Mixed effects models

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

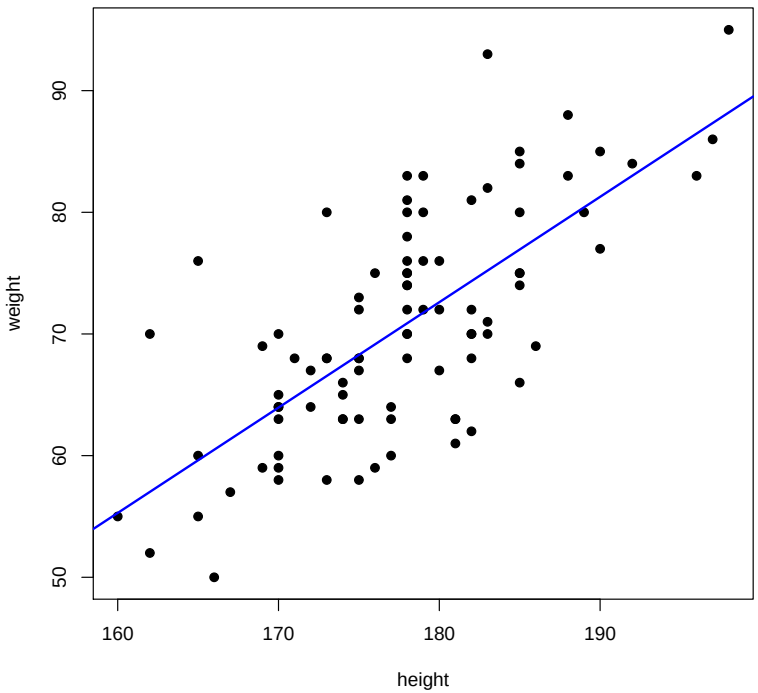
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

jan.graffelman@upc.edu

October 24, 2022

Contents

Classical linear regression



Theoretical model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Usual assumptions:

- $E(\varepsilon_i) = 0$.
- $V(\varepsilon_i) = \sigma^2$ (constant variance).
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ (independent observations).
- $\varepsilon_i \sim N(0, \sigma^2)$.

Summarized:

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma)$$

Correlated errors/observations

- Standard regression methods assume uncorrelated errors
- This assumption can be unwarranted in several circumstances. E.g.:
  - Student grades when students are grouped in schools
  - Biochemical markers of individuals inside families
  - Treatment variables of patients inside hospitals
  - Repeated measurements of individuals over time (longitudinal data)
  - ....
- Mixed effect models allow for correlation among observations in clusters
- Mixed effect models are also known as
  - Random coefficient models
  - Variance component models
  - Hierarchical models
  - Multilevel models

Random intercept model

- Let y_{ij} represent observation j in cluster i
- Random intercept model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + \varepsilon_{ij} \qquad i = 1, \ldots, M \quad j = 1, \ldots, n_i$$

- with

$$u_i \sim N(0, \sigma_u^2) \qquad \varepsilon_{ij} \sim N(0, \sigma^2) \qquad u_i, \varepsilon_{ij} \text{ independent}$$

- β_1 is a fixed effect, u_i is a random effect.

$$V\left(y_{ij}\right) = V\left(u_i + \varepsilon_{ij}\right) = \sigma_u^2 + \sigma^2$$

- The correlation between two error terms of the same individual, the intraclass correlation, is

$$Cor(u_i + \varepsilon_{ij}, u_i + \varepsilon_{ik}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}$$

- Alternatively,

$$y_{ij} = \alpha_i + \beta_1 x_{ij} + \varepsilon_{ij} \qquad \alpha_i = \beta_0 + u_i \qquad \alpha_i \sim N(\beta_0, \sigma_u^2)$$

Mixed model estimation

- Two methods that are used to estimate mixed models
  - Maximum likelihood estimation (ML)
  - Restricted maximum likelihood estimation (REML)
- ML estimators are known to underestimate variance components
- REML have been developed to compensate for this
- In practice both methods are used and their estimates compared
- REML estimates for the variance components are typically larger

