Assignment 2 Lachlan Bassi 22975276

Introduction

This report presents an analysis of the dataset presented in the 2005 paper *Multilevel and Longitudinal modelling using Stata by Sophia Rabe-Hesketh and Anders Skondal* that contains information about the weight of children from Asian families. The dataset includes the child's unique identifier (ID), integer indicating which occasion/visit the measurement took place (OCC), age at measurement (AGE), weight of child in kilograms (WEIGHT), birth weight (BRTHWT), and gender either boy or girl (GENDER). The aim of the analysis is to investigate how the response variable, weight of children, develops as they grow older. The feature variables of interest will be gender, age, and ID.

Methodology

First the data was imported and then a line graph for both genders for each id as well as a boxplot for both genders was generated to visualise the data. They analysed the relationship of weight against age for boys and girls. From here an initial linear mixed effects model was fitted.

$$\mathbf{weight_{ti} = \beta_0 + \beta_1 \times age_{ti} + \beta_2 \times age^2_{ti} + \beta_3 \times gender_i + \beta_4 \times gender_i \times age_{ti} + u_{0i} + u_{1i} \times age_{ti} + \varepsilon_{ti}}$$

The model is fitted in R with a residual covariate structure using varIdent. It specifies the variance of the weight measurements is modelled as a linear function of gender. This allows for different variances of the weight measurements to be estimated for boys and girls rather than assuming the variances are equal across both genders.

From here a hypothesis test was conducted to determine if the random effects structure in the linear mixed effects model is needed for age. The null hypothesis assumes that random effects for age is 0. The alternative hypothesis is that the random effect for age is non-zero. To test this hypothesis an anova test was carried out and the p-value was used to evaluate the performance, a value less than 0.05 was considered significant.

Next another hypothesis test was conducted to determine if the weight function is specified as heterogeneous with respect to gender. The null hypothesis is that this weight function is not a significantly better fit with the weight structure. The alternative hypothesis is that the weights for the residual variance are unequal for boys and girls, indicating the weight function is heterogeneous with respect to gender.

Additionally, a hypothesis test was conducted to analyse whether the mean structure could be simplified by conducting a hypothesis test to determine if the gender*age variable should be removed. Null hypothesis: there will not be a significant difference when removing the variable from the model. Alternative hypothesis: there will be a significant difference when removing the variable from the model.

After the hypothesis tests, the following diagnostic plots were created: residuals vs fitted values, Q-Q plot of residuals, residuals vs fitted values for each gender, scale-location plot of fitted values vs the square root of residuals, and an autocorrelation plot of residuals.

Results

The first hypothesis test, which compared the random effects structure of age, found that there was a significant difference when the random effect was removed. Comparing the full (including the random effect) and reduced models (not including the random effect) using an ANOVA test in R, the output had a p-value of <0.0001, which is much smaller than the significance threshold of 0.05. This result indicates a significant difference. The model including the random effects was determined to provide a better fit due to its smaller AIC (523 < 545) and smaller BIC (556 < 571) values and a larger log-likelihood value (-251 > -265).

The next hypothesis test was measuring if including a weights structure for gender improved the model. The initial hypothesis test compared a reduced model with no weights applied to the original model that has weights = varIdent(form = ~1 | gender) which allows for different residual variances for boys and girls while assuming that these variances are constant across the levels of the other covariates in the model.

The p-value was 0.0901 implying the null hypothesis was true meaning there was no difference between the two models and because of this the simpler model was selected, reduced model, as it had one less variable. This hypothesis test was then carried out with two other weight structures against this reduced model, *varExp(form = ~ age | gender) and varPower(form = ~ age | gender),* and both were very significant with p-values (6e-4) and (7e-4) respectively. These two models still address the same hypothesis test which aims to determine whether different residual variances for gender improves the model fit.

