Linear regression models in R (session 1)

Tom Price

3 March 2009

Linear regression model

$$y = b_0 + b_1 x_1 + \ldots + b_k x_k + e$$

Where

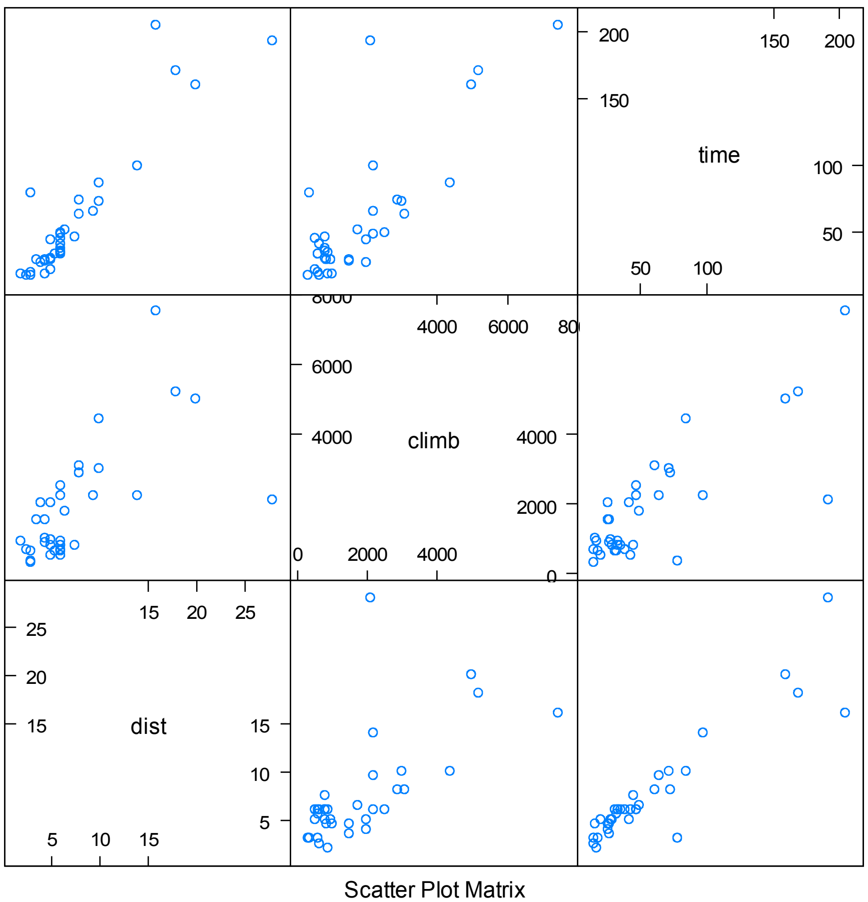
y	is the dependent variable
$x_1 \ldots x_k$	are independent variables (predictors)
$b_0 \ldots b_k$	are the regression coefficients
e	denotes the residuals

- The residuals are assumed to be identically and independently Normally distributed with mean 0.
- The coefficients are usually estimated by the "least squares" technique – choosing values of $b_0 \ldots b_k$ that minimise the sum of the squares of the residuals e.

Scottish hill races

Set of data on record times of Scottish hill races against distance and total height climbed.

```
library(MASS)
?hills
data(hills)
library(lattice)
splom(~hills)
```



Formula

`?formula`

Specifies the model e.g.

`y ~ a`	y is dependent var, a is independent var
`y ~ factor( a )`	dummy coded
`y ~ -1 + factor( a )`	no intercept
`y ~ a + b + c`	3 independent variables
`y ~ a * b + c`	includes one interaction term
`y ~ a * b * c`	includes all interactions terms
`y ~ ( a + b + c )^3`	same as above
`y ~ ( a + b + c )^2`	includes all 2-way interaction terms