Example data set: pig growth

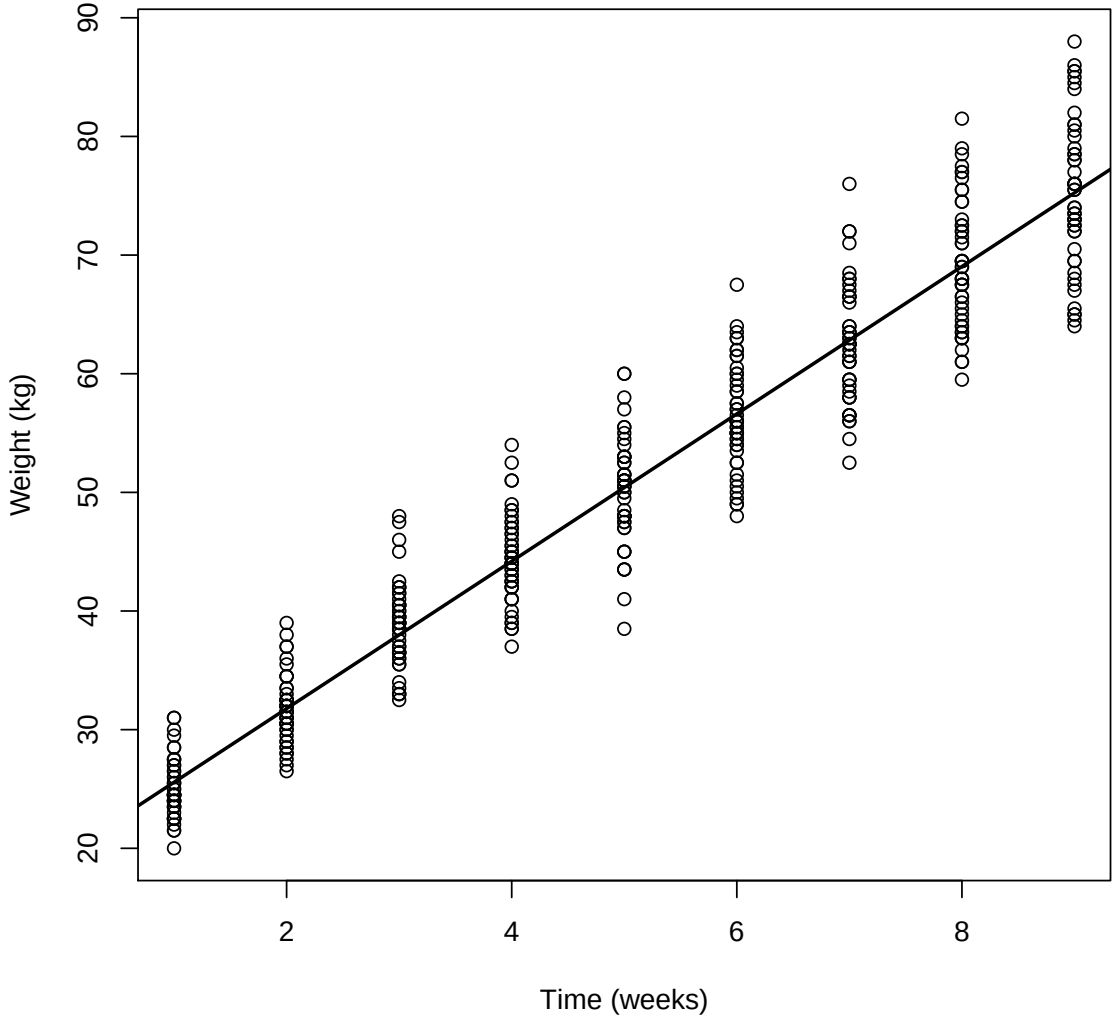
In wide format

	1	2	3	4	5	6	7	8	9
1	24.0	32.0	39.0	42.5	48.0	54.5	61.0	65.0	72.0
2	22.5	30.5	40.5	45.0	51.0	58.5	64.0	72.0	78.0
3	22.5	28.0	36.5	41.0	47.5	55.0	61.0	68.0	76.0
4	24.0	31.5	39.5	44.5	51.0	56.0	59.5	64.0	67.0
5	24.5	31.5	37.0	42.5	48.0	54.0	58.0	63.0	65.5
6	23.0	30.0	35.5	41.0	48.0	51.5	56.5	63.5	69.5
7	22.5	28.5	36.0	43.5	47.0	53.5	59.5	67.5	73.5
8	23.5	30.5	38.0	41.0	48.5	55.0	59.5	66.5	73.0
9	20.0	27.5	33.0	39.0	43.5	49.0	54.5	59.5	65.0
10	25.5	32.5	39.5	47.0	53.0	58.5	63.0	69.5	76.0
11	24.5	31.0	40.5	46.0	51.5	57.0	62.5	69.5	76.0
12	24.0	29.0	39.0	44.0	50.5	57.0	61.5	68.0	73.5
13	23.5	30.5	36.5	42.0	47.0	55.0	59.0	65.5	73.0
14	21.5	30.5	37.0	42.5	48.0	52.5	58.5	63.0	69.5
15	25.0	32.0	38.5	44.0	51.0	59.0	66.0	75.5	86.0
16	21.5	28.5	34.0	39.5	45.0	51.0	58.0	64.5	72.5
17	31.0	38.0	48.0	54.0	60.0	62.0	66.5	75.5	84.0
18	27.5	32.5	36.0	43.0	49.5	52.5	56.0	61.0	64.0
19	30.0	37.0	45.0	51.0	58.0	63.0	67.5	74.5	81.0
20	26.0	32.0	40.5	45.5	52.5	55.5	62.5	69.5	74.0
21	26.0	32.5	39.5	44.0	48.0	54.5	58.0	66.0	73.0
22	28.5	35.5	41.5	47.5	54.0	59.5	63.5	71.0	78.5
23	26.5	34.5	42.0	48.5	55.5	62.0	68.0	76.5	85.0
24	27.5	33.5	41.0	45.0	50.5	56.0	62.5	71.0	78.0
25	22.5	27.0	33.5	38.5	41.0	49.0	56.0	64.0	68.0
26	22.0	26.5	32.5	38.5	43.5	50.5	56.5	63.5	68.5
27	23.5	29.0	35.5	40.0	45.0	50.0	56.5	63.0	67.5
28	22.5	29.5	36.5	42.0	45.0	55.0	61.0	68.0	72.0
29	27.5	34.5	42.0	47.5	53.0	63.0	72.0	79.0	85.5
30	23.5	28.0	33.0	37.0	38.5	48.0	52.5	62.0	64.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
48	28.5	36.0	42.5	49.0	55.0	63.5	72.0	78.5	85.5

In long format

	subject	time	weight
1	1	1	24.0
2	2	1	22.5
3	3	1	22.5
4	4	1	24.0
5	5	1	24.5
6	6	1	23.0
7	7	1	22.5
8	8	1	23.5
9	9	1	20.0
10	10	1	25.5
11	11	1	24.5
12	12	1	24.0
13	13	1	23.5
14	14	1	21.5
15	15	1	25.0
16	16	1	21.5
17	17	1	31.0
18	18	1	27.5
19	19	1	30.0
20	20	1	26.0
21	21	1	26.0
22	22	1	28.5
23	23	1	26.5
24	24	1	27.5
25	25	1	22.5
26	26	1	22.0
27	27	1	23.5
28	28	1	22.5
29	29	1	27.5
30	30	1	23.5
⋮	⋮	⋮	⋮
432	48	9	85.5

