Code

Multiple Regression with R

R. Nicholls / D.-L. Couturier / M. Fernandes

Last modified: 04 Mar 2019

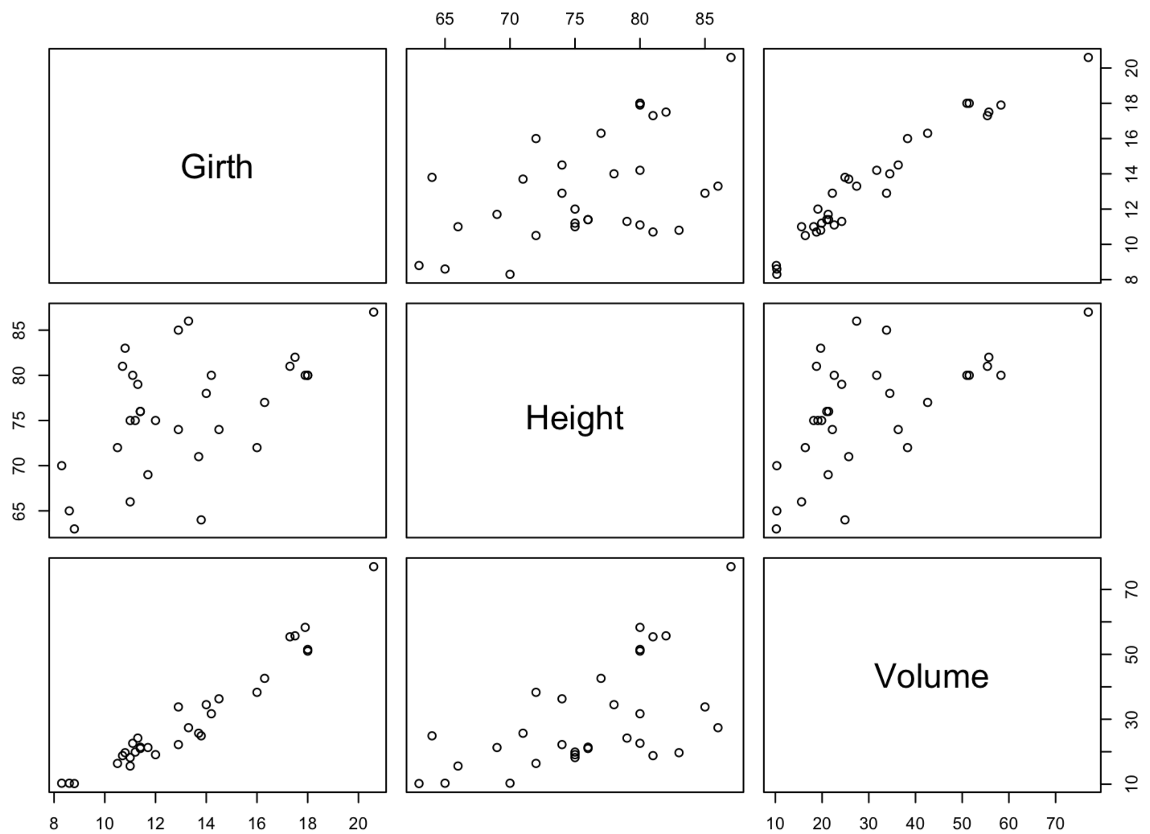
Section 1: Multiple Regression

The in-built dataset `trees` contains data pertaining to the `Volume`, `Girth` and `Height` of 31 felled black cherry trees. In the Simple Regression session, we constructed a simple linear model for `Volume` using `Girth` as the independent variable. Now we will expand this by considering `Height` as another predictor.

Start by plotting the dataset:

Hide

```
plot(trees)
```



This plots all variables against each other, enabling visual information about correlations within the dataset.

Re-create the original model of `Volume` against `Girth`:

Hide

```
m1 = lm(Volume~Girth,data=trees)
summary(m1)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Now include `Height` as an additional variable:

Hide

```
m2 = lm(Volume~Girth+Height,data=trees)
summary(m2)
```

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Note that the R^2 has improved, yet the `Height` term is less significant than the other two parameters.

Try including the interaction term between `Girth` and `Height` :

Hide

```
m3 = lm(Volume~Girth*Height,data=trees)
summary(m3)
```

```
##
## Call:
## lm(formula = Volume ~ Girth * Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5821 -1.0673  0.3026  1.5641  4.6649
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.39632   23.83575   2.911  0.00713 **
## Girth        -5.85585    1.92134  -3.048  0.00511 **
## Height       -1.29708    0.30984  -4.186  0.00027 ***
## Girth:Height  0.13465    0.02438   5.524 7.48e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.709 on 27 degrees of freedom
## Multiple R-squared:  0.9756, Adjusted R-squared:  0.9728
## F-statistic: 359.3 on 3 and 27 DF,  p-value: < 2.2e-16
```

All terms are highly significant. Note that the `Height` is more significant than in the previous model, despite the introduction of an additional parameter.

We'll now try a different functional form - rather than looking for an additive model, we can explore a multiplicative model by applying a log-log transformation (leaving out the interaction term for now).

Hide

```
m4 = lm(log(Volume)~log(Girth)+log(Height),data=trees)
summary(m4)
```

```
##
## Call:
## lm(formula = log(Volume) ~ log(Girth) + log(Height), data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168561 -0.048488  0.002431  0.063637  0.129223
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
## log(Girth)   1.98265    0.07501  26.432  < 2e-16 ***
## log(Height)  1.11712    0.20444   5.464 7.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08139 on 28 degrees of freedom
## Multiple R-squared:  0.9777, Adjusted R-squared:  0.9761
## F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

All terms are significant. Note that the residual standard error is much lower than for the previous models. However, this value cannot be compared with the previous models due to transforming the response variable. The R^2 value has increased further, despite reducing the number of parameters from four to three.

