Linear regression models in R (session 2)

Tom Price

10 March 2009

Your regression model is wrong

"All models are wrong,
but some are useful"

- How wrong is your model?
- How useful is your model?

Sir George Box (b. 1919)

Exercise 2

Investigate the dataset “trees” in the MASS package.

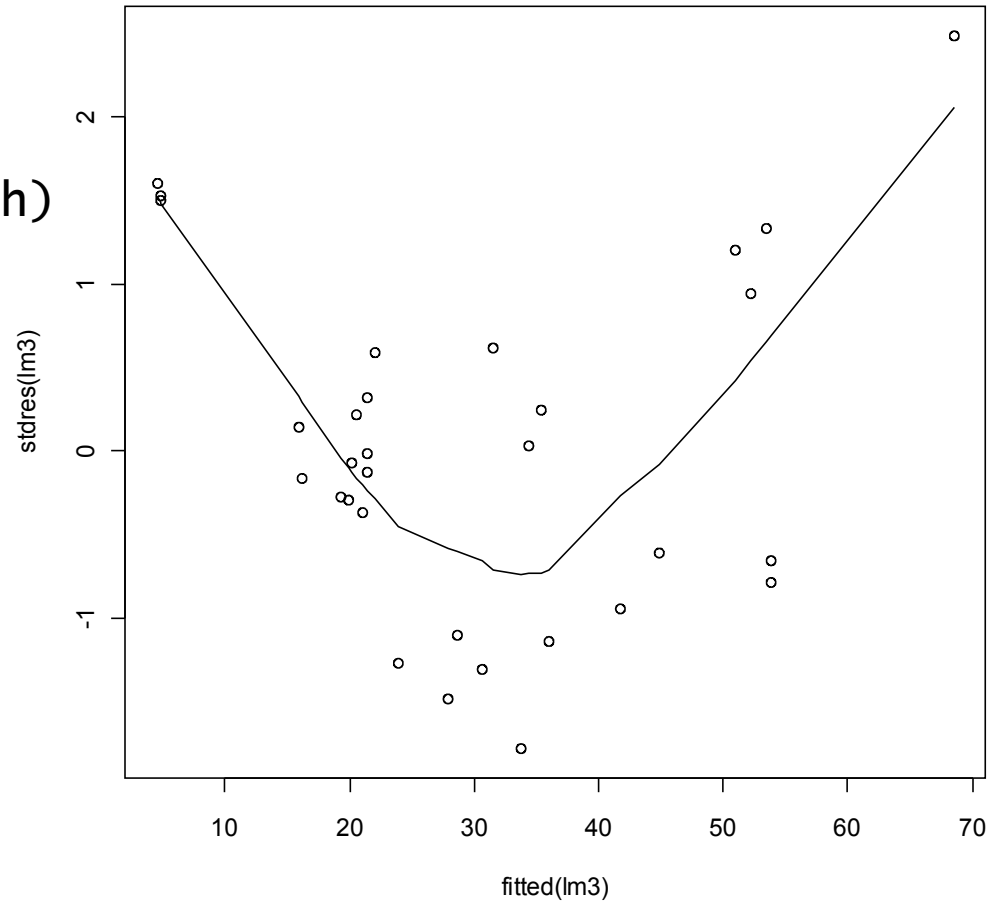
- How does Volume depend on Height and Girth? Try some models and examine the residuals to assess model fit.
- Transforming the data can give better fitting models, especially when the residuals are heteroscedastic. Try log and cube root transforms for Volume. Which do you think works better? How do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

Some useful commands:

```
library(MASS)

?trees
?lm
?formula
?stdres
?fitted

?boxcox
```