Linear model

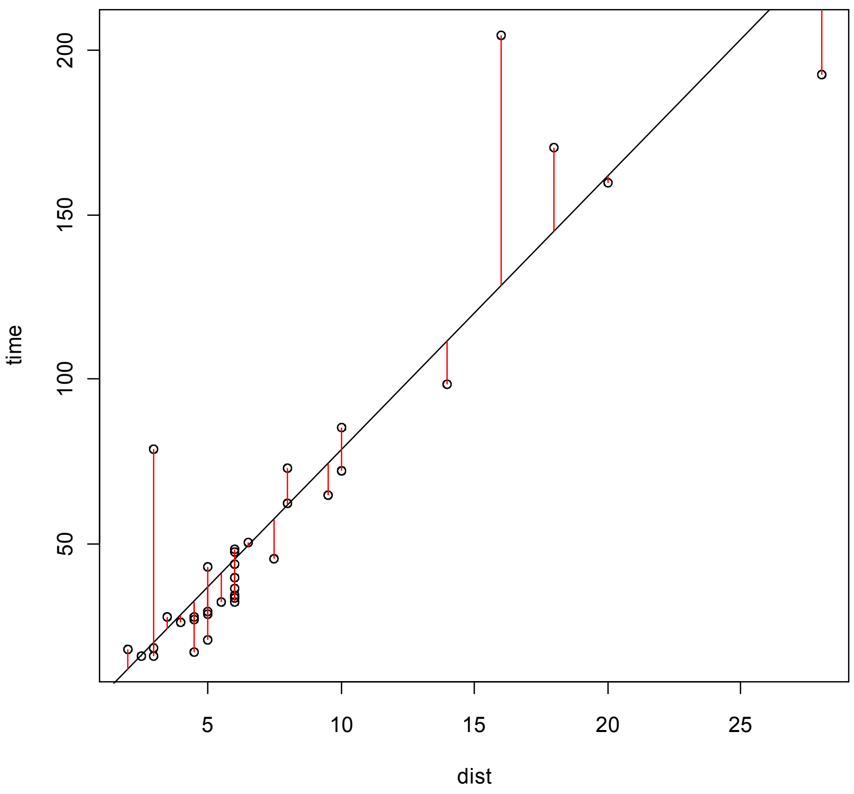
```
?lm
lm1=lm(time~dist,data=hills)
summary(lm1)
```

Linear model

```
lm1=lm(time~dist,data=hills)
summary(lm1)

Call:
lm(formula = time ~ dist)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.745  -9.037  -4.201   2.849  76.170 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -4.8407     5.7562  -0.841    0.406    
dist          8.3305     0.6196  13.446 6.08e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 19.96 on 33 degrees of freedom
Multiple R-squared: 0.8456,     Adjusted R-squared: 0.841 
F-statistic: 180.8 on 1 and 33 DF,  p-value: 6.084e-15
```

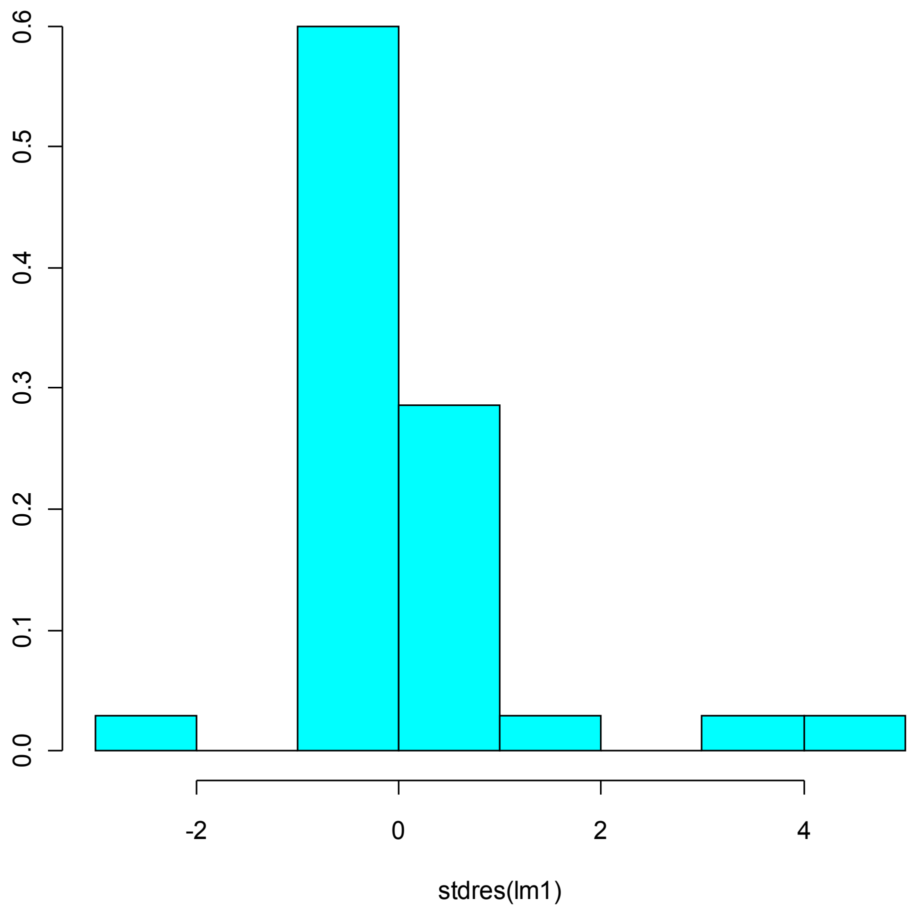
Fitted points

```
attach(hills)
(c1=coef(lm1))

(Intercept)        dist
  -4.840720    8.330456

plot(dist,time)
abline(c1)
f1=fitted(lm1)
for(i in 1:35)
  lines(c(dist[i],dist[i]),
  c(time[i],f1[i]),col="red")
```