Since both varExp and varPower were significant and both possess smaller AIC and BIC values then the model with no weights structure. They were compared to each other and varExp was selected due to its slightly smaller AIC value (513.24 < 513.45) and BIC value (549.41 < 549.62) and a slightly bigger log likelihood (-245.6 > 245.7) indicating it is a slightly better fit.

The final hypothesis test, which tested the significance of the interaction term of age and gender proved the null hypothesis true as the p-value was 0.0593 which is greater than the threshold value of 0.05. Because of this there is no significant difference between the two models and so the simpler model was selected, the model without the interaction term between gender and age as it had one less variable.

The final model equation is: **3.67 + 7.90*age + -1.75 * age^2 -0.495*gender_girl + b0_id +b1_id * age + error**

Weight difference between boys and girls at 1 is 0.495114 and at 2 is 0.495114 using the fixed effects.

Discussion

Analysing the diagnostic plots generated, the residuals vs fitted plot shows the majority of the residual falling in the range of -1 and 2 which is an ideal range implying low variance. There does appear to be some grouping which can be seen in the residual's vs fitted and the square root-residuals vs fitted values plots which implies there may be subgroups within the data. These sub-groups should be investigated in further analysis potentially through the use of a different functional form. Another way to investigate these subgroups would be to see if the variables occ and brthwt are significant as they were left out of the analysis. The normal Q-Q plot of residuals mostly follows a normal distribution but shows the larger theoretical quantities at the far-right end start to deviate up from the line meaning they have larger values than expected. This implies the residuals are positively skewed and have a heavier right tail.

Looking at the summary statistics all the fixed effects coefficients are significant as they all have p-values much less than 0.05. The positive coefficient for age, 7.90 implies that as age increases by a year then weight increases by 7.90kg. The negative coefficient for age squared being -1.75 suggests that the rate of weight increase with age slows down over time at a rate of -1.75kg per year. Additionally, the negative coefficient for the gender effect, gendergirl = -0.496 implies that on average girls tend to weigh 490 grams less than boys.

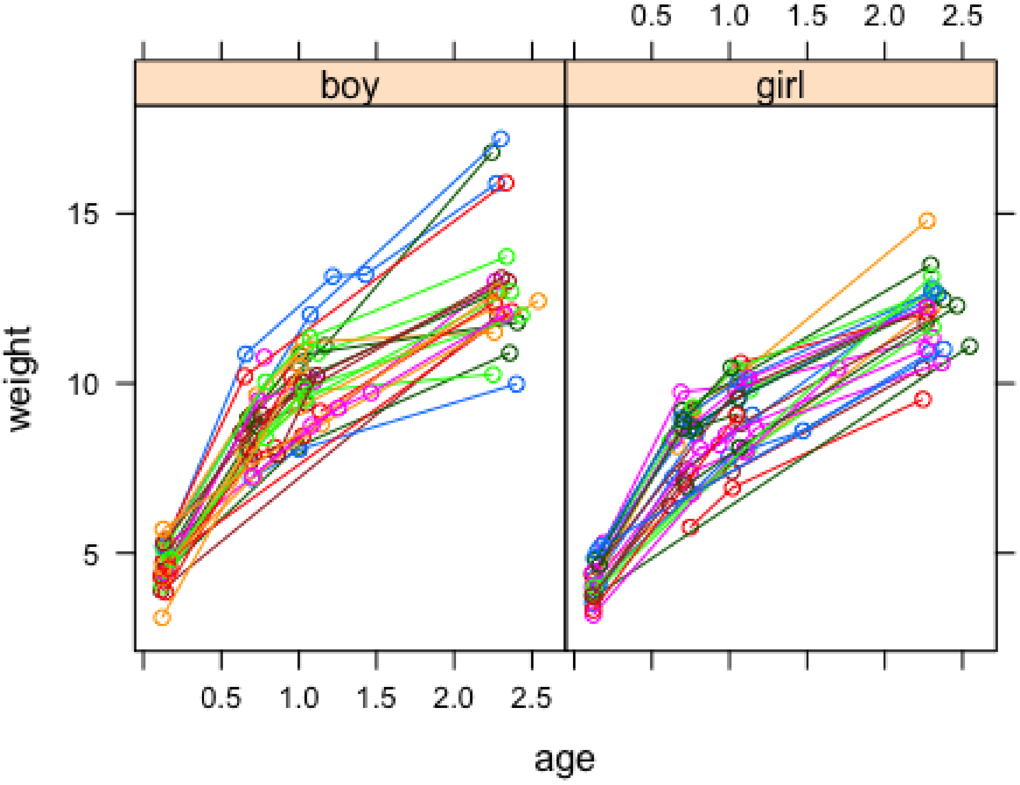
The random effect's structure calculates the standard deviation between id's for the intercept (0.521) and age (0.393). These standard deviation values imply there is considerable variability in both the initial weight and the rate that weight increases across Asian individuals. The positive correlation between random intercept and random slope (0.674) suggests that individuals with higher initial weights tend to have steeper weight increase slopes over time.