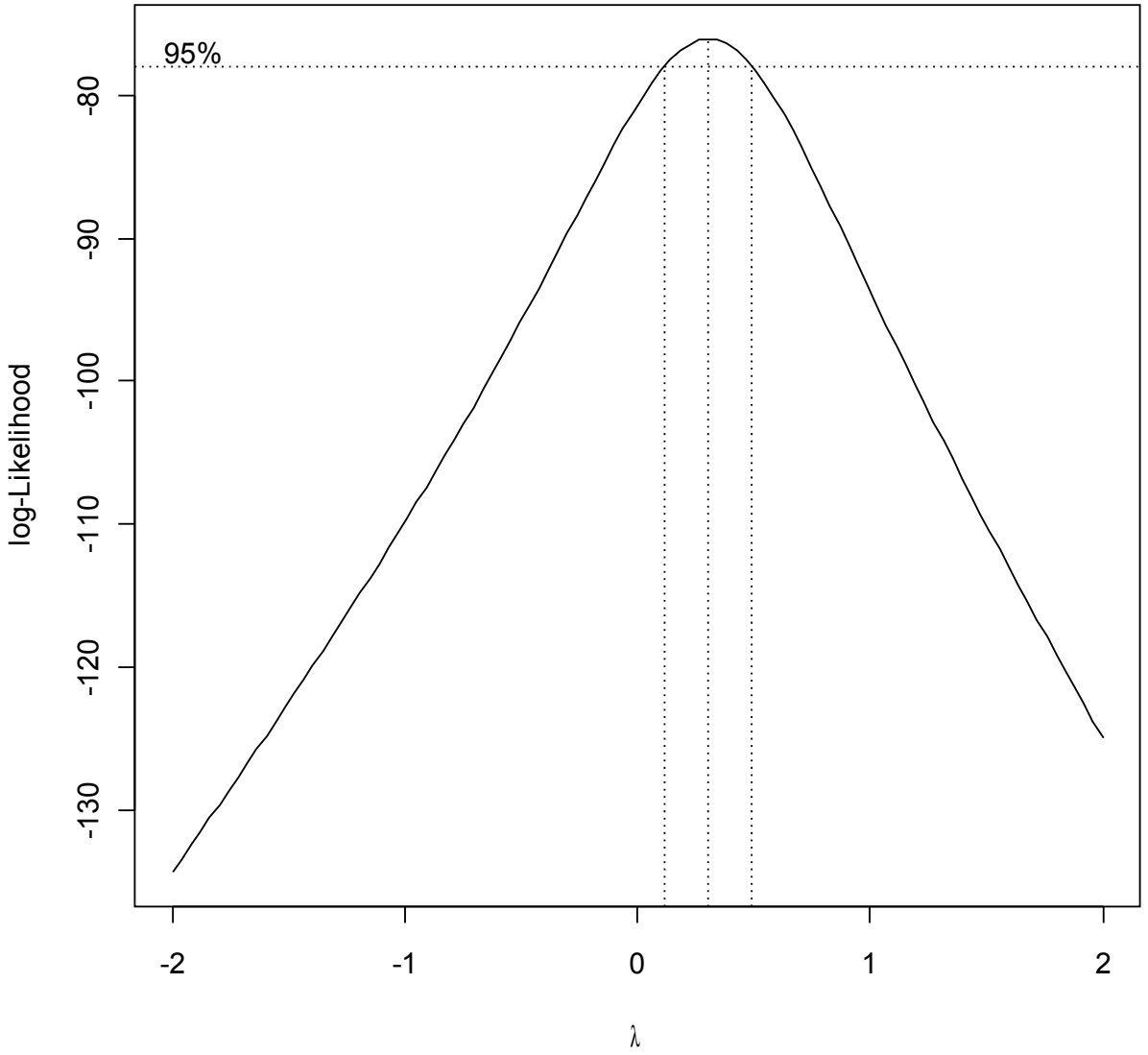
trees

```
library(MASS)
data(trees)
attach(trees)
lm3=lm(Volume~Height+Girth)
sr3=stdres(lm3)
f3=fitted(lm3)
plot(f3,sr3)
lines(lowess(f3,sr3))
```



