Data analysis with R

Lecture 10

More on statistical modelling

Jouni Junnila

Regression with a single predictor

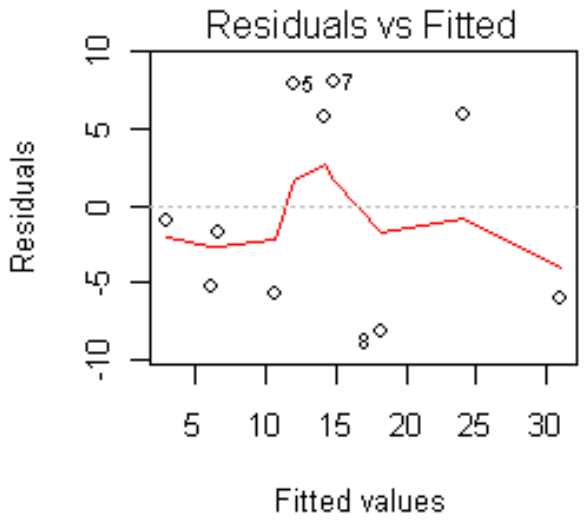
- On Monday we handled models, where the explanatories were factors.
- When the explanatories are numeric variables, we fit regression models.
- Simplest possible regression model is a model, where there is only one predictor.
- An example of this is the roller data previously investigated, where we want to explain depression with weight of the roller.

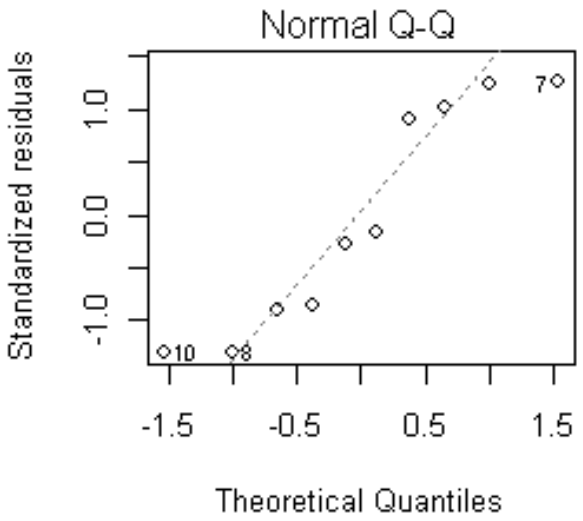
Example regression model in R

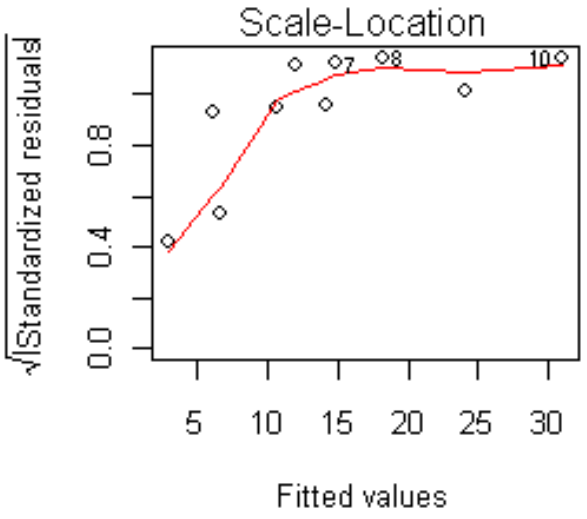
```
model <- lm(depression~weight, data=roller);summary(model)

Call:
lm(formula = depression ~ weight, data = roller)
Residuals:
   Min     1Q Median     3Q    Max
-8.180 -5.580 -1.346  5.920  8.020
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.0871     4.7543  -0.439  0.67227
weight        2.6667     0.7002   3.808  0.00518 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.735 on 8 degrees of freedom
Multiple R-squared: 0.6445,     Adjusted R-squared: 0.6001
F-statistic:  14.5 on 1 and 8 DF,  p-value: 0.005175
```