Finally, the variance functions parameters for boys are 0.456 and girls is 0.301 implying that the residual variances vary between genders with boys having higher residual variance then girls. This can also be visualised in the xyplot of the residuals and fitted values of boys and girls with the boys having more scattered residuals. Finally, the bars in the auto correlation plot all mostly fall within the blue shaded region implying the residuals are not significantly autocorrelated which implies a good fit.

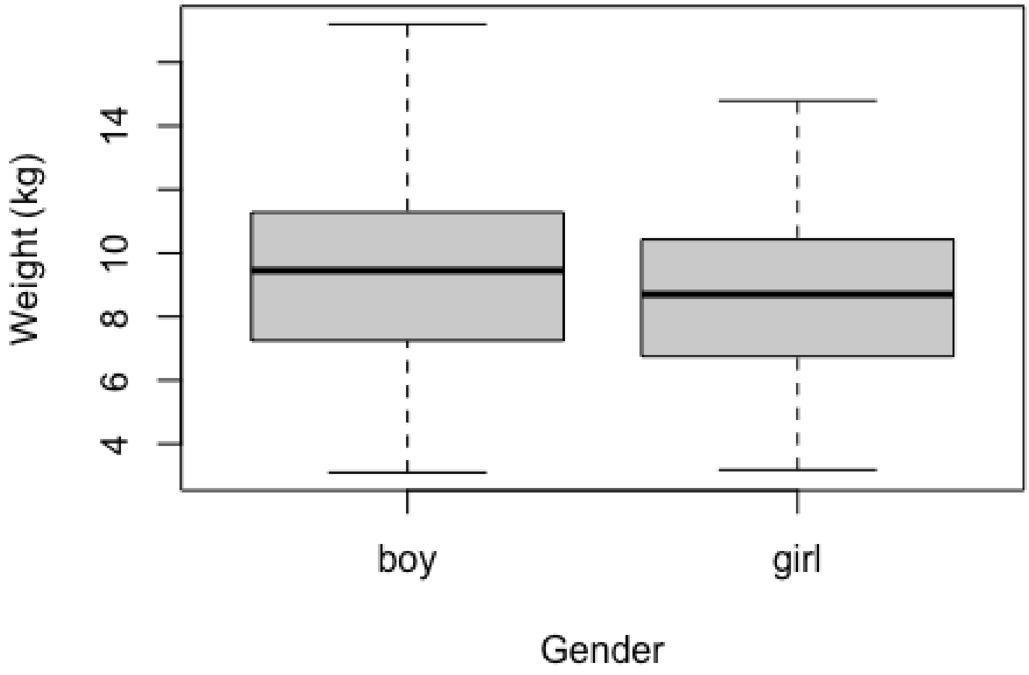
Assignment 2

2023-04-26

```
library(nlme)

load("asian.rda")

#visualise data


library(lattice)
xyplot(weight~age|gender, groups=id, dat=asian, type="b")
```



```
boxplot(asian$weight ~ asian$gender, xlab="Gender", ylab="Weight (kg)")
```



```
library(nlme)

model = lme(weight ~ age + I(age^2) + gender + gender:age,
                  random = ~ 1 + age | id, weights = varIdent(form = ~1 | gender
),
                  data = asian)

summary(model)

## Linear mixed-effects model fit by REML
##   Data: asian
##        AIC      BIC    logLik
##   535.1032 567.7301 -257.5516
##
## Random effects:
##  Formula: ~1 + age | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 0.6405865 (Intr)
## age         0.4880728 0.146
## Residual    0.6413159
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | gender
##  Parameter estimates:
##       boy      girl
## 1.0000000 0.7765769
## Fixed effects:  weight ~ age + I(age^2) + gender + gender:age
##                     Value  Std.Error  DF    t-value p-value
## (Intercept)      3.769710 0.18085990 127  20.843258  0.0000
## age              7.754957 0.25311341 127  30.638269  0.0000
## I(age^2)        -1.636271 0.08656815 127 -18.901542  0.0000
## gendergirl      -0.510547 0.21636454  66  -2.359660  0.0213
## age:gendergirl  -0.220595 0.17179912 127  -1.284026  0.2015
##  Correlation:
##                 (Intr) age    I(g^2) gndrgr
## age             -0.563
## I(age^2)         0.457 -0.863
## gendergirl      -0.658  0.135  0.007
## age:gendergirl   0.249 -0.378  0.002 -0.331
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -1.95378355 -0.45893134 -0.07457757  0.46065239  2.93290629
##
## Number of Observations: 198
## Number of Groups: 68
```