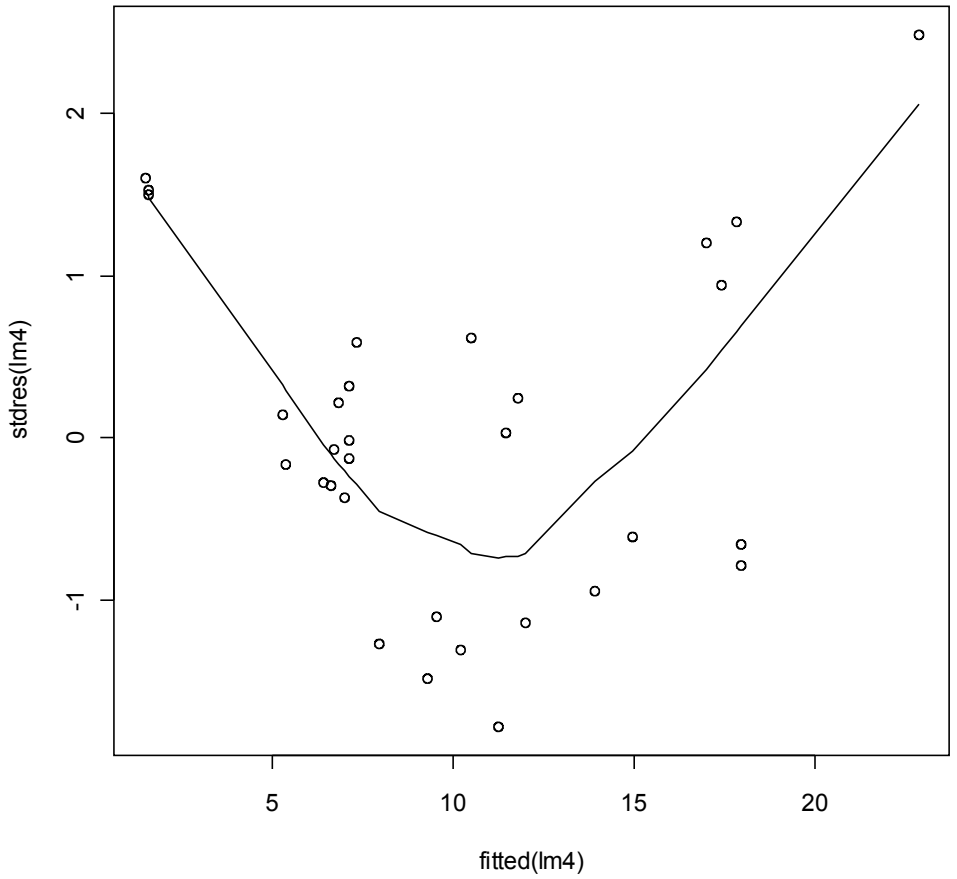
trees

```
boxcox(lm3)
```



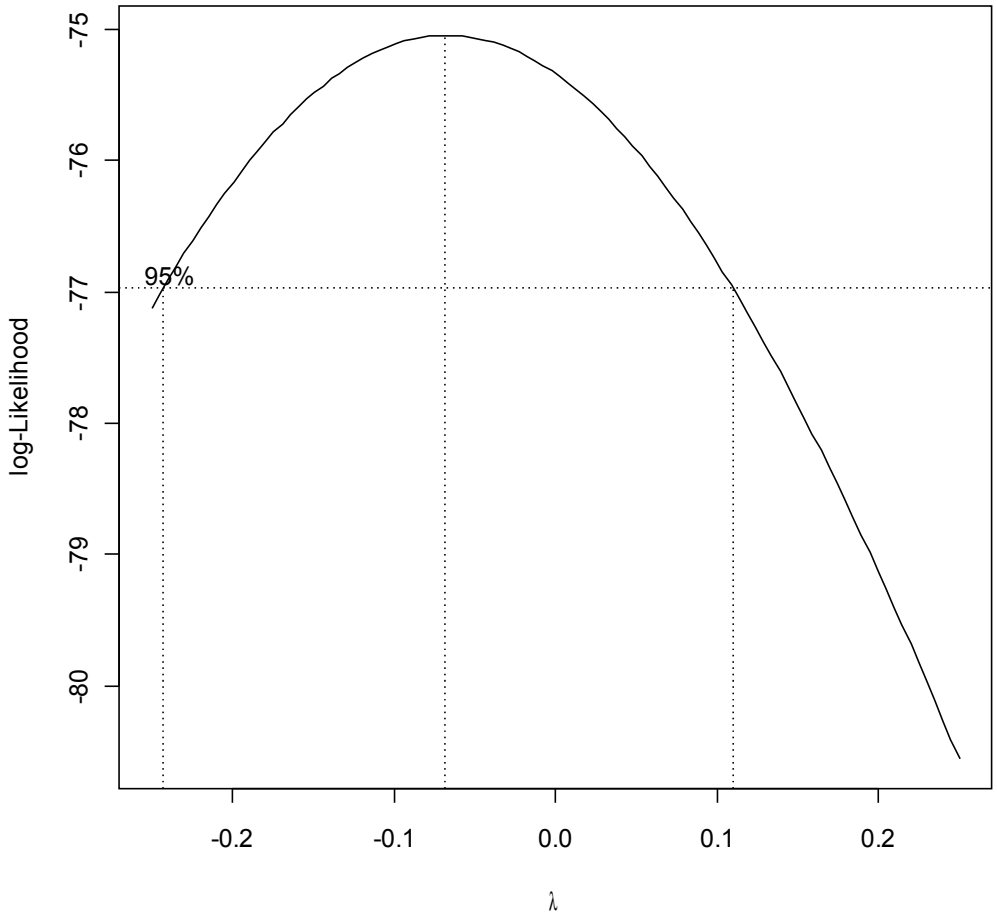
trees

```
lm4=lm(Volume^1/3~Height+Girth)
sr4=stdres(lm4)
f4=fitted(lm4)
plot(f4,sr4)
lines(lowess(f4,sr4))
```



trees

```
boxcox( Volume~log(Height)
   +log(Girth),
   lambda=seq(-.25,.25,.05))
```