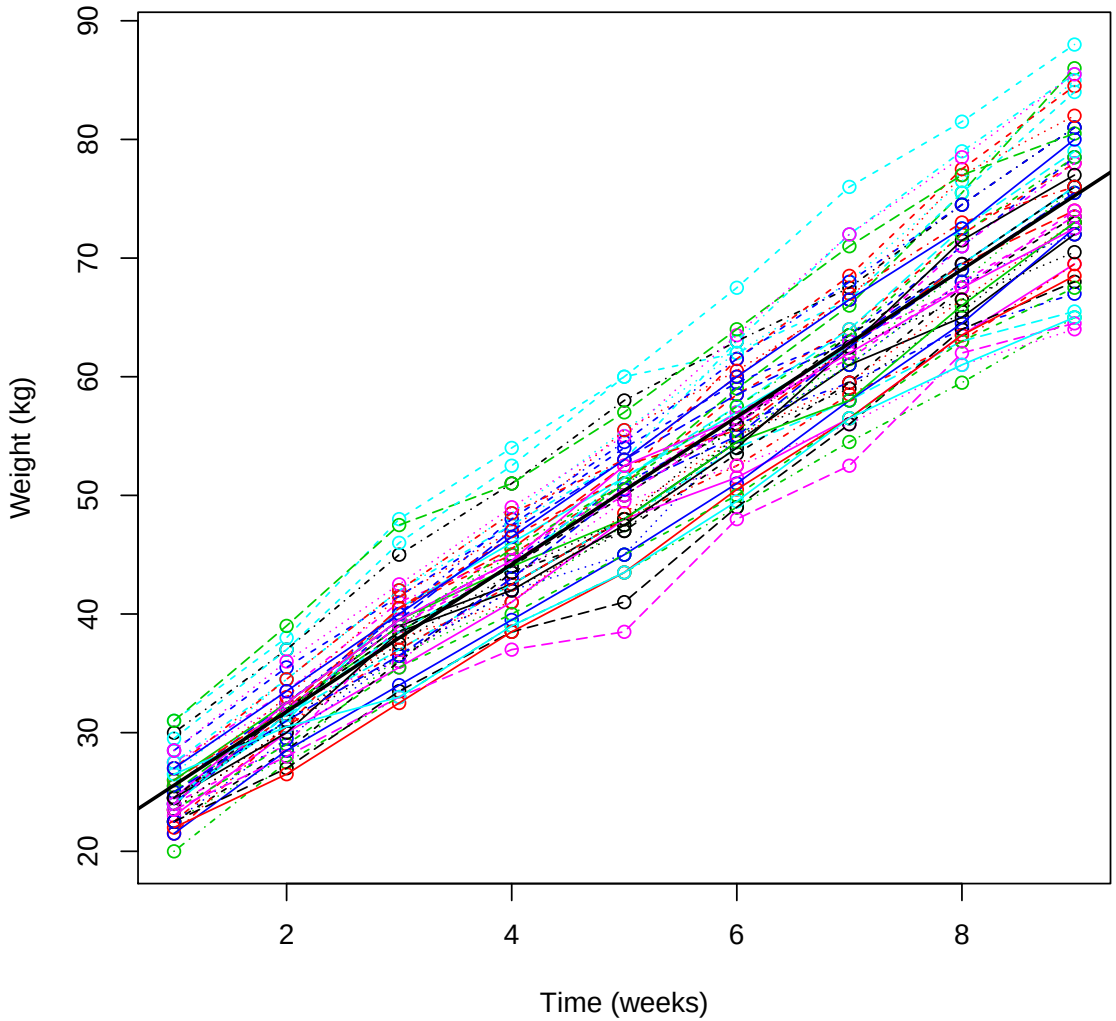
OLS regression



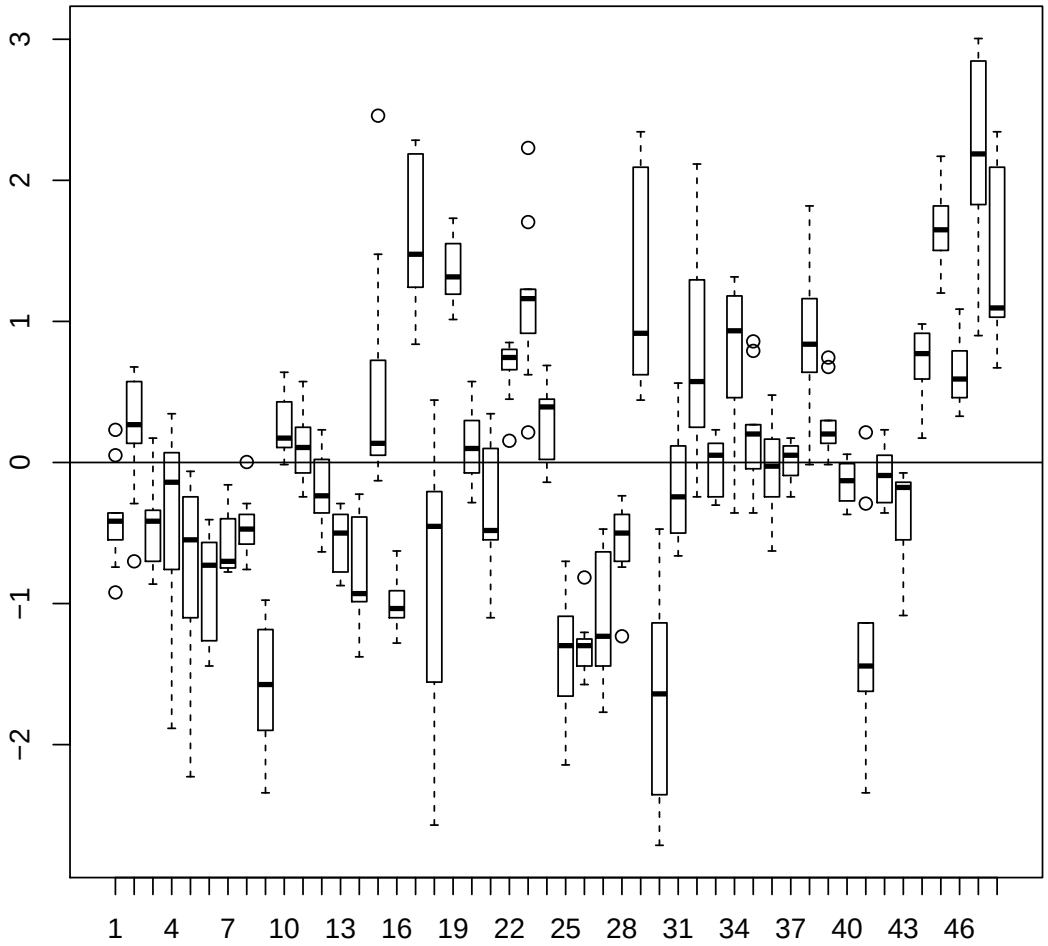
OLS regression

```
> model.0 <- lm(weight~time,data=Pigs)
> summary(model.0)

Call:
lm(formula = weight ~ time, data = Pigs)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9051  -2.5348  -0.1952   2.5949  13.1751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.35561    0.46054   42.03   <2e-16 ***
time         6.20990    0.08184   75.88   <2e-16 ***
---

Residual standard error: 4.392 on 430 degrees of freedom
Multiple R-squared:  0.9305,Adjusted R-squared:  0.9303
F-statistic:  5757 on 1 and 430 DF,  p-value: < 2.2e-16

>
```