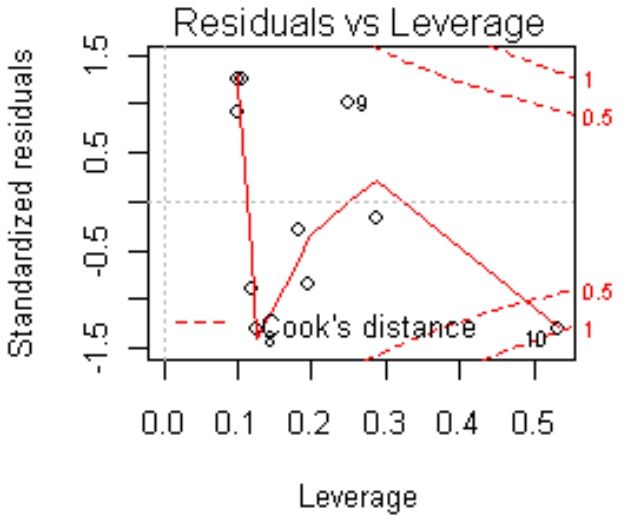
Diagnostic plots

- If we use the *plot*-function on to our *model*-object, we'll get a set of four diagnostic plots.
  - One is already familiar to us (the normal QQ-plot), other diagnostic what will be printed are "Residuals vs fitted", "Scale-Location", "Residuals vs Leverage"
  - Investigation of the three other diagnostic plots is left out from this course.
  - *plot(model)*