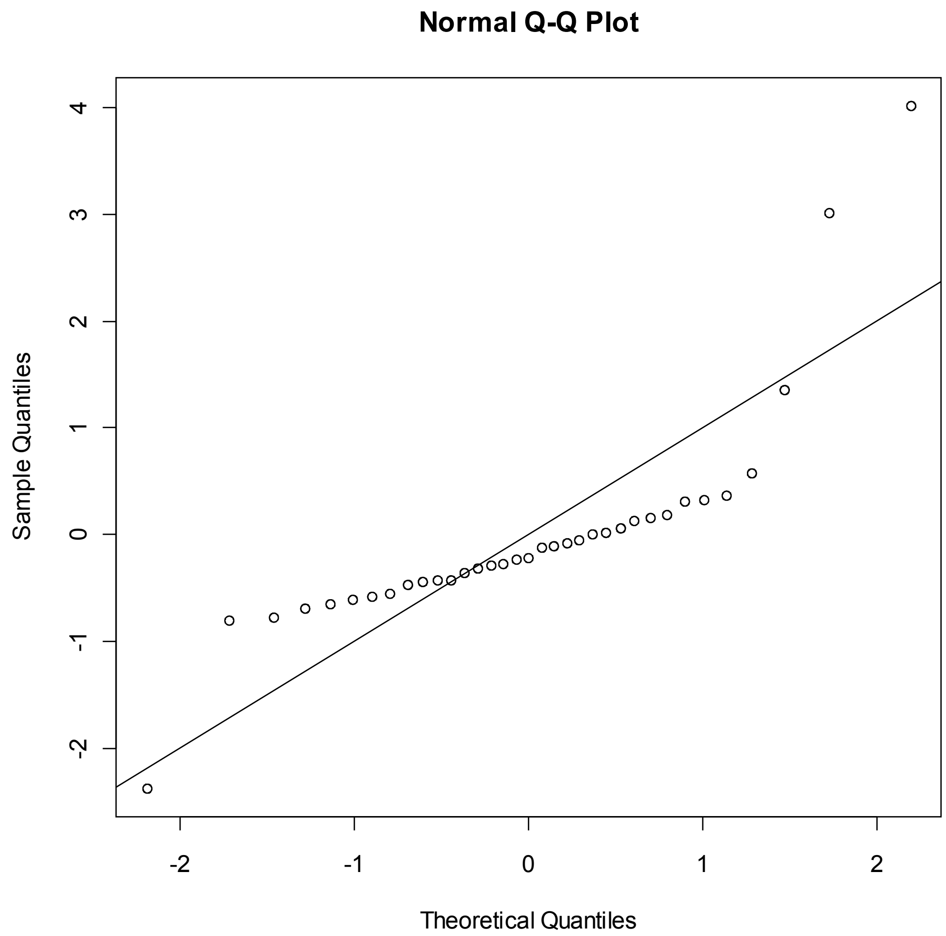
Standardised residuals

```
sr1=stdres(lm1)
truehist(sr1)
```