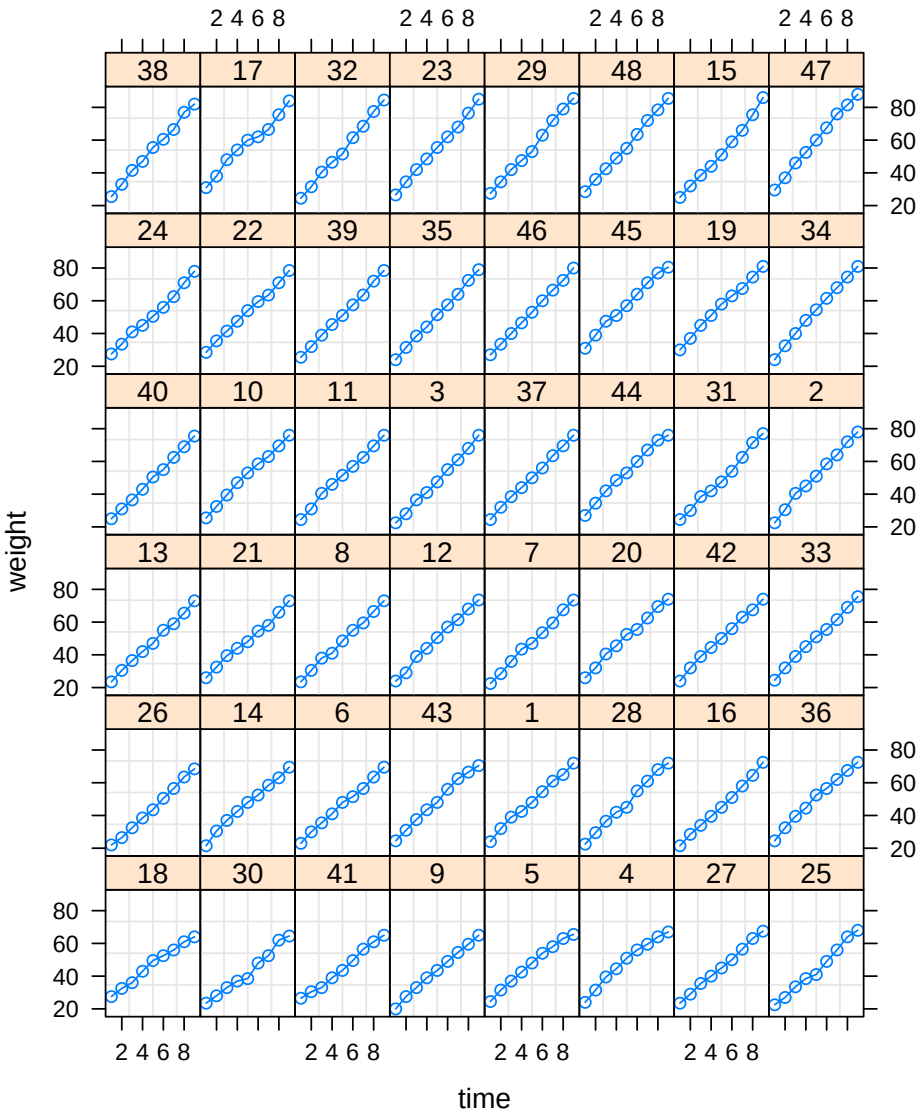
Assessing fit graphically



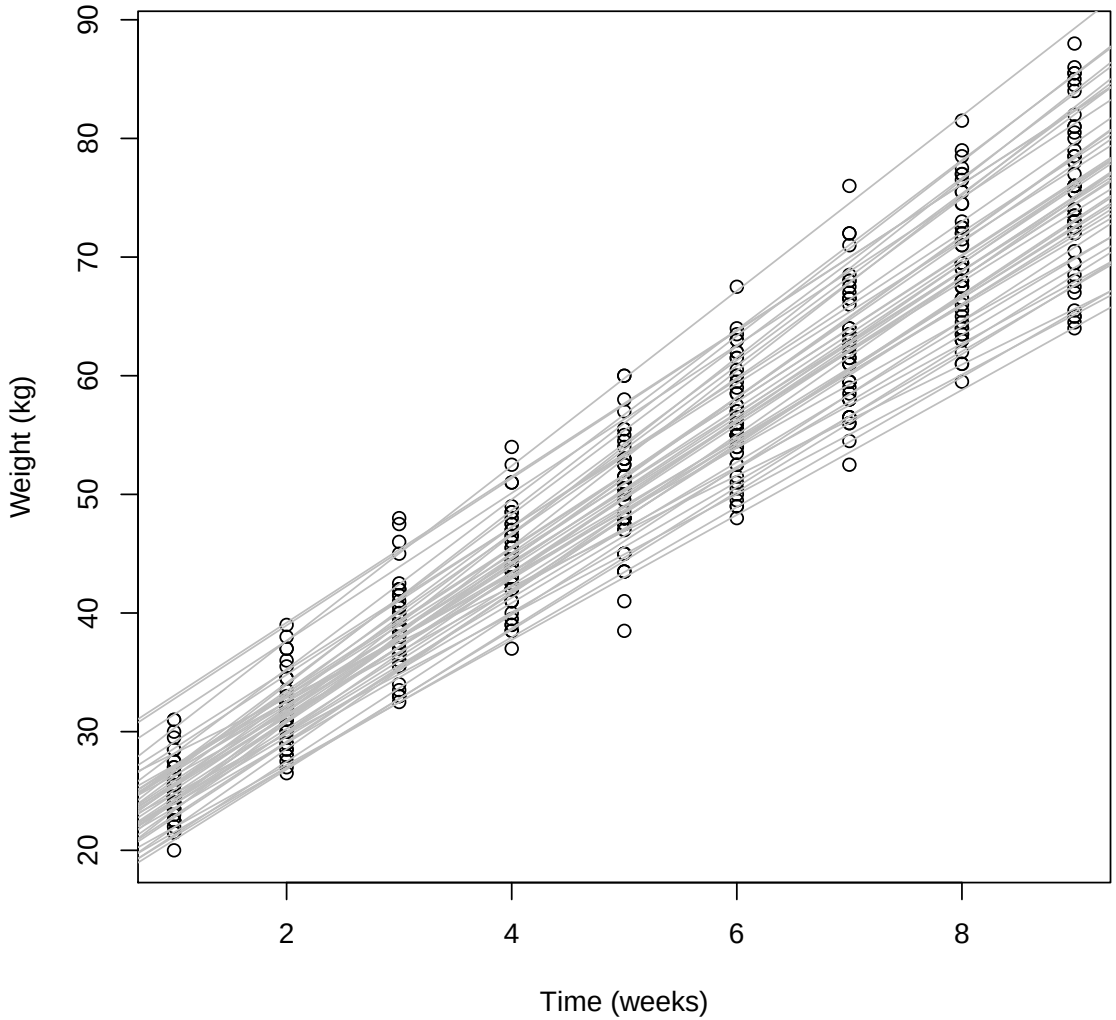
Plotting residuals



Re-plotting the data



Separate regressions



R packages for mixed models

- There are a least two packages for estimation of mixed models in R.
  - Package `lme4` with function `lmer`
  - Package `nmle` with function `lme`
- The examples in this module are made with `nmle` and `lme`.

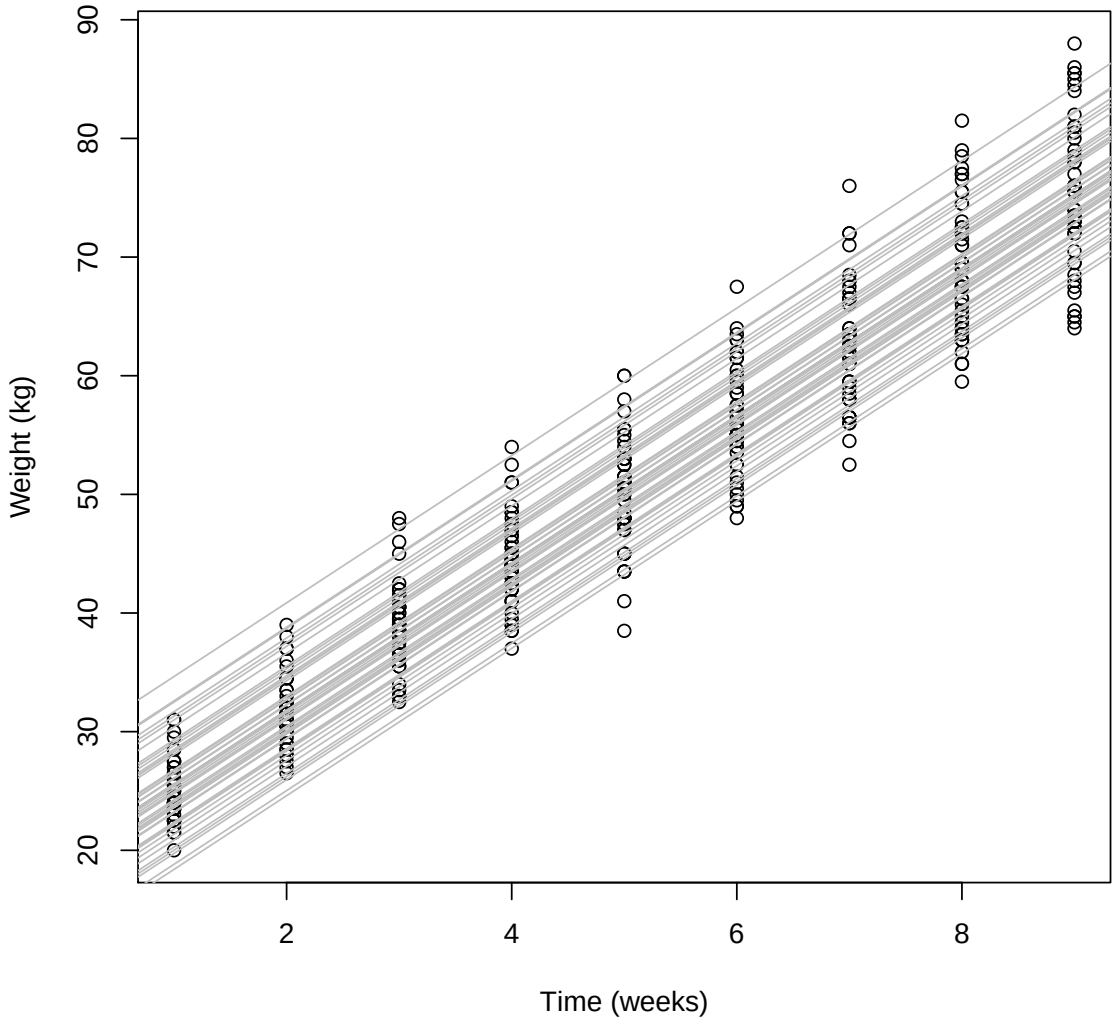
Fitting the random intercept model

```
> library(nlme)
> model.1 <- lme(weight~time,data=Pigs,random=~1|subject)
> summary(model.1)
Linear mixed-effects model fit by REML
 Data: Pigs
       AIC      BIC    logLik
  2041.797 2058.052 -1016.898

Random effects:
 Formula: ~1 | subject
        (Intercept) Residual
StdDev:    3.891253 2.096356

Fixed effects: weight ~ time
                Value Std.Error  DF   t-value p-value
(Intercept) 19.355613 0.6031390 383  32.09146       0
time         6.209896 0.0390633 383 158.97012       0
 Correlation:
     (Intr)
time -0.324

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.73902210 -0.54562381  0.01835208  0.51221200  3.93133783

Number of Observations: 432
Number of Groups: 48
>
```

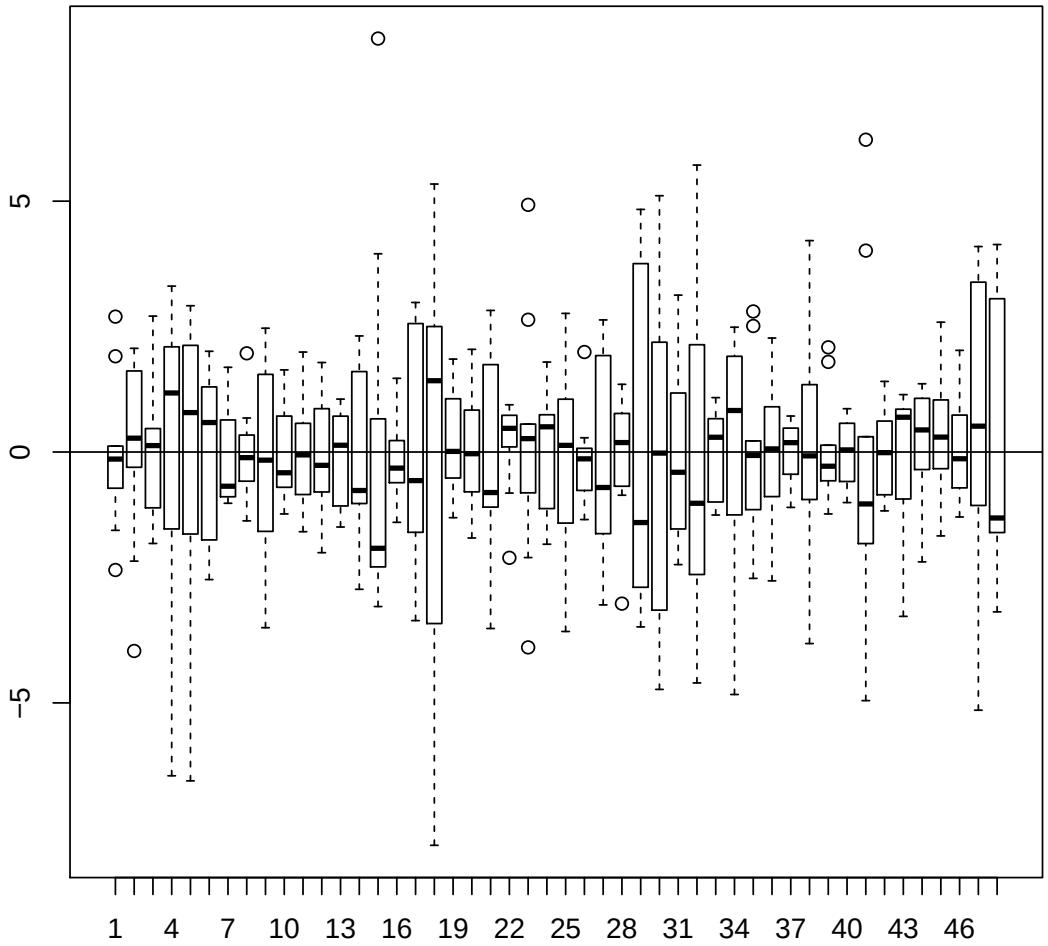
```
> intervals(model.1)
Approximate 95% confidence intervals

 Fixed effects:
                lower      est.     upper
(Intercept) 18.16974 19.355613 20.541492
time         6.13309  6.209896  6.286701
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: subject
                       lower     est.   upper
sd((Intercept)) 3.158269 3.891253 4.79435

 Within-group standard error:
   lower     est.    upper
1.953029 2.096356 2.250202
```