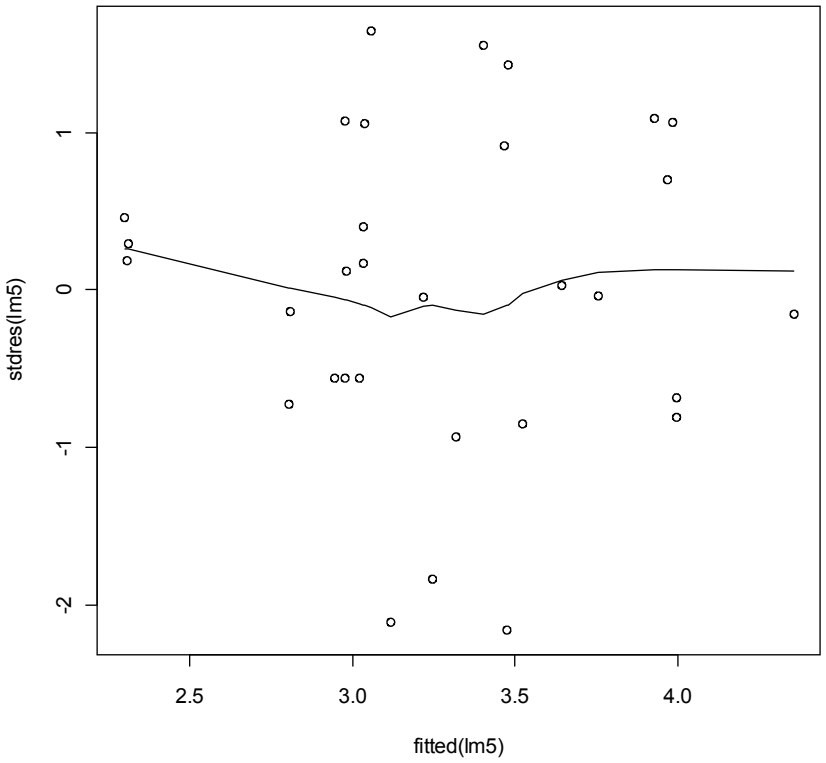


trees

```
lm5=lm(log(Volume)~log(Height)
   +log(Girth))
sr5=stdres(lm5)
f5=fitted(lm5)
plot(f5,sr5)
lines(lowess(f5,sr5))
coef(lm5)
```

```
(Intercept) log(Height)  log(Girth)
  -6.631617    1.117123    1.982650
```

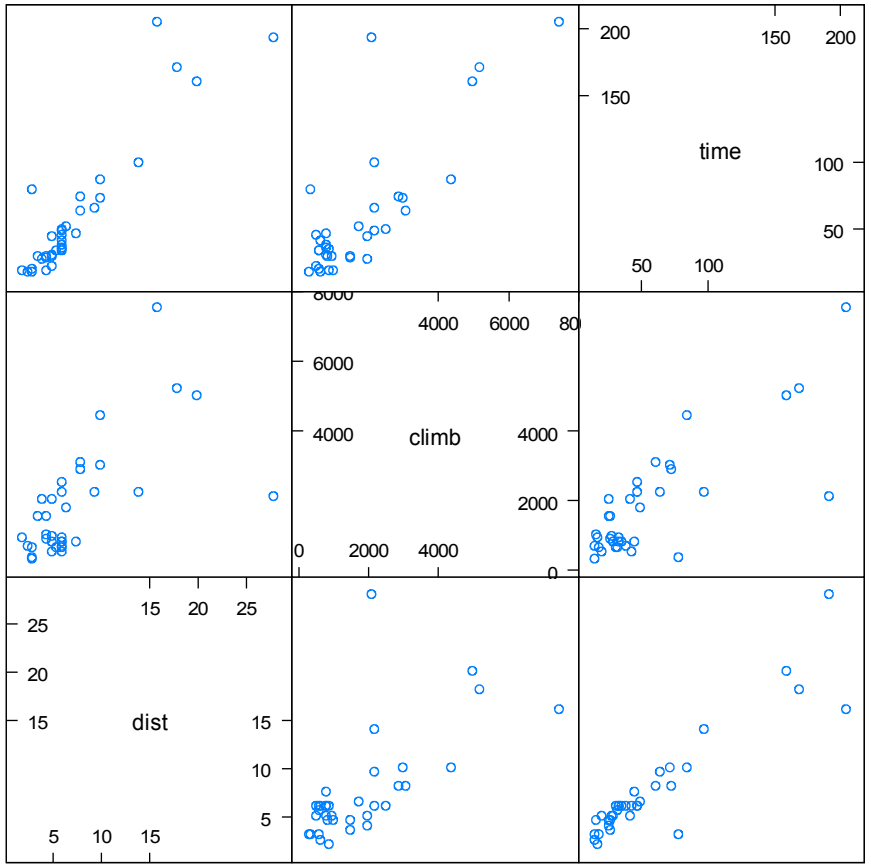
- i.e. Vol ≈ c*Height*Girth^2
- See also ?trees