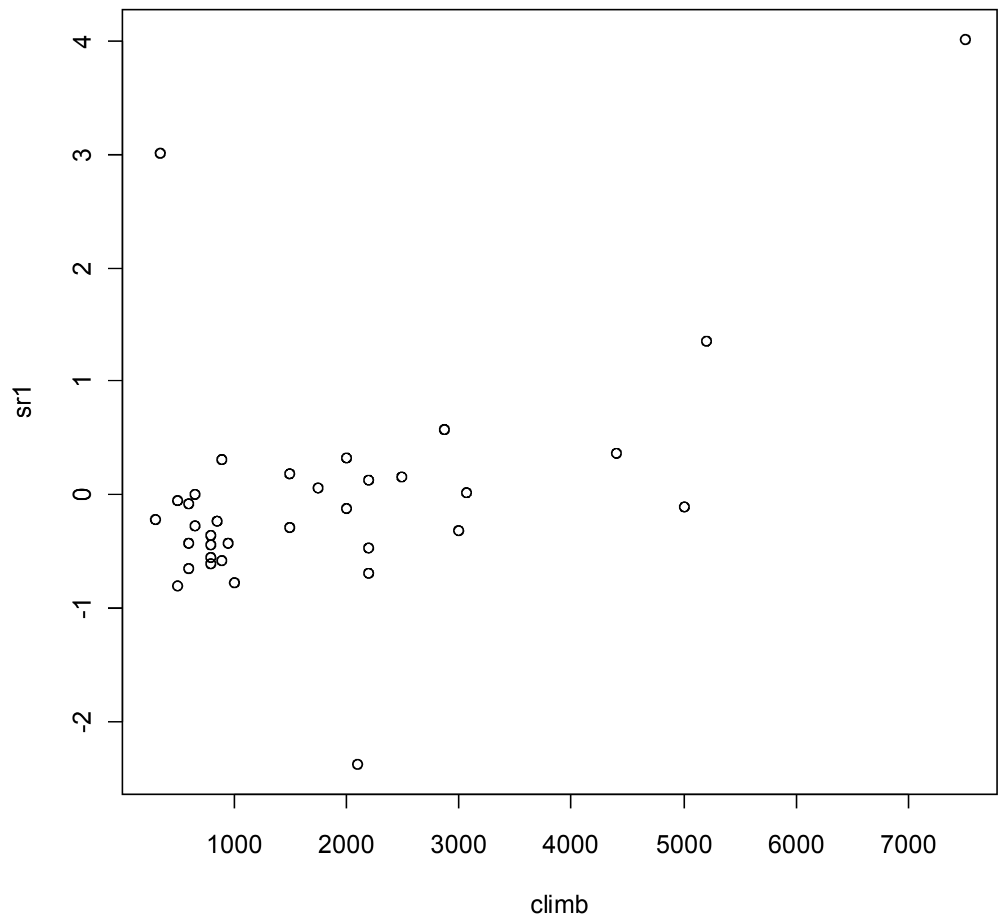
Standardised residuals

```
qqnorm(sr1)
abline(0,1)
```



Standardised residuals

```
plot(climb,sr1)
```



Exercise 1

- Using the Scottish hill races dataset, model the time as a function of both distance and total height climbed.
  - What can you learn from looking at the fitted points and the standardised residuals?

- Some useful commands:

```
library(MASS)
?hills
data(hills)
attach(hills)
?lm
?formula
?fitted
?stdres
?truehist
?plot
?qqnorm
```

Model fit

- The linear model assumes that residuals are independently identically Normally distributed with mean 0.

- To assess whether the model fits the data, look at the residuals.