$$\hat{\rho} = \frac{(3.89)^2}{(3.89)^2 + (2.096)^2} = 0.775$$

The fitted model



Residuals random intercept model



Comparing models

- There are several criteria to compare linear and linear mixed models
- Let k be the difference in number of parameters between two models.
- Difference in deviance (likelihood ratio test) between general model L_2 and restricted model L_1

$$G^2 = 2\ln\left(\frac{L_2}{L_1}\right) = 2\ln(L_2) - 2\ln(L_1) = D_1 - D_2 \sim \chi^2_k \text{ under } H_0$$

- Akaike information criterion (AIC)

$$AIC = 2k - 2\ln\left(L(\hat{\theta})\right)$$

- Bayesian information criterion (BIC)

$$BIC = k\ln(N) - 2\ln\left(L(\hat{\theta})\right)$$

- Smaller AIC and BIC indicate better fit

The two models for the Pigs data

```
> anova(model.1,model.0)
        Model df      AIC      BIC    logLik   Test L.Ratio p-value
model.1     1  4 2041.797 2058.052 -1016.898
model.0     2  3 2512.945 2525.136 -1253.472 1 vs 2 473.148  <.0001
>
```

Random slope and intercept model

Random slope and intercept model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + v_i x_{ij} + \varepsilon_{ij} \qquad i = 1, \ldots, M \quad j = 1, \ldots, n_i$$

$$u_i \sim N(0, \sigma_u^2), \quad v_i \sim N(0, \sigma_v^2), \quad Cov(u, v) = \sigma_{u,v}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Alternatively,

$$y_{ij} = \alpha_i + \gamma_i x_{ij} + \varepsilon_{ij}$$

with

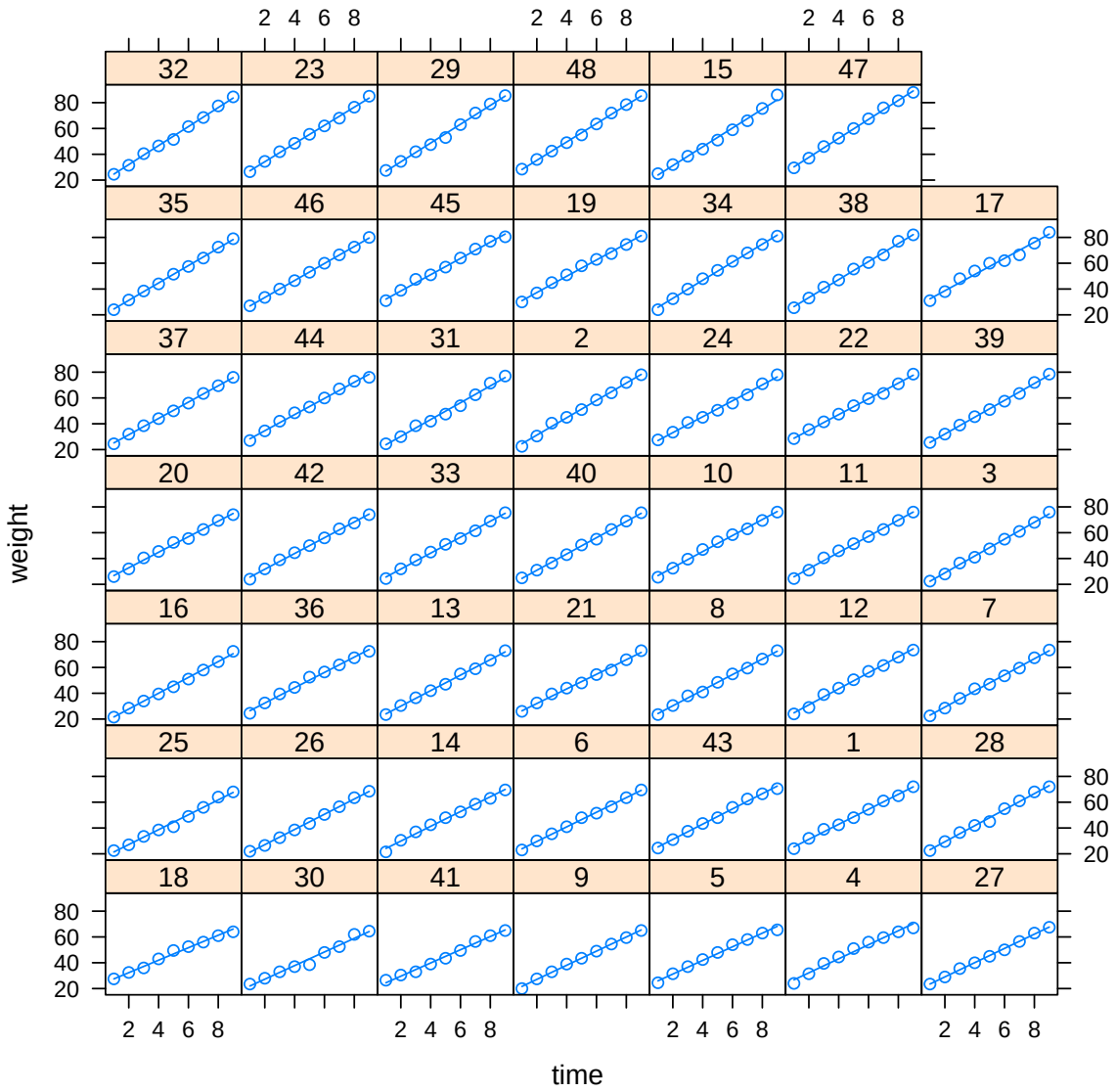
$$\alpha_i = \beta_0 + u_i \quad \text{and} \quad \gamma_i = \beta_1 + v_i$$

$$\alpha_i \sim N(\beta_0, \sigma_u^2) \quad \gamma_i \sim N(\beta_1, \sigma_v^2)$$

