat, name="pseudo R-square") {

  n     <- length(stiffness.data$density)
  TSS   <- (n-1)*var(stiffness.data$density)
  SSR   <- sum((stiffness.data$density-stiffness.data$stiffness)^2)
  pRsq  <- 1-(SSR/TSS)
  names(pRsq) <- name
  return(pRsq)
}

model.stats <- data.frame(
  model = c("straight line", "log response", "non-linear LS"),
  CD = signif(
    c(summary(sl.model)$r.squared,
      pseudoRsq(stiffness.data$prog, exp(log.model$fitted.values)),
      pseudoRsq(stiffness.data$prog, predict(nls.model))
      ), digits=3),
  predicted = signif(
    c(predict(sl.model, newdata=stiffness20data, interval="none"),
      exp(predict(log.model, newdata=stiffness20data, interval="none")),
      predict(nls.model, newdata=stiffness20data, interval="none")
      ), digits=3
  )
)

kbl(model.stats, caption="Model Comparison") %>%
  kable_styling(full_width=FALSE)
```

Model Comparison

model	CD	predicted
straight line	7.99e-01	52500
log response	-5.59e+07	47000
non-linear LS	-5.59e+07	47500