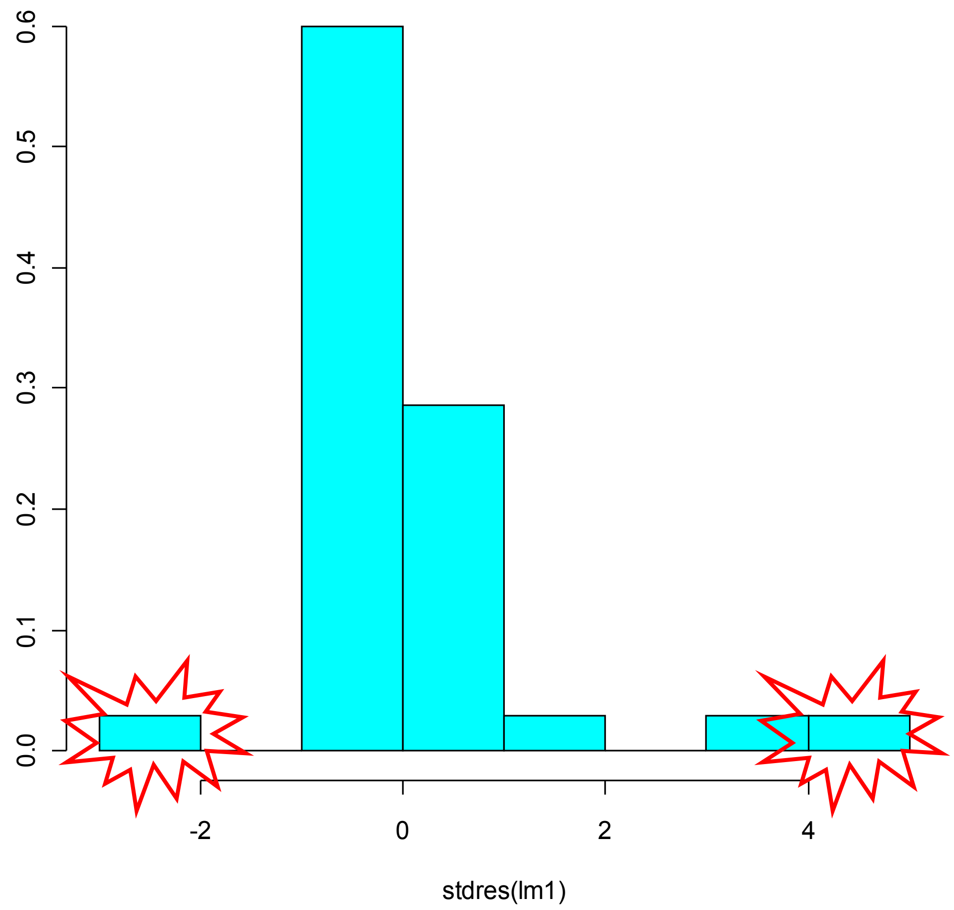
Model fit

- If the model does not fit, it may be because of:
  - Outliers
  - Unmodelled covariates
  - Heteroscedasticity (residuals have unequal variance)
  - Clustering (residuals have lower variance within subgroups)
  - Autocorrelation (correlation between residuals at successive time points)

- All of these can be detected by looking for patterns in the residuals.
- In the next session we will look at some ways to find better fitting models.