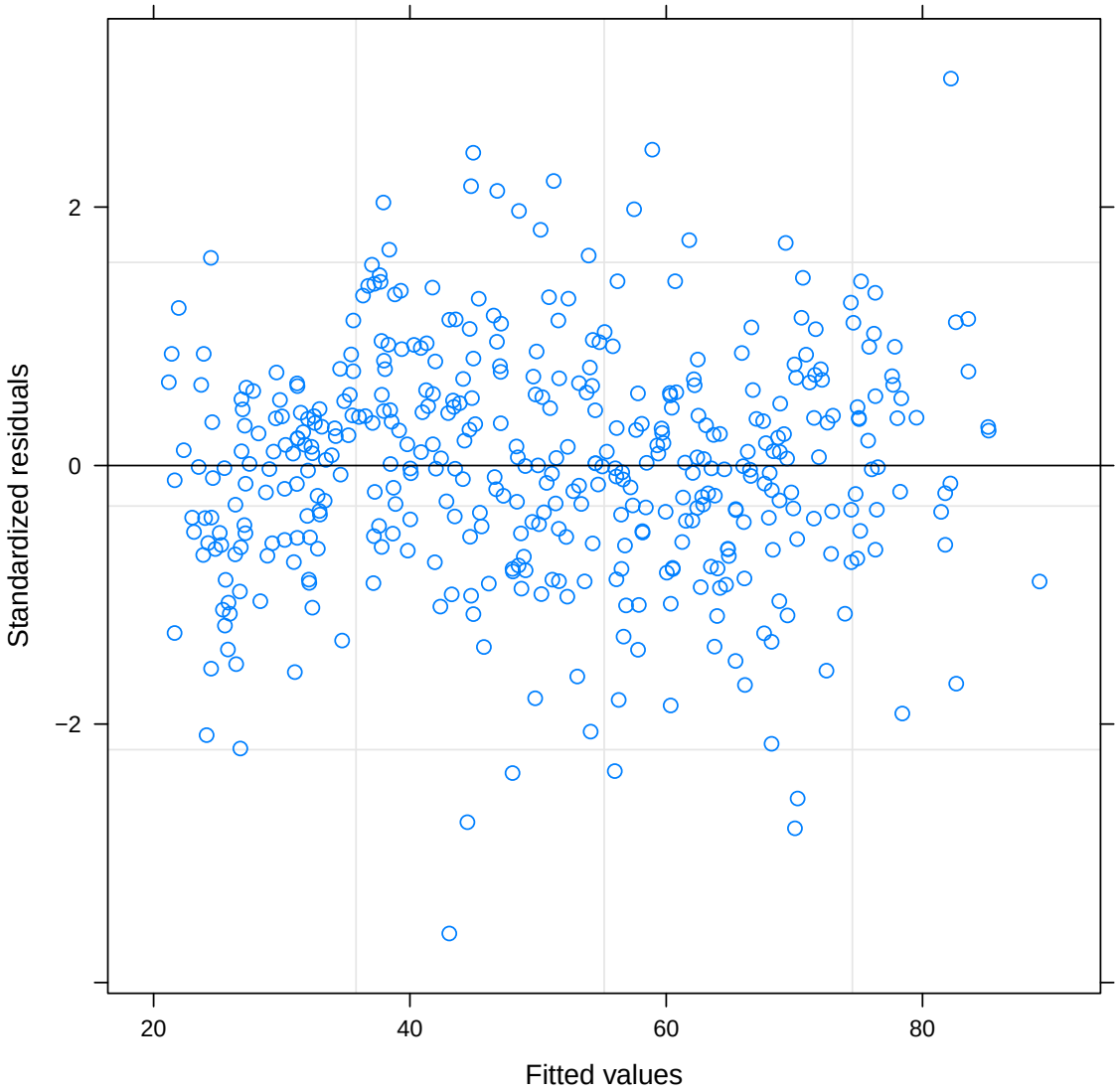
Fitting the random slope and intercept model

```
> model.2 <- lme(weight~time,data=Pigs,random=~time|subject)
> summary(model.2)
Linear mixed-effects model fit by REML
 Data: Pigs
       AIC      BIC    logLik
  1752.871 1777.254 -870.4356

Random effects:
 Formula: ~time | subject
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 2.6431920 (Intr)
time        0.6164379 -0.063
Residual    1.2636572

Fixed effects: weight ~ time
                Value Std.Error  DF  t-value p-value
(Intercept) 19.355613 0.4038676 383 47.92564       0
time         6.209896 0.0920382 383 67.47085       0
 Correlation:
     (Intr)
time -0.133

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.62018844 -0.54735954  0.01503617  0.54855117  2.99391406

Number of Observations: 432
Number of Groups: 48
```

Fitted random slope and intercept model



Residuals random slope and intercept

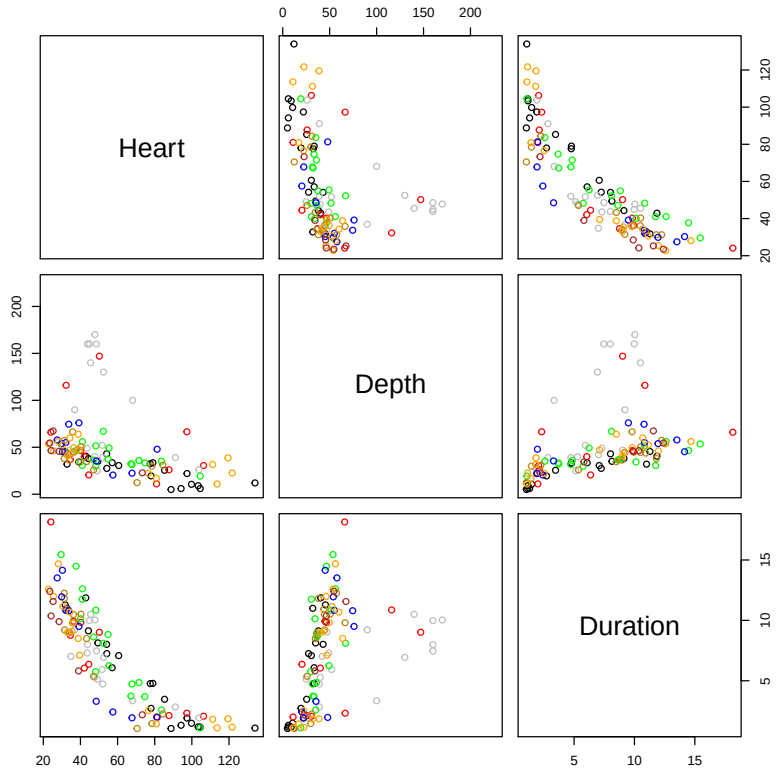


Comparing models

```
> anova(model.1,model.2)
        Model df      AIC      BIC     logLik   Test  L.Ratio p-value
model.1     1  4 2041.797 2058.052 -1016.8984
model.2     2  6 1752.871 1777.254  -870.4356 1 vs 2 292.9256  <.0001
```

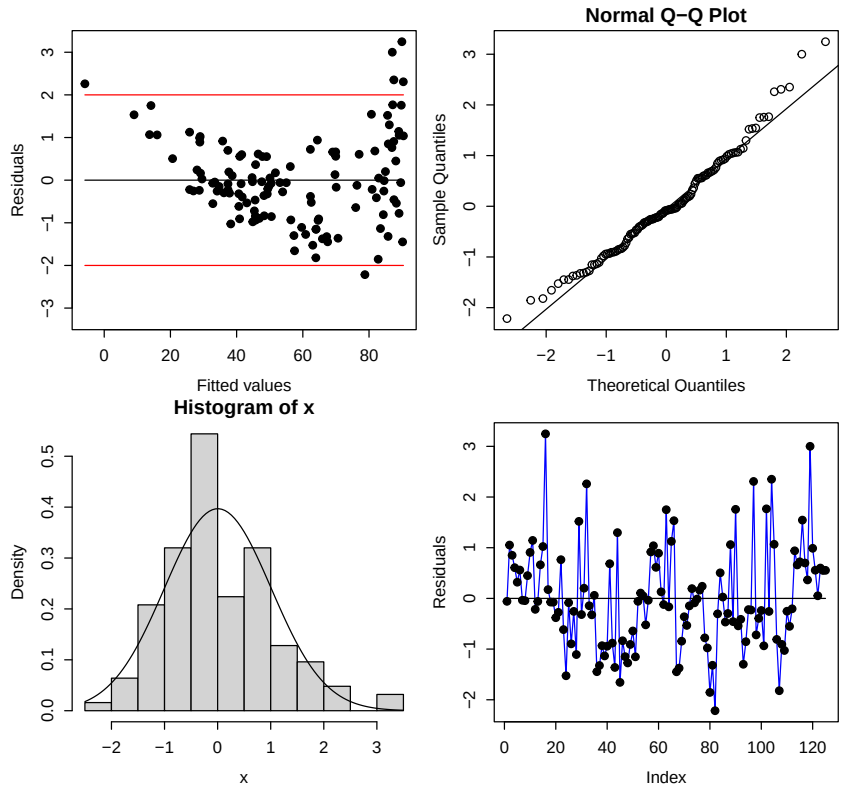
Example: Penguin data

- Data on 125 dives of Emperor penguins.
- Registered variables: Heart rate (bpm), Depth of dive (m), Duration of dive (min).
- Dives made by 9 penguins.