Scottish hill races

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

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

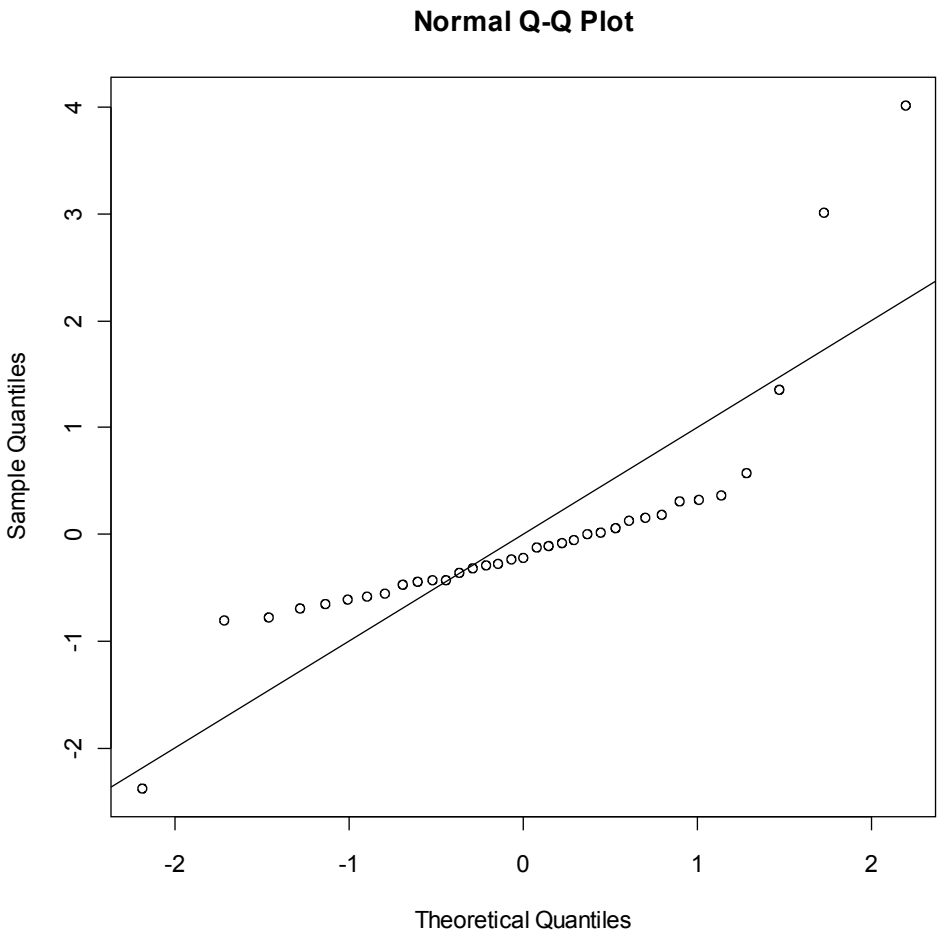


Linear model

```
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
```

Standardised residuals

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



Influence measures

```
Influence.measures(lm1)

Influence measures of
         lm(formula = time ~ dist, data = hills) :

                 dfb.1_  dfb.dist    dffit cov.r   cook.d    hat inf
Greenmantle      0.00115 -0.000794  0.00117 1.123 7.06e-07 0.0529
Carnethy         0.02247 -0.007754  0.02869 1.096 4.24e-04 0.0308
Craig Dunain    -0.08088  0.027913 -0.10326 1.075 5.44e-03 0.0308
Ben Rha         -0.06136  0.000546 -0.10395 1.070 5.51e-03 0.0286
Ben Lomond       0.00206  0.000345  0.00400 1.095 8.25e-06 0.0288
Goatfell         0.05083  0.008532  0.09890 1.073 4.99e-03 0.0288
Bens of Jura    -0.66496  1.533213  1.82255 0.307 8.75e-01 0.0977   *
Cairnpapple     -0.06162  0.021267 -0.07868 1.084 3.17e-03 0.0308
Scolty          -0.05884  0.028395 -0.06741 1.093 2.33e-03 0.0347
Traprain        -0.03779  0.013043 -0.04825 1.092 1.20e-03 0.0308
Lairig Ghru      1.42026 -2.170105 -2.24554 1.287 2.15e+00 0.4325   *
…