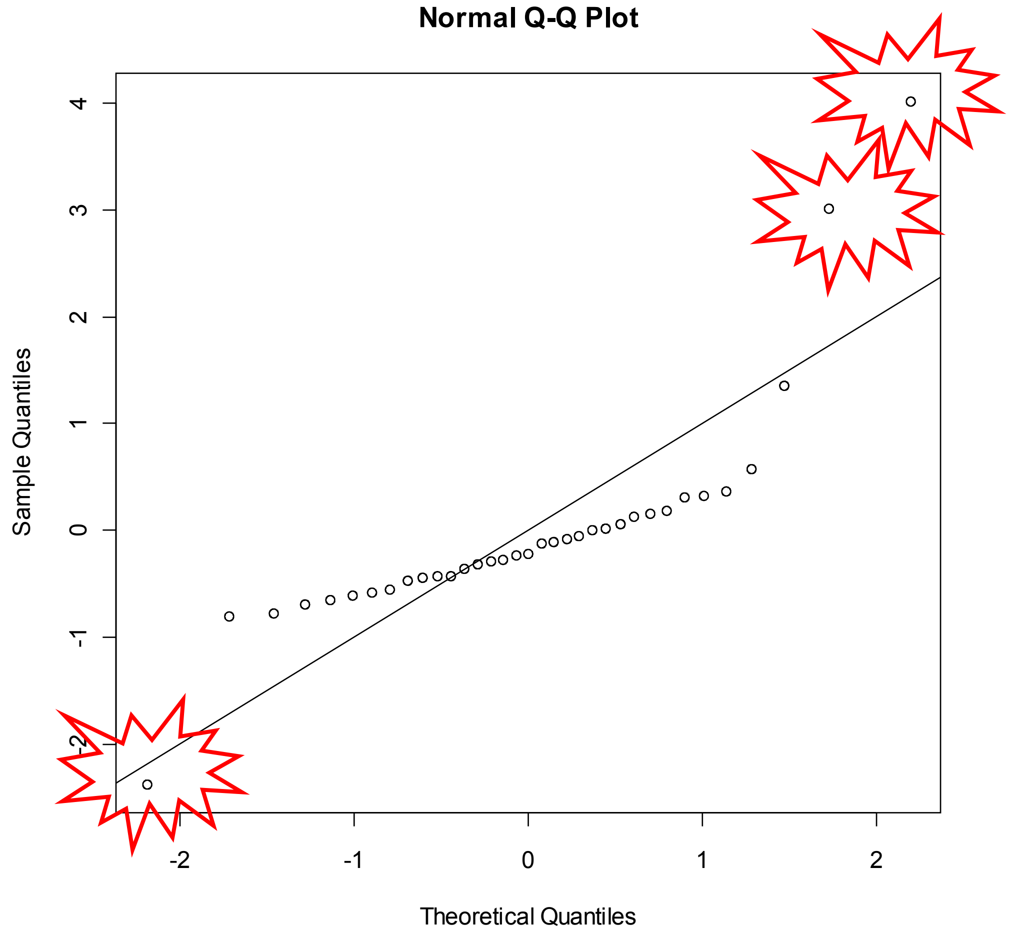
Outliers

```
sr1=stdres(lm1)
truehist(sr1)
```



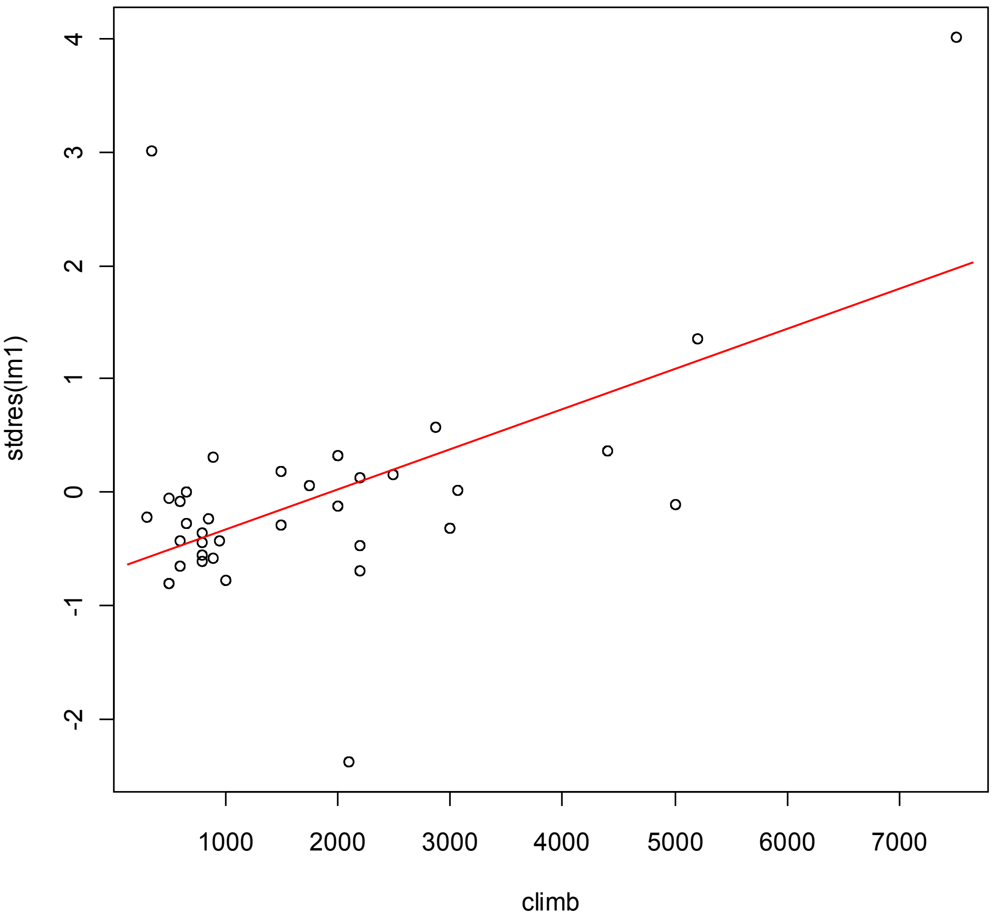
Outliers

```
qqnorm(sr1)
abline(0,1)
```