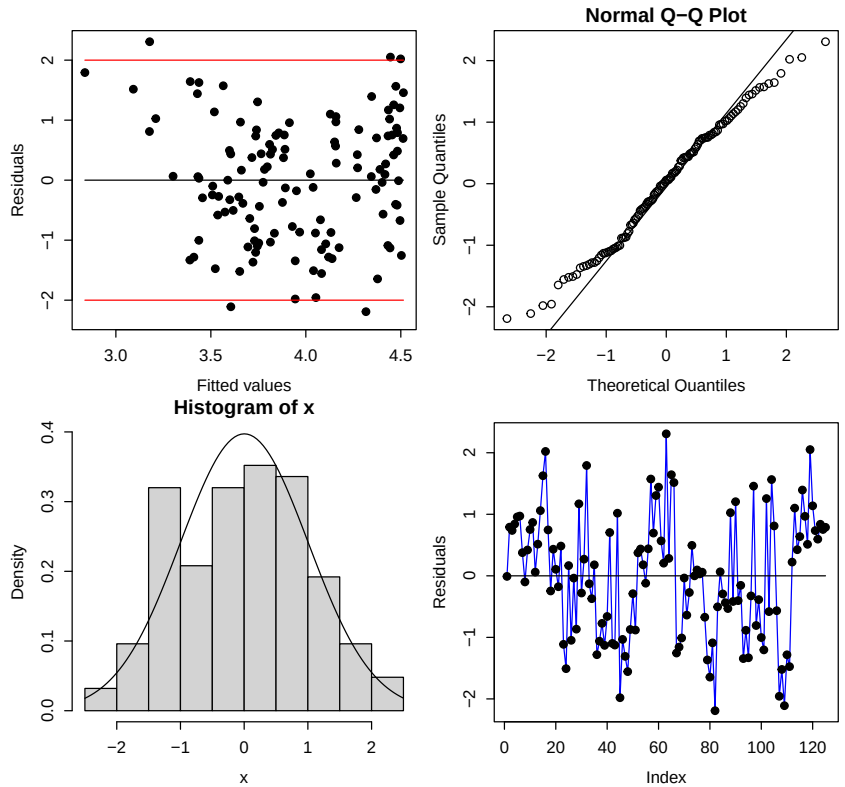
OLS regression (1/2)

```
> model.1 <- lm(Heart~Depth+Duration,data=X)
> summary(model.1)

Call:
lm(formula = Heart ~ Depth + Duration, data = X)

Residuals:
    Min      1Q  Median      3Q     Max
-30.259  -9.861  -1.158   8.389  44.070

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 95.34639    2.60156  36.650  < 2e-16 ***
Depth        0.07733    0.02865   2.699  0.00793 **
Duration    -5.85600    0.33571 -17.444  < 2e-16 ***

Residual standard error: 13.77 on 122 degrees of freedom
Multiple R-squared:  0.7314,Adjusted R-squared:  0.7269
F-statistic: 166.1 on 2 and 122 DF,  p-value: < 2.2e-16
```

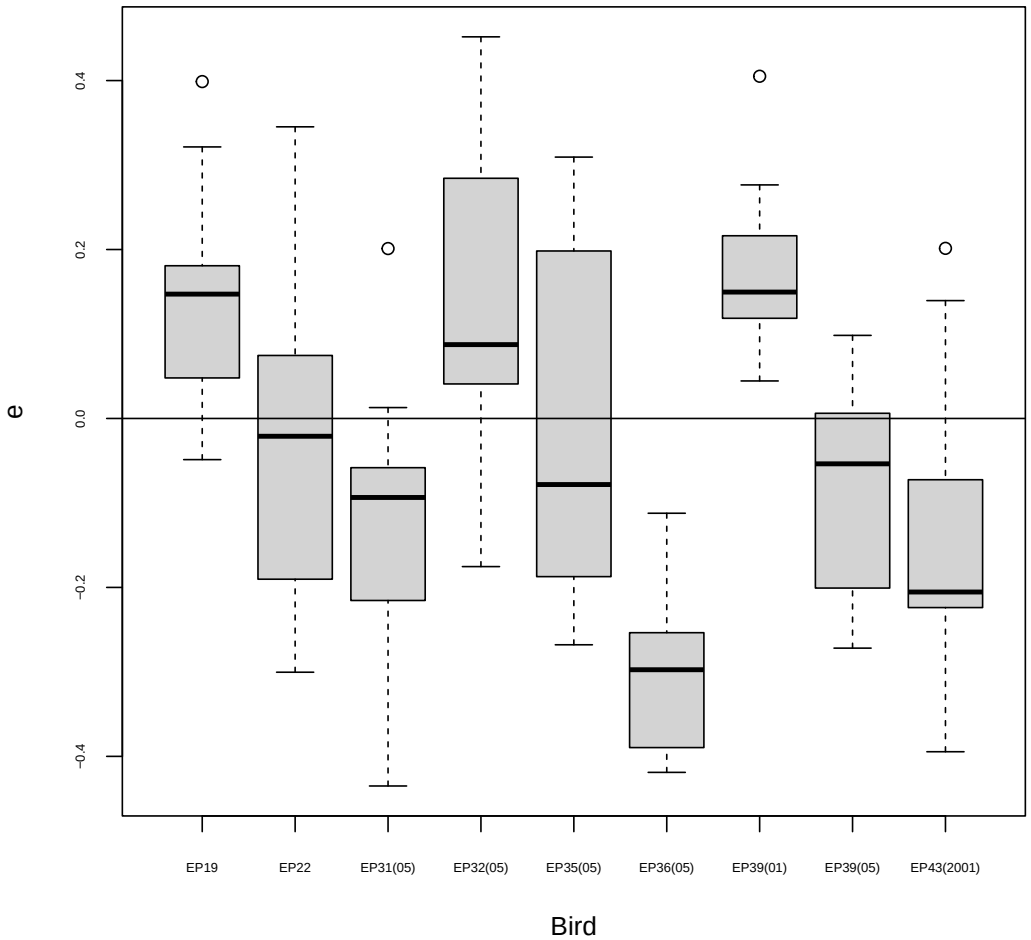


OLS regression (2/2)

```
> summary(model.1)

Call:
lm(formula = logHeart ~ Depth + Duration, data = X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43507 -0.17164  0.01166  0.14850  0.45176

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.5870524  0.0378152 121.302  < 2e-16 ***
Depth        0.0020416  0.0004164   4.903 2.94e-06 ***
Duration    -0.1038470  0.0048798 -21.281  < 2e-16 ***

Residual standard error: 0.2001 on 122 degrees of freedom
Multiple R-squared:  0.7955,Adjusted R-squared:  0.7921
F-statistic: 237.3 on 2 and 122 DF,  p-value: < 2.2e-16
```