Knock Hill       0.75801 -0.500397  0.78251 0.590 2.29e-01 0.0483   *
…
Moffat Chase     0.02496 -0.045021 -0.04912 1.294 1.24e-03 0.1785   *
```

Cook's distance: look out for values near 1

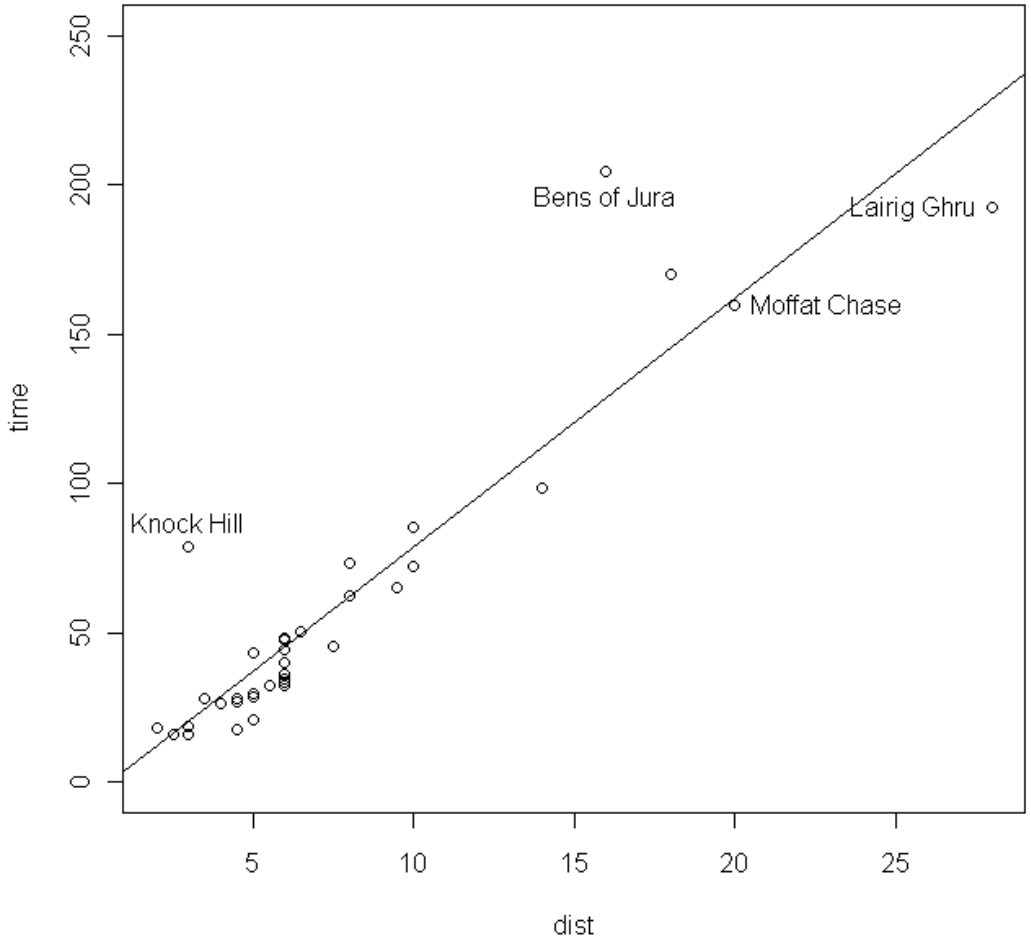
Influence measures

```
Influence.measures(lm1)

Influence measures of
         lm(formula = time ~ dist, data = hills) :

                 dfb.1_  dfb.dist    dffit cov.r   cook.d    hat inf
Greenmantle      0.00115 -0.000794  0.00117 1.123 7.06e-07 0.0529
Carnethy         0.02247 -0.007754  0.02869 1.096 4.24e-04 0.0308
Craig Dunain    -0.08088  0.027913 -0.10326 1.075 5.44e-03 0.0308
Ben Rha         -0.06136  0.000546 -0.10395 1.070 5.51e-03 0.0286
Ben Lomond       0.00206  0.000345  0.00400 1.095 8.25e-06 0.0288
Goatfell         0.05083  0.008532  0.09890 1.073 4.99e-03 0.0288
Bens of Jura    -0.66496  1.533213  1.82255 0.307 8.75e-01 0.0977   *
Cairnpapple     -0.06162  0.021267 -0.07868 1.084 3.17e-03 0.0308
Scolty          -0.05884  0.028395 -0.06741 1.093 2.33e-03 0.0347
Traprain        -0.03779  0.013043 -0.04825 1.092 1.20e-03 0.0308
Lairig Ghru      1.42026 -2.170105 -2.24554 1.287 2.15e+00 0.4325   *
…
Knock Hill       0.75801 -0.500397  0.78251 0.590 2.29e-01 0.0483   *
…
Moffat Chase     0.02496 -0.045021 -0.04912 1.294 1.24e-03 0.1785   *
```