```
# Full model
full_model <- lme(weight ~ age + I(age^2) + gender + gender:age,
                  random = ~ 1 + age | id,
                  data = asian, weights = varIdent(form = ~1 | gender),
                  method = "ML")

# Reduced model without random age effect
reduced_model <- lme(weight ~ age + I(age^2) + gender + gender:age,
                     random = ~ 1 | id,
                     data = asian, weights = varIdent(form = ~1 | gender),
                     method = "ML")

# Likelihood ratio test
anova(full_model, reduced_model)
```

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## full_model        1 10 523.1175 556.0002 -251.5588
## reduced_model     2  8 545.0594 571.3656 -264.5297 1 vs 2 25.94191  <.0001
```

```
full_model <- lme(weight ~ age + I(age^2) + gender + gender:age,
                  random = ~ 1 + age | id,
                  data = asian, weights = varIdent(form = ~1 | gender),
                  method = "ML")
no_weights_model <- lme(weight ~ age + I(age^2) + gender + gender:age,
                  random = ~ 1 + age | id,
                  data = asian,
                  method = "ML")



anova(full_model, no_weights_model)
```

```
##                   Model df      AIC      BIC    logLik   Test L.Ratio p-value
## full_model            1 10 523.1175 556.0002 -251.5588                       
## no_weights_model      2  9 523.9897 553.5841 -252.9949 1 vs 2  2.8722  0.0901
```

```
varExp_model <- lme(weight ~ age + I(age^2) + gender + gender:age,
                    random = ~ 1 + age | id,
                    weights = varExp(form = ~ age | gender),
                    data = asian,
                    method = "ML")

varPower_model <- lme(weight ~ age + I(age^2) + gender + gender:age,
                      random = ~ 1 + age | id,
                      weights = varPower(form = ~ age | gender),
                      data = asian,
                      method = "ML")

anova(no_weights_model, varExp_model)
```

```
##                   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## no_weights_model      1  9 523.9897 553.5841 -252.9949                        
## varExp_model          2 11 513.2407 549.4117 -245.6204 1 vs 2 14.74897   6e-04
```