Residuals vs Fitted
Residuals
Fitted values
5
7
8


Normal Q-Q
Standardized residuals
Theoretical Quantiles
7
10
8

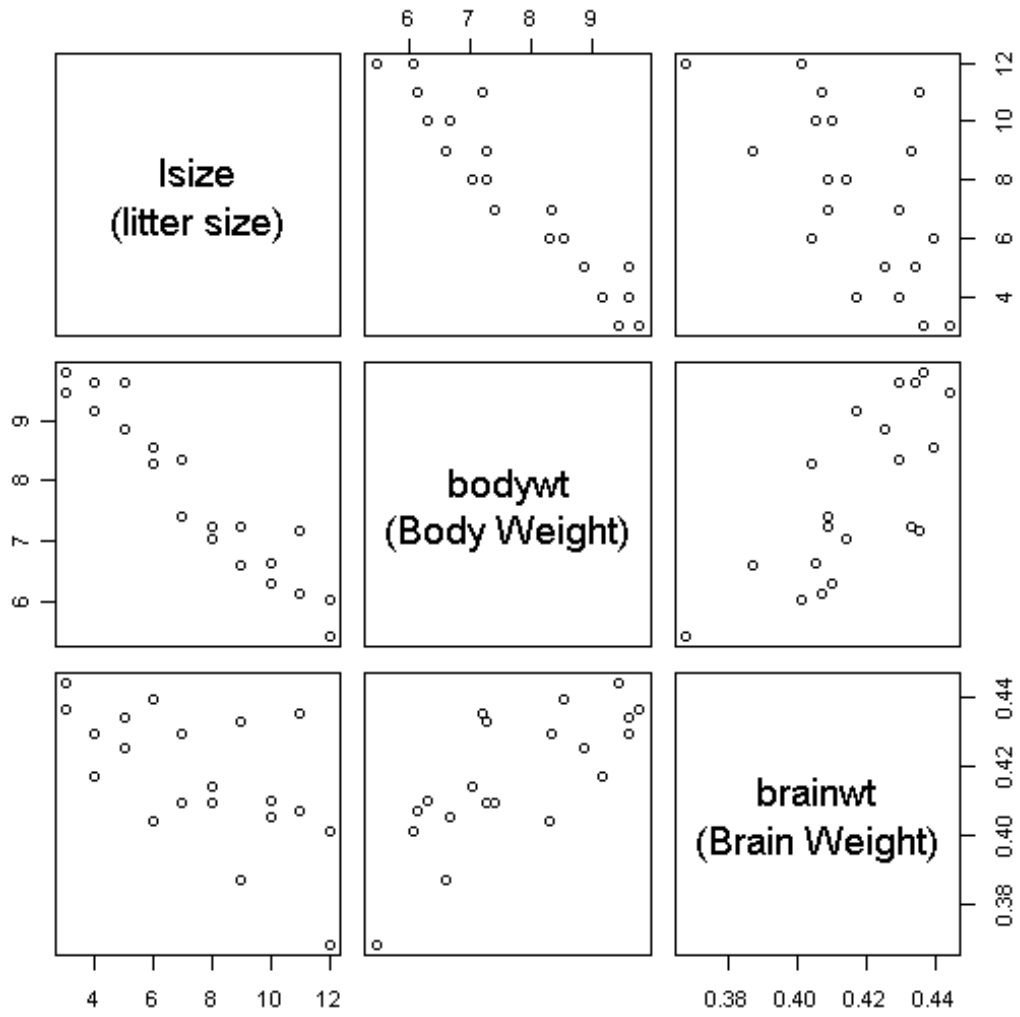

Scale-Location
√|Standardized residuals|
Fitted values
7
8
10


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1
0.5
0.5
1
9
8
10

Multiple linear regression

- Multiple linear regression generalizes methodology of simple linear regression to allow multiple predictor variables.
- Let's consider an example of multiple regression.
  - Dataset *litters* has data on the variables *lsize* (litter size), *bodywt* (body weight) and *brainwt* (brain weight), for 20 mice.
  - Our goal is to explain the brain weight with litter size and body weight.
- Let's first plot the data with matrix plot.

```
pairs(litters, labels=c("lsize\n(litter size)",
"bodywt\n(Body Weight)","brainwt\n(Brain
Weight)"))
```



Interpretating the plots

- From the plots we can see, that there could be some linear connection between brain weight and both litter size and body weight.
  - It seems to sensible to fit a regression model, with the two predictors.
  - We can use the *lm*-function here again and type
    - `lm(brainwt ~ lsize + bodywt + lsize*bodywt, data=litters)`

Modifying the model

- Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2443473  0.0862005   2.835   0.0120 *
lsize        -0.0022600  0.0069043  -0.327   0.7477
bodywt        0.0161056  0.0086876   1.854   0.0823 .
lsize:bodywt  0.0011854  0.0008212   1.443   0.1682
```

- The interaction term is not significant. Let's remove that and fit the model again.
  - `lm(brainwt ~ lsize + bodywt, data=litters)`

Interpretation

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.178247   0.075323   2.366  0.03010 *
lsize       0.006690   0.003132   2.136  0.04751 *
bodywt      0.024306   0.006779   3.586  0.00228 **
```

- Both *lsize* and *bodywt* seems to be significant explanatories.
- The estimates seem to be very small, but still they are significant, why?
- The estimate gives us a value, that how much will a change of one unit in the predictor effect the response.
- The response values are 0.368-0.439. So small in general and also small variation. That's why our estimates are also small.

Interpretation (2)

- In multiple regression it's not as straight forward to interpret the coefficients as with simple linear regression.
- In our example, the coefficients for *lsize* estimates the change in *brainwt* with *lsize* when *bodywt* is held constant. Same applies naturally with *bodywt*.
- For any particular value of *bodywt*, *brainwt* increases with *lsize*. This is a significant finding for the purpose of the study.