These are all various ways of quantifying the effect of deleting the data point on the results

Effect of Outliers

```
attach(hills)
plot(dist,time,
   ylim=c(0,250))
abline(coef(lm1))
identify(dist,time,
   labels=rownames(hills))
```



Dealing with outliers

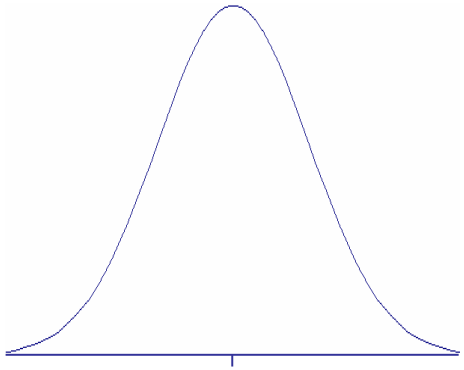
- Data-driven methods of deleting outliers are inherently dubious
- Usually what you need is a better model

How wrong is your model?

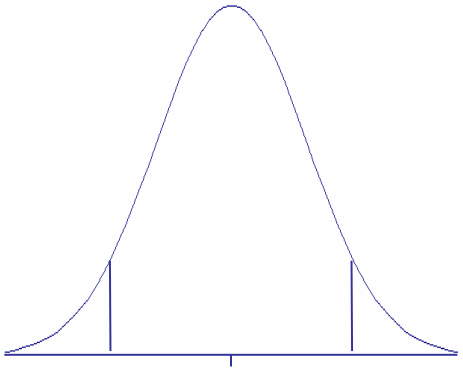
How useful is your model?

  - What do outliers tell you about model fit?
  - Does fitting your model to part of the data make it more or less useful?