Hide

```
confint(m4)
```

```
##                 2.5 %    97.5 %
## (Intercept) -8.269912 -4.993322
## log(Girth)   1.828998  2.136302
## log(Height)  0.698353  1.535894
```

Looking at the confidence intervals for the parameters reveals that the estimated power of `Girth` is around 2, and `Height` around 1. This makes a lot of sense, given the well-known dimensional relationship between `Volume`, `Girth` and `Height`!

For completeness, we'll now add the interaction term.

Hide

```
m5 = lm(log(Volume)~log(Girth)*log(Height),data=trees)
summary(m5)
```

```
##
## Call:
## lm(formula = log(Volume) ~ log(Girth) * log(Height), data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.165941 -0.048613  0.006384  0.062204  0.132295
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.6869     7.6996  -0.479    0.636
## log(Girth)               0.7942     3.0910   0.257    0.799
## log(Height)              0.4377     1.7788   0.246    0.808
## log(Girth):log(Height)   0.2740     0.7124   0.385    0.704
##
## Residual standard error: 0.08265 on 27 degrees of freedom
## Multiple R-squared:  0.9778, Adjusted R-squared:  0.9753
## F-statistic: 396.4 on 3 and 27 DF,  p-value: < 2.2e-16
```

The R^2 value has increased (of course, as all we've done is add an additional parameter), but interestingly none of the four terms are significant. This means that none of the individual terms alone are vital for the model - there is duplication of information between the variables. So we will revert back to the previous model.

Given that it would be reasonable to expect the power of `Girth` to be 2, and Height to be 1, we will now fix those parameters, and instead just estimate the one remaining parameter.

Hide

```
m6 = lm(log(Volume)-log((Girth^2)*Height)~1,data=trees)
summary(m6)
```

```
##
## Call:
## lm(formula = log(Volume) - log((Girth^2) * Height) ~ 1, data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168446 -0.047355 -0.003518  0.066308  0.136467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.16917    0.01421  -434.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0791 on 30 degrees of freedom
```

Note that there is no R^2 (as only the intercept was included in the model), and that the Residual Standard Error is incomparable with previous models due to changing the response variable.

We can alternatively construct a model with the response being y, and the error term additive rather than multiplicative.

Hide

```
m7 = lm(Volume~0+I(Girth^2):Height,data=trees)
summary(m7)
```

```
##
## Call:
## lm(formula = Volume ~ 0 + I(Girth^2):Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6696 -1.0832 -0.3341  1.6045  4.2944
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## I(Girth^2):Height 2.108e-03  2.722e-05   77.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.455 on 30 degrees of freedom
## Multiple R-squared:  0.995,  Adjusted R-squared:  0.9949
## F-statistic:  5996 on 1 and 30 DF,  p-value: < 2.2e-16
```