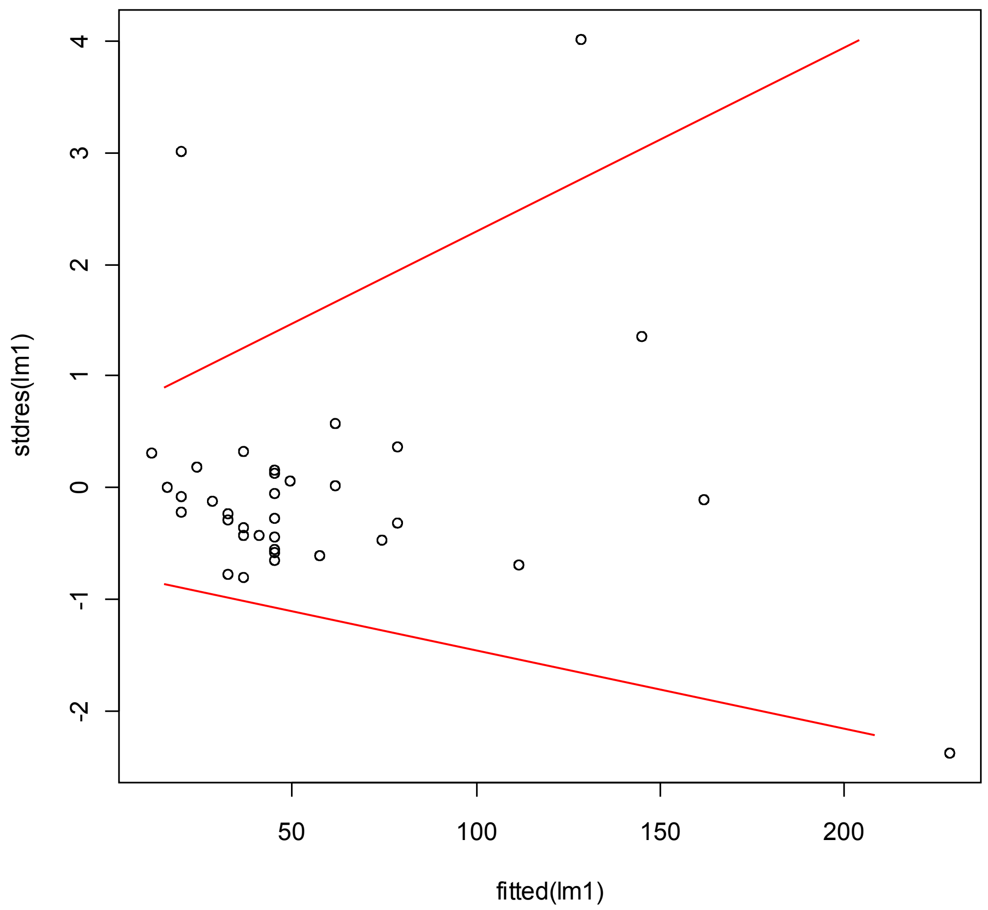
Unmodelled covariate

```
plot(climb,sr1)
```



Heteroscedasticity

```
plot(f1,sr1)
```



Exercise 2

Investigate the dataset "trees" in the MASS package.

- How does Volume depend on Height and Girth? Try some models and examine the residuals to assess model fit.
- Transforming the data can give better fitting models, especially when the residuals are heteroscedastic. Try log and cube root transforms for Volume. Which do you think works better? How do you interpret the results?

Some useful commands:

```
library(MASS)

?trees
?lm
?formula
?stdres
?fitted

?boxcox
```

Reading

Venables & Ripley, “Modern Applied Statistics with S”, chapter 6.