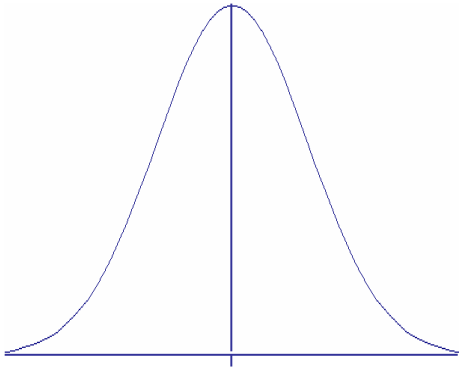
Simple M estimators



Mean



Trimmed Mean



Median

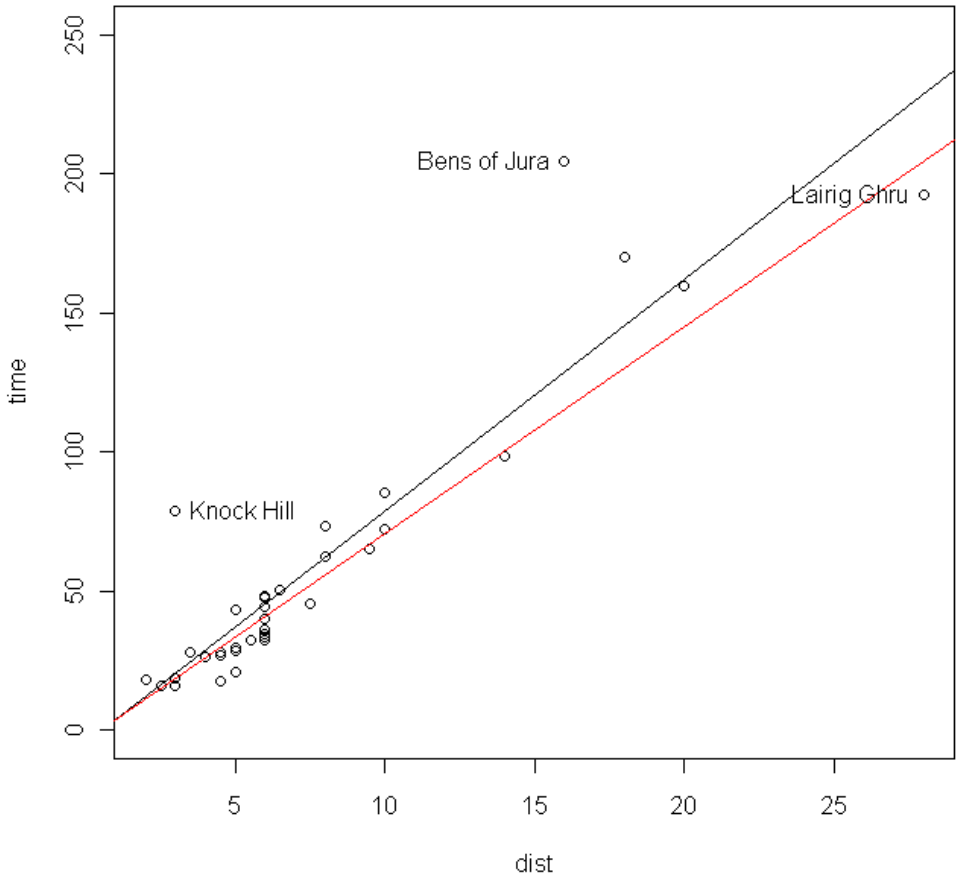
Robust linear model

```
rlm1=rlm(time~dist,data=hills,method="MM")
summary(rlm1)

Call: rlm(formula = time ~ dist, method = "MM")
Residuals:
      Min        1Q    Median        3Q       Max
-12.49415  -4.54489   0.08618   6.76252  89.51255

Coefficients:
            Value   Std. Error t value
(Intercept) -3.6742  2.4831    -1.4796
dist         7.4237  0.2673    27.7764
```

- See also ?rlm

Effect of Outliers

```
attach(hills)
plot(dist,time,
   ylim=c(0,250))
abline(coef(lm1))
abline(coef(rlm1),col="red")
identify(dist,time,
   labels=rownames(hills))
```



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.

Tests of Significance

```
lm0=lm(time~1,data=hills)
lm1=lm(time~dist,data=hills)
summary(lm1)
…
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

anova(lm0,lm1)
Analysis of Variance Table

Model 1: time ~ 1
Model 2: time ~ dist
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     34 85138
2     33 13142  1     71997 180.79 6.084e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tests of Significance

```
rlm0=rlm(time~1,data=hills,method="MM")
rlm1=rlm(time~dist,data=hills,method="MM")
summary(rlm1)
…
Coefficients:
            Value   Std. Error t value
(Intercept) -6.3603  2.8655    -2.2196
dist         8.0510  0.3084    26.1040

Residual standard error: 8.432 on 33 degrees of freedom

anova(rlm0,rlm1)
Analysis of Variance Table

Model 1: time ~ 1
Model 2: time ~ dist
  Res.Df   RSS Df Sum of Sq F Pr(>F)
1        89922
2        13682        76239
```

ANOVA 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 dummy coded 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.
- Estimated in exactly the same way as the linear regression model

Random effects ANOVA

$$y_{ijk} = b_0 + b_1 u_i + b_2 v_{ij} + e_{ijk}$$

Where

y_{ijk}	test score for individual k in school i with teacher j
b_0, b_1, b_2	regression coefficients
$u_i \sim N(0,\sigma^2_u)$	random effect of school i
$v_{ij} \sim N(0,\sigma^2_v)$	random effect of teacher j in school i
$e_{ijk} \sim N(0,\sigma^2_e)$	residual for individual k

- Also called a multilevel model or hierarchical linear model (HLM), or mixed model if there are covariates (fixed effects)
- Estimated using `lme4` in *R* or using other software

Reading

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