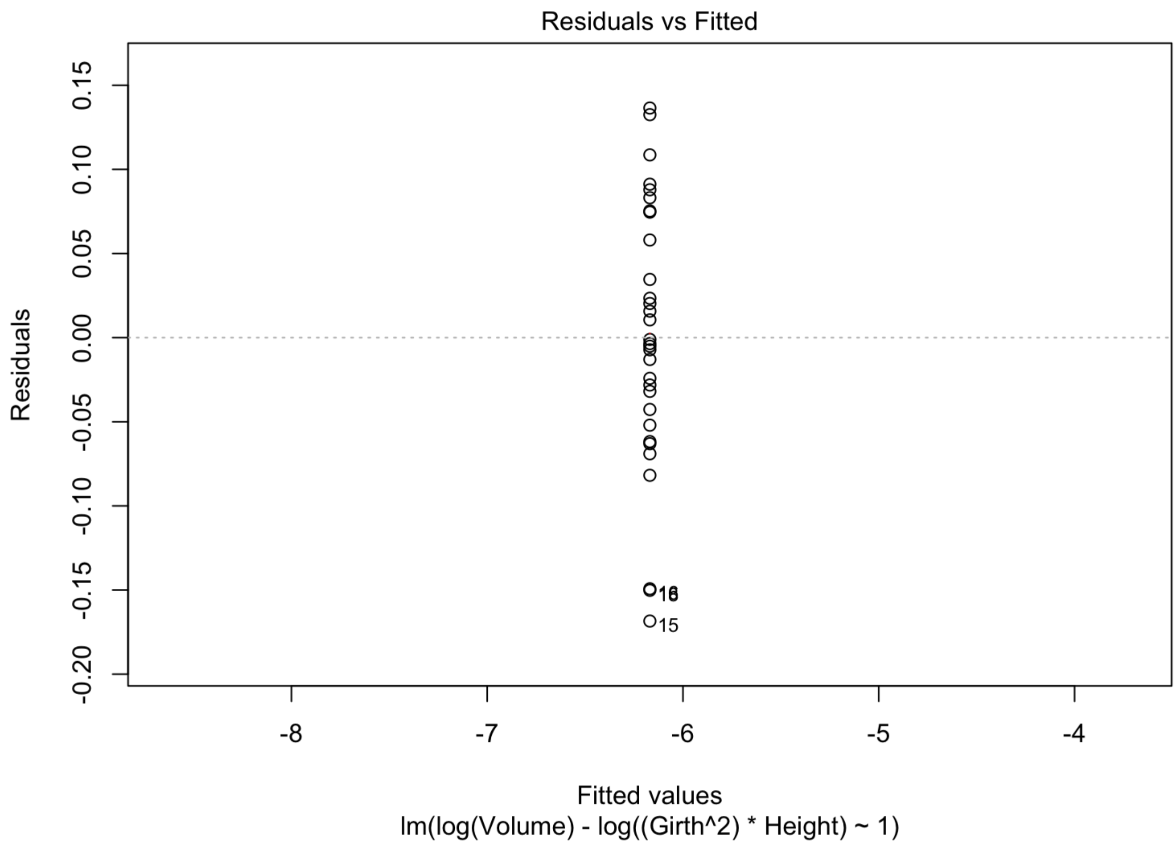
Note that the parameter estimates for the last two models are slightly different… this is due to differences in the error model.

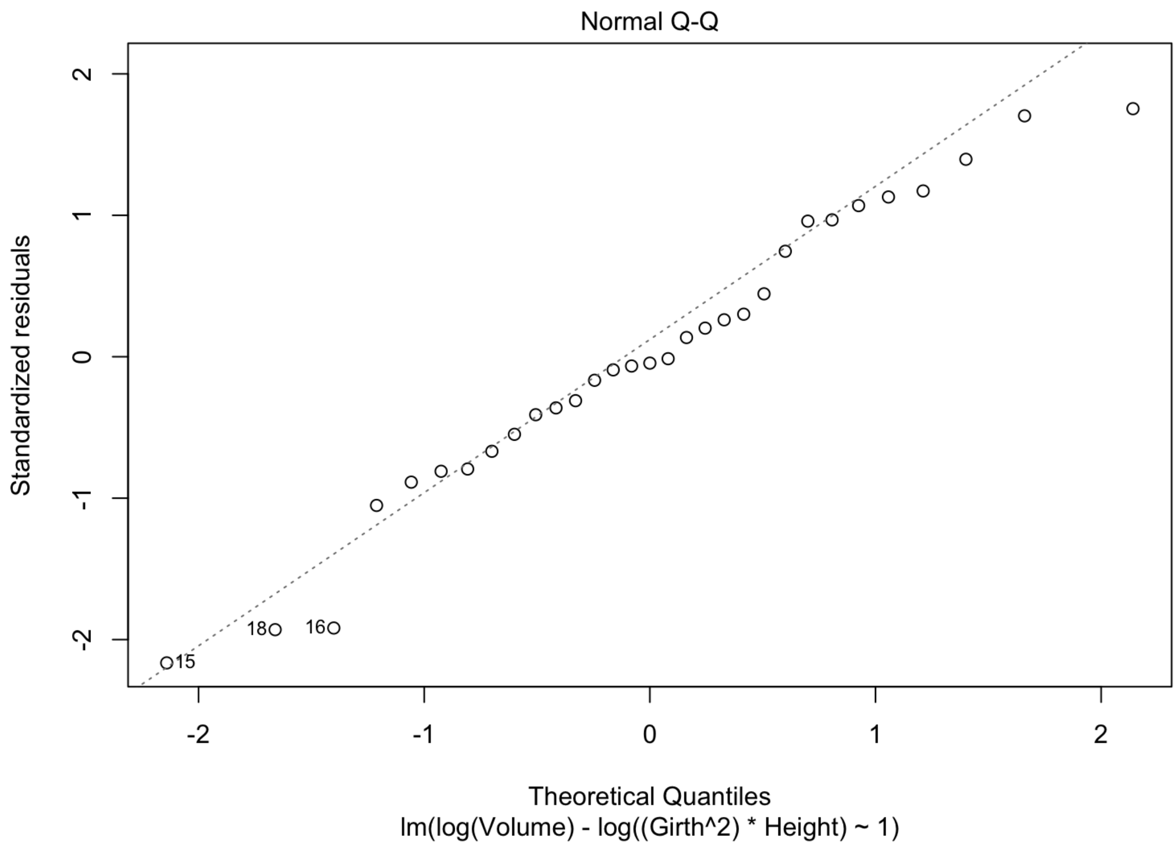
Section 2: Model Selection

Of the last two models, the one with the log-Normal error model would seem to have the more Normal residuals. This can be inspected by looking at diagnostic plots, by and using the `shapiro.test()`:

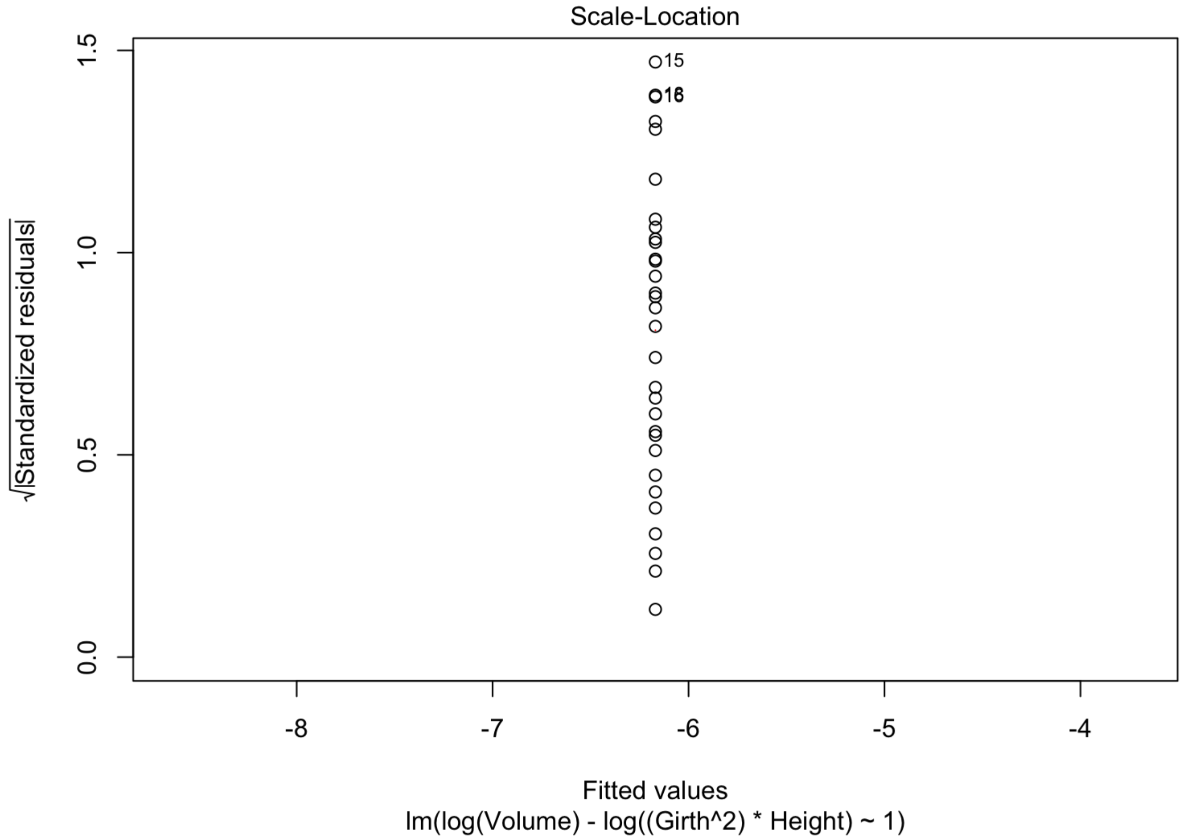
Hide

```
plot(m6)
```



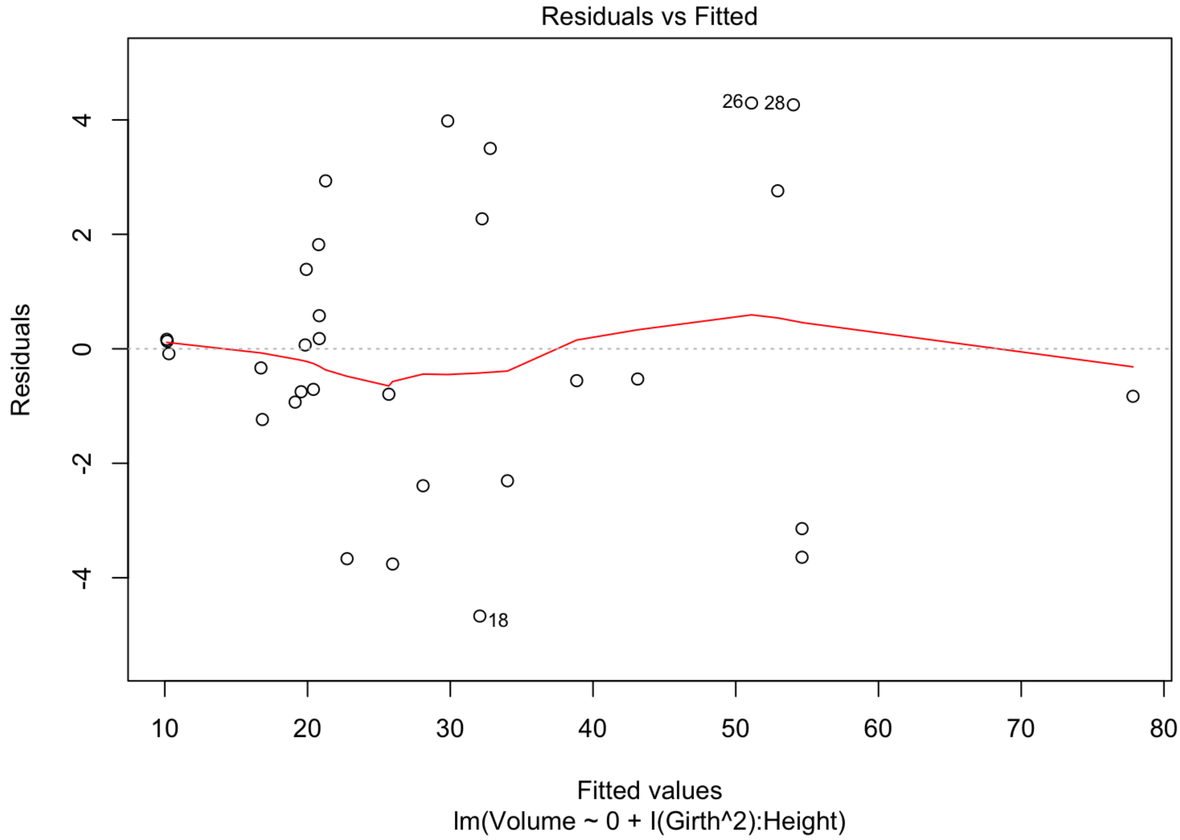


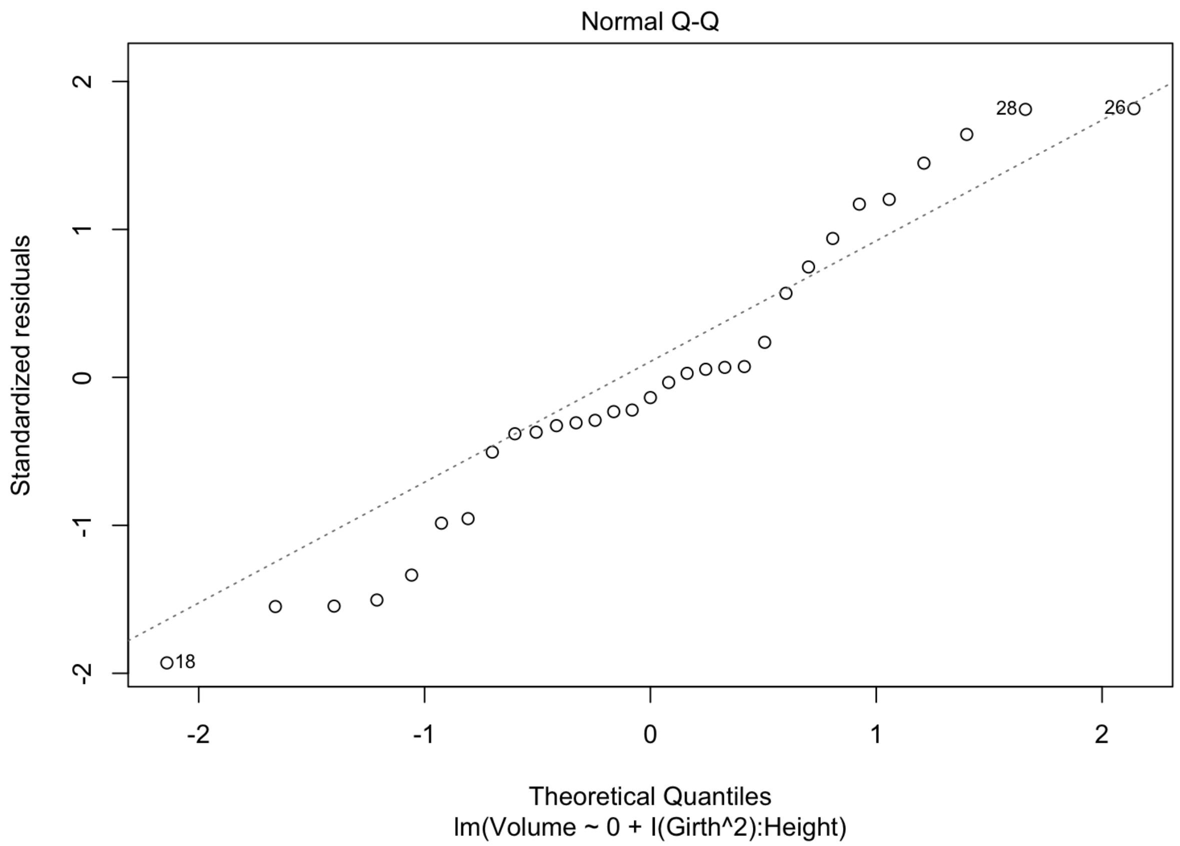

```
## hat values (leverages) are all = 0.03225806
##  and there are no factor predictors; no plot no. 5
```