Linear mixed models

- So far we have had just fixed effects in the model.
- Quite often we have designs, that require using random effects as well.
- If we have both random and fixed effects in the model, we call the model a linear mixed model.
- Convinient functions for fitting mixed models are *lme* (*nlme*-library) and *lmer* (lme4-library)
- *lmer4* is useful also for fitting generalized linear mixed models and nonlinear mixed models.

lme-example

- The *lme*-function is specially good, if we have nested random effects.
  - Function allows them, and they are also easy to code.
- The Orthodont data has 108 rows and 4 columns of the change in an orthdontic measurement over time.
  - *distance*. A numeric vector of distances from the pituitary to the pterygomaxillary fissure (mm).
  - *age*. A numeric vector of ages of the subjects.
  - *subject*. A factor for the subject ID.
  - *sex*. A factor with levels Male and Female

Model construction

- Our goal is to explain *distance* with *age*. We also have to consider that we have several samples from the same subject. Thus, subject should be consider as a random variable.

```
f1 <- lme(distance ~ age, data=Orthodont,
random=~1|Subject)
```

Summary(lme)

```
>summary(f1)
Linear mixed-effects model fit by REML
 Data: Orthodont
       AIC      BIC    logLik
  454.6367 470.6173 -221.3183

Random effects:
 Formula: ~age | Subject
 Structure: General positive-definite, Log-Cholesky
   parametrization
             StdDev    Corr
(Intercept) 2.3270340 (Intr)
age         0.2264278 -0.609
Residual    1.3100397
```

Summary(lme) [2]

```
Fixed effects: distance ~ age
                Value Std.Error DF   t-value p-value
(Intercept) 16.761111 0.7752460 80 21.620377       0
age          0.660185 0.0712533 80  9.265333       0
 Correlation:
    (Intr)
age -0.848
Standardized Within-Group Residuals:
Min         Q1          Med         Q3          Max
-3.2231061 -0.4937611  0.0073166  0.4721511  3.9160332

Number of Observations: 108
Number of Groups: 27
```

Linear mixed models & nesting

- Often we have also nesting structures in our design.
- Obviously we have to consider also this when we form our statistical model.
- Nesting is most commonly related to random variables.
  - Eg. Subject nested in centers.
- Next, we'll consider an example about linear mixed effects model, with nesting structure in the random effects.

Example-data

- The data is called *kiwishade*. It's from a designed experiment that compared different kiwifruit shading treatments.
- There are four different shading treatments.
- There are four vines in each plot and four plots (one for each treatment) in each of the three blocks (north,west,east)
  - Vine is nested in plot
  - Plot is nested in block.

Constructing the model

- We can define a suitable model in many different ways and in actual situation we would naturally test several possibilities.
- However, here we'll go through one example of a suitable and easy to code model.
- `mo <- lme(yield~shade, random=~1 | block/plot, data=kiwishade)`

summary

- > summary(mo)

```
Linear mixed-effects model fit by REML
 Data: kiwishade
       AIC      BIC    logLik
  265.9663 278.4556 -125.9831

Random effects:
 Formula: ~1 | block
        (Intercept)
StdDev:    2.019373

 Formula: ~1 | plot %in% block
        (Intercept) Residual
StdDev:    1.478623 3.490381
```

Summary (2)

```
Fixed effects: yield ~ shade
                 Value Std.Error DF  t-value p-value
(Intercept)  100.20250  1.761617 36 56.88098  0.0000
shadeAug2Dec   3.03083  1.867621  6  1.62283  0.1558
shadeDec2Feb -10.28167  1.867621  6 -5.50522  0.0015
shadeFeb2May  -7.42833  1.867621  6 -3.97743  0.0073
 Correlation:
             (Intr) shdA2D shdD2F
shadeAug2Dec -0.53
shadeDec2Feb -0.53   0.50
shadeFeb2May -0.53   0.50   0.50

Number of Observations: 48
Number of Groups:
          block plot %in% block
              3              12
```

Interpretation

- From the results we are now able to determine
  1) How does different treatments effect the yield.
  2) How much variation does the block create
  3) How much variation does the plot create
  4) How much variation there is between different vines.