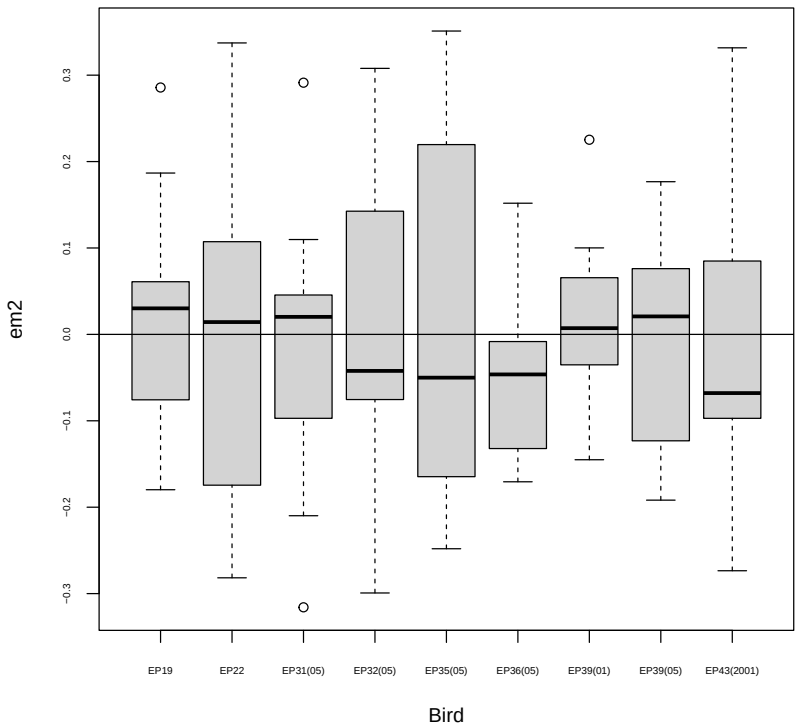
Residuals by Penguin

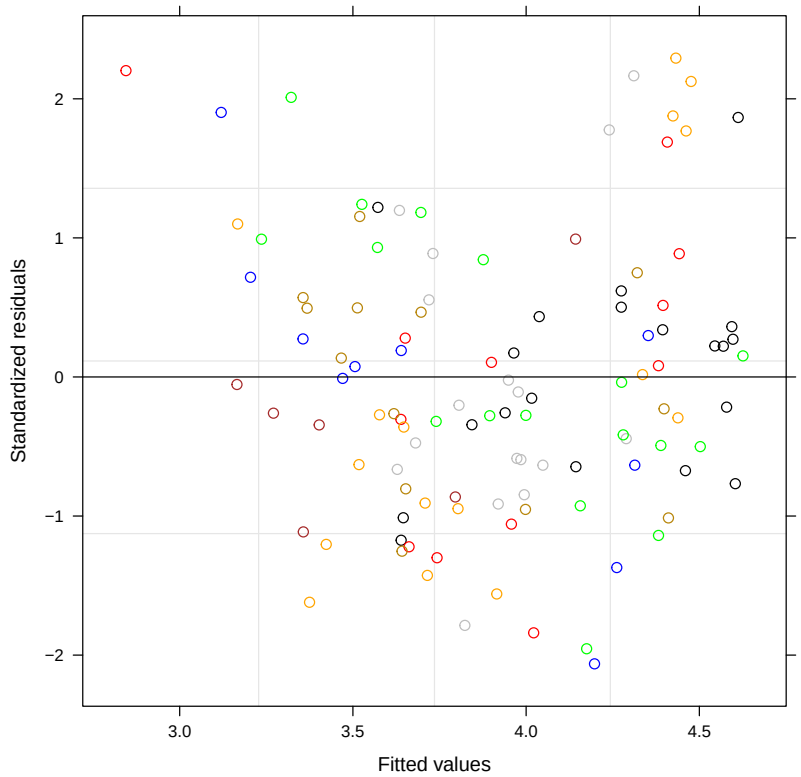


Random intercept model

```
> summary(model.2)
Linear mixed-effects model fit by REML
  Data: X
       AIC       BIC  logLik
  -56.7974 -42.77729 33.3987

Random effects:
 Formula: ~1 | Bird
        (Intercept)  Residual
StdDev:   0.1480502 0.1531288

Fixed effects:  logHeart ~ Depth + Duration
                Value  Std.Error  DF   t-value p-value
(Intercept)  4.560688 0.05872885 114  77.65669   0e+00
Depth        0.001657 0.00043084 114   3.84527   2e-04
Duration    -0.100821 0.00395385 114 -25.49932   0e+00
 Correlation:
         (Intr) Depth
Depth    -0.183
Duration -0.310 -0.467

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.06274654 -0.63446745 -0.03841057  0.50174882  2.29256217

Number of Observations: 125
Number of Groups: 9
```

Residuals mixed model





Random intercept and model

```
model.3 <- lme(logHeart~Depth+Duration,data=X,random=~Depth|Bird)
> summary(model.3)
Linear mixed-effects model fit by REML
  Data: X
       AIC       BIC  logLik
  -52.7974 -33.16925 33.3987

Random effects:
 Formula: ~Depth | Bird
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev       Corr
(Intercept) 1.480502e-01 (Intr)
Depth       7.838200e-08 0
Residual    1.531288e-01

Fixed effects:  logHeart ~ Depth + Duration
                Value  Std.Error  DF   t-value p-value
(Intercept)  4.560688 0.05872885 114  77.65669   0e+00
Depth        0.001657 0.00043084 114   3.84527   2e-04
Duration    -0.100821 0.00395385 114 -25.49932   0e+00
 Correlation:
         (Intr) Depth
Depth    -0.183
Duration -0.310 -0.467

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.06274654 -0.63446745 -0.03841057  0.50174882  2.29256217

Number of Observations: 125
Number of Groups: 9
> anova(model.2,model.3)
        Model df      AIC       BIC  logLik   Test      L.Ratio p-value
model.2     1  5 -56.7974 -42.77729 33.3987
model.3     2  7 -52.7974 -33.16925 33.3987 1 vs 2 1.468885e-08       1
>
```

Random intercept and model

```
model.4 <- lme(logHeart~Depth+Duration,data=X,random=~Duration|Bird)
> summary(model.4)
Linear mixed-effects model fit by REML
  Data: X
        AIC       BIC   logLik
  -68.40076 -48.77261 41.20038

Random effects:
 Formula: ~Duration | Bird
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 0.18477344 (Intr)
Duration    0.01734109 -0.621
Residual    0.13638843

Fixed effects:  logHeart ~ Depth + Duration
                Value  Std.Error  DF   t-value p-value
(Intercept)  4.556507 0.06845923 114  66.55797       0
Depth        0.001943 0.00042340 114   4.58870       0
Duration    -0.102910 0.00701107 114 -14.67815       0
 Correlation:
         (Intr) Depth
Depth    -0.116
Duration -0.604 -0.304

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-2.025972345 -0.717177975 -0.006164678  0.509625140  2.544984792

Number of Observations: 125
Number of Groups: 9
> anova(model.2,model.4)
         Model df       AIC       BIC   logLik   Test  L.Ratio p-value
model.2      1  5 -56.79740 -42.77729 33.39870
model.4      2  7 -68.40076 -48.77261 41.20038 1 vs 2 15.60336   4e-04
```

References

- Pinheiro, J. C. & Bates, D. M. (2000) Mixed-effects models in S and S-plus, Springer, New York

Exercise

The dataset `Oxboys` in package `nlme` contains height, subject number and standardized age for boys from Oxford. We will use linear and mixed models to study the relationship between height and age.

1. Load the file `oxford.rda` into the R environment. The data first has to be formatted for its use by the functions of the `nlme` package. This can be done with the instruction:
   `oxford <- groupedData(height age | subject, data=X)`
2. Use the instructions `class(oxford)`, `formula(oxford)` and `colnames(oxford)` to see the required structure of the data.
3. How many height measurements were made on each boy?
4. Use the instruction `plot(oxford)` to inspect the data. Do you think there is evidence for variability in intercept and growth rates?
5. Do the regression of height on age. Make standard plots of the residuals of the regression (histogram, residuals versus fitted values, normal probability plot). Do you observe any problems?
6. Make boxplots of the residuals for each boy. Do you observe any problems?
7. Do separate regressions for each boy using the `lmList` instruction. Extract the intercepts and the slopes, and make a boxplof of each. Do you think intercepts and slopes vary significantly across boys?
8. Create all 95% confidence intervals for the intercepts and the slopes, using the `intervals` function. Display all intervals in a graph. Do you think intercepts and slopes vary significantly across boys?
9. Fit a random intercept model to the data with `lme`. Use the output to obtain an estimate of the intraclass correlation coefficient.
10. Compare the regression model with the random intercept model. Which model fits the data better?
11. Fit a mixed model with random intercept and random slope. Does this model fit better than a random intercept only model?
12. Investigate the residuals of this model. Make boxplots of the residuals per individual. What do you observe?
13. Make a normal probability plot of the residuals. Is the normality assumption reasonable?
14. Plot the fitted model for each individual with the `plot(augPred(model))` instruction. Does the linear model fit well for all boys?
15. Plot the residuals of the model per subject as a function of age (`plot(model,resid(.) age|subject)`). What do you observe?
16. Use `lmList` again to fit quadratic regressions for each boy and plot their confidence intervals.
17. Fit a new model with a quadratic as a fixed effect, and another one including it as a random effect. Assess the residuals and the fit of these models. What is your final model for the data?