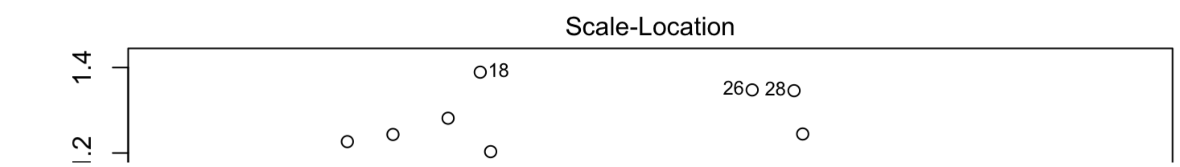

Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
15
16
-8
-7
-6
-5
-4
Fitted values
lm(log(Volume) - log((Girth^2) * Height) ~ 1)

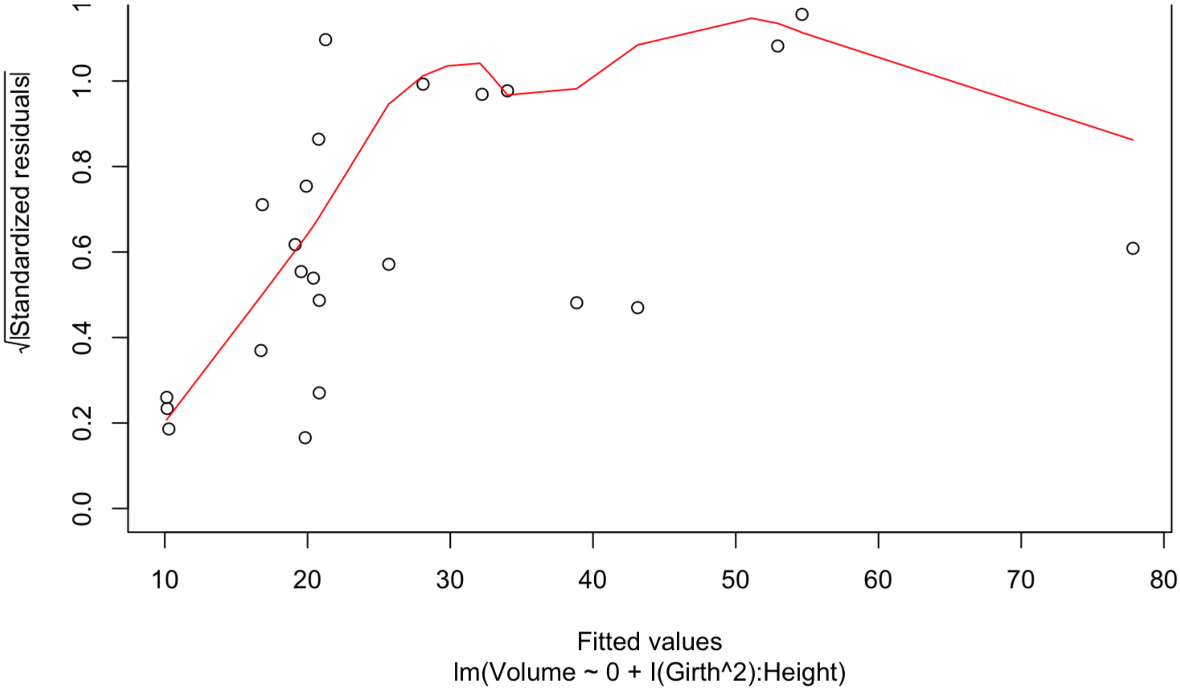
Hide

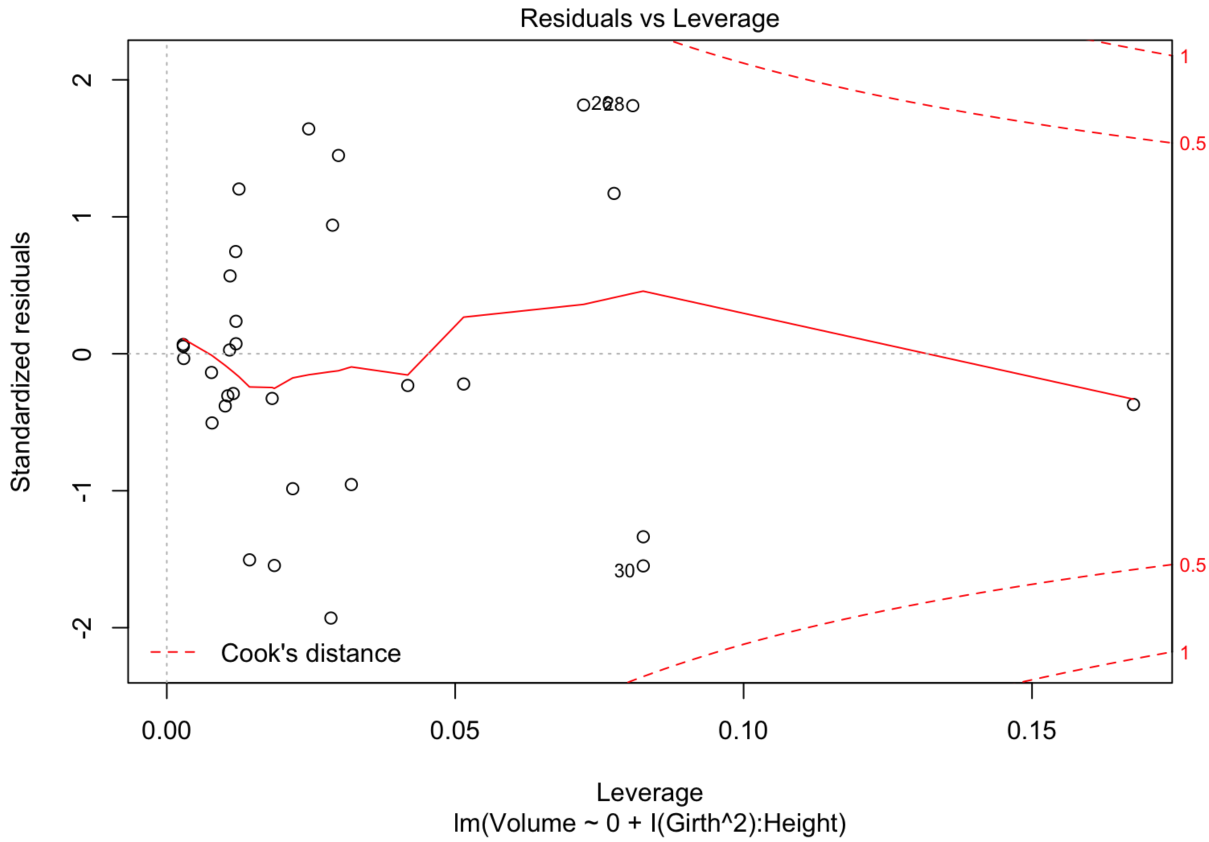
```
plot(m7)
```


Residuals vs Fitted
Residuals
-4
-2
0
2
4
26
28
18
10
20
30
40
50
60
70
80
Fitted values
lm(Volume ~ 0 + I(Girth^2):Height)


Normal Q-Q
Standardized residuals
-2
-1
0
1
2
28
26
18
-2
-1
0
1
2
Theoretical Quantiles
lm(Volume ~ 0 + I(Girth^2):Height)


Scale-Location
1.4
1.2
18
26
28





Hide

```
shapiro.test(residuals(m6))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(m6)
## W = 0.97013, p-value = 0.5225
```

Hide

```
shapiro.test(residuals(m7))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(m7)
## W = 0.95846, p-value = 0.2655
```

The Akaike Information Criterion (AIC) can help to make decisions regarding which model is the most appropriate. Now calculate the AIC for each of the above models:

Hide

```
summary(m1)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Hide

```
AIC(m1)
```

```
## [1] 181.6447
```

Hide

```
summary(m2)
```

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Hide

```
AIC(m2)
```

```
## [1] 176.91
```

Hide

```
summary(m3)
```

```
##
## Call:
## lm(formula = Volume ~ Girth * Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5821 -1.0673  0.3026  1.5641  4.6649
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.39632   23.83575   2.911  0.00713 **
## Girth        -5.85585    1.92134  -3.048  0.00511 **
## Height       -1.29708    0.30984  -4.186  0.00027 ***
## Girth:Height  0.13465    0.02438   5.524 7.48e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.709 on 27 degrees of freedom
## Multiple R-squared:  0.9756, Adjusted R-squared:  0.9728
## F-statistic: 359.3 on 3 and 27 DF,  p-value: < 2.2e-16
```

Hide

```
AIC(m3)
```

```
## [1] 155.4692
```

Hide

```
summary(m4)
```

```
##
## Call:
## lm(formula = log(Volume) ~ log(Girth) + log(Height), data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168561 -0.048488  0.002431  0.063637  0.129223
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
## log(Girth)   1.98265    0.07501  26.432  < 2e-16 ***
## log(Height)  1.11712    0.20444   5.464 7.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08139 on 28 degrees of freedom
## Multiple R-squared:  0.9777, Adjusted R-squared:  0.9761
## F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

Hide

```
AIC(m4)
```

```
## [1] -62.71125
```

Hide

```
summary(m5)
```

```
##
## Call:
## lm(formula = log(Volume) ~ log(Girth) * log(Height), data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.165941 -0.048613  0.006384  0.062204  0.132295
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.6869     7.6996  -0.479    0.636
## log(Girth)               0.7942     3.0910   0.257    0.799
## log(Height)              0.4377     1.7788   0.246    0.808
## log(Girth):log(Height)   0.2740     0.7124   0.385    0.704
##
## Residual standard error: 0.08265 on 27 degrees of freedom
## Multiple R-squared:  0.9778, Adjusted R-squared:  0.9753
## F-statistic: 396.4 on 3 and 27 DF,  p-value: < 2.2e-16
```

Hide

```
AIC(m5)
```

```
## [1] -60.88061
```

Hide

```
summary(m6)
```

```
##
## Call:
## lm(formula = log(Volume) - log((Girth^2) * Height) ~ 1, data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168446 -0.047355 -0.003518  0.066308  0.136467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.16917    0.01421  -434.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0791 on 30 degrees of freedom
```

Hide

```
AIC(m6)
```

```
## [1] -66.34198
```

Hide

```
summary(m7)
```

```
##
## Call:
## lm(formula = Volume ~ 0 + I(Girth^2):Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6696 -1.0832 -0.3341  1.6045  4.2944
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## I(Girth^2):Height 2.108e-03  2.722e-05   77.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.455 on 30 degrees of freedom
## Multiple R-squared:  0.995,  Adjusted R-squared:  0.9949
## F-statistic:  5996 on 1 and 30 DF,  p-value: < 2.2e-16
```

Hide

```
AIC(m7)
```

```
## [1] 146.6447
```

Whilst the AIC can help differentiate between similar models, it cannot help deciding between models that have different responses. Which model would you select as the most appropriate?

Section 3: Stepwise Regression

The in-built dataset `swiss` contains data pertaining to fertility, along with a variety of socioeconomic indicators. We want to select a sensible model using stepwise regression. First regress `Fertility` agains all available indicators:

Hide

```
m8 = lm(Fertility~.,data=swiss)
summary(m8)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

Are all terms significant?

Now use stepwise regression, performing backward elimination in order to automatically remove inappropriate terms:

Hide

```
library(MASS)
summary(stepAIC(m8))
```

```
## Start:  AIC=190.69
## Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## - Examination       1     53.03 2158.1 189.86
## <none>                          2105.0 190.69
## - Agriculture       1    307.72 2412.8 195.10
## - Infant.Mortality  1    408.75 2513.8 197.03
## - Catholic          1    447.71 2552.8 197.75
## - Education         1   1162.56 3267.6 209.36
##
## Step:  AIC=189.86
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## <none>                          2158.1 189.86
## - Agriculture       1    264.18 2422.2 193.29
## - Infant.Mortality  1    409.81 2567.9 196.03
## - Catholic          1    956.57 3114.6 205.10
## - Education         1   2249.97 4408.0 221.43
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6765  -6.0522   0.7514   3.1664  16.1422
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
## Agriculture      -0.15462    0.06819  -2.267  0.02857 *
## Education        -0.98026    0.14814  -6.617 5.14e-08 ***
## Catholic          0.12467    0.02889   4.315 9.50e-05 ***
## Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.168 on 42 degrees of freedom
## Multiple R-squared:  0.6993, Adjusted R-squared:  0.6707
## F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

Are all terms significant? Is this model suitable? What are the pro's and con's of this approach?

Section 4: Non-Linear Models

The in-built dataset `trees` contains data pertaining to the `Volume`, `Girth` and `Height` of 31 felled black cherry trees.

In the Simple Regression session, we constructed a simple linear model for `Volume` using `Girth` as the independent variable. Using Multiple Regression we trialled various models, including some that had multiple predictor variables and/or involved log-log transformations to explore power relationships.

However, due to limitations of the method, we were not able to explore other options such as a parameterised power relationship with an additive error model. We will now attempt to fit this model:

[Math Processing Error]

Parameters for non-linear models may be estimated using the `nls` package in R.

Hide

```
volume = trees$Volume
height = trees$Height
girth = trees$Girth
m9 = nls(volume~beta0*girth^beta1*height^beta2,start=list(beta0=1,beta1=2,beta2=1))
summary(m9)
```

```
##
## Formula: volume ~ beta0 * girth^beta1 * height^beta2
##
## Parameters:
##       Estimate Std. Error t value Pr(>|t|)
## beta0 0.001449   0.001367   1.060 0.298264
## beta1 1.996921   0.082077  24.330  < 2e-16 ***
## beta2 1.087647   0.242159   4.491 0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.533 on 28 degrees of freedom
##
## Number of iterations to convergence: 5
## Achieved convergence tolerance: 8.255e-07
```

Note that the parameters `beta0`, `beta1` and `beta2` weren't defined prior to the function call - `nls` knew what to do with them. Also note that we had to provide starting points for the parameters. What happens if you change them?

Are all terms significant? Is this model appropriate? What else could be tried to achieve a better model?

Section 5: Practical Exercises

Puromycin

The in-built R dataset `Puromycin` contains data regarding the reaction velocity versus substrate concentration in an enzymatic reaction involving untreated cells or cells treated with Puromycin.

- Plot `conc` (concentration) against `rate`. What is the nature of the relationship between `conc` and `rate`?
- Find a transformation that linearises the data and stabilises the variance, making it possible to use linear regression. Create the corresponding linear regression model. Are all terms significant?
- Add the `state` term to the model. What type of variable is this? Is the inclusion of this term appropriate?

- Now add a term representing the interaction between `rate` and `state`. Are all terms significant? What can you conclude?
- Given this information, create the regression model you believe to be the most appropriate for modelling `conc`. Regenerate the plot of `conc` against `rate`. Draw curves corresponding to the fitted values of the final model onto this plot (note that two separate curves should be drawn, corresponding to the two levels of `state`).

Attitude

The in-built R dataset `attitude` contains data from a survey of clerical employees.

- Create a linear model regressing `rating` on `complaints`, and store the model in a variable.
- Use the step function to perform forward selection stepwise regression, in order to automatically add appropriate terms, using a command similar to:
  `new_model = step(original_model,.~.+privileges+learning+raises+critical+advance)`
- Which term(s) were added? What is Akaike's Information Criterion (AIC) corresponding to the final model? Are all terms in the resulting model significant? Check diagnostic plots. Do you think this is a suitable model?