```
anova(no_weights_model, varPower_model)
```

```
##                   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## no_weights_model      1  9 523.9897 553.5841 -252.9949                        
## varPower_model        2 11 513.4505 549.6214 -245.7252 1 vs 2 14.53923   7e-04
```

```
anova(varExp_model, varPower_model)
```

```
##                 Model df      AIC      BIC    logLik
## varExp_model        1 11 513.2407 549.4117 -245.6204
## varPower_model      2 11 513.4505 549.6214 -245.7252
```

```
# Model without gender:age interaction
no_interaction_model <- lme(weight ~ age + I(age^2) + gender,
                            random = ~ 1 + age | id,
                            weights = varExp(form = ~ age | gender),
                            data = asian,
                            method = "ML")

# Likelihood ratio test
anova(varExp_model, no_interaction_model)
```

```
##                       Model df      AIC      BIC    logLik   Test  L.Ratio
## varExp_model              1 11 513.2407 549.4117 -245.6204                
## no_interaction_model      2 10 514.7966 547.6793 -247.3983 1 vs 2 3.555898
##                       p-value
## varExp_model                 
## no_interaction_model   0.0593
```

```
final_model = no_interaction_model
summary(final_model)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: asian
##        AIC      BIC    logLik
##   514.7966 547.6793 -247.3983
##
## Random effects:
##  Formula: ~1 + age | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 0.5213937 (Intr)
## age         0.3932396 0.674
## Residual    0.4011110
##
## Variance function:
##  Structure: Exponential of variance covariate, different strata
##  Formula: ~age | gender
##  Parameter estimates:
##       boy      girl
## 0.4562713 0.3010983
## Fixed effects:  weight ~ age + I(age^2) + gender
##                 Value  Std.Error  DF   t-value p-value
## (Intercept)  3.673643 0.13988165 128  26.26251  0.0000
## age          7.899976 0.21450150 128  36.82947  0.0000
## I(age^2)    -1.750631 0.08964916 128 -19.52758  0.0000
## gendergirl  -0.495114 0.17912036  66  -2.76414  0.0074
##  Correlation:
##            (Intr) age    I(g^2)
## age        -0.408
## I(age^2)    0.388 -0.919
## gendergirl -0.633 -0.038  0.056
##
## Standardized Within-Group Residuals:
##           Min            Q1           Med            Q3           Max
## -2.416387e+00 -5.562761e-01 -4.588499e-05  5.220651e-01  2.865226e+00
##
## Number of Observations: 198
## Number of Groups: 68
```

To calculate the estimated weight difference between a boy and a girl at different ages, we can use the model equation:

Weight = β0 + β1 * age + β2 * age^2 + β3 * gender_girl + b0_id + b1_id * age + ε

For a 1-year-old boy: Weight_boy_1 = β0 + β1 * 1 + β2 * 1^2 Weight_boy_1 = 3.673643 + 7.899976 * 1 - 1.750631 * 1^2

```
Weight_boy_1 = 3.673643 + 7.899976 - 1.750631
Weight_boy_1

## [1] 9.822988
```

For a 1-year-old girl: Weight_girl_1 = β0 + β1 * 1 + β2 * 1^2 + β3 * 1 Weight_girl_1 = 3.673643 + 7.899976 * 1 - 1.750631 * 1^2 - 0.495114

```
Weight_girl_1 = 3.673643 + 7.899976 - 1.750631 - 0.495114
Weight_girl_1
```

Assignment 2 Lachlan Bassi 22975276

```
## [1] 9.327874
```

Weight difference at 1 year old:

```
Weight_boy_1 - Weight_girl_1

## [1] 0.495114
```

For a 2-year-old boy: Weight_boy_2 = β0 + β1 * 2 + β2 * 2^2 Weight_boy_2 = 3.673643 + 7.899976 * 2 - 1.750631 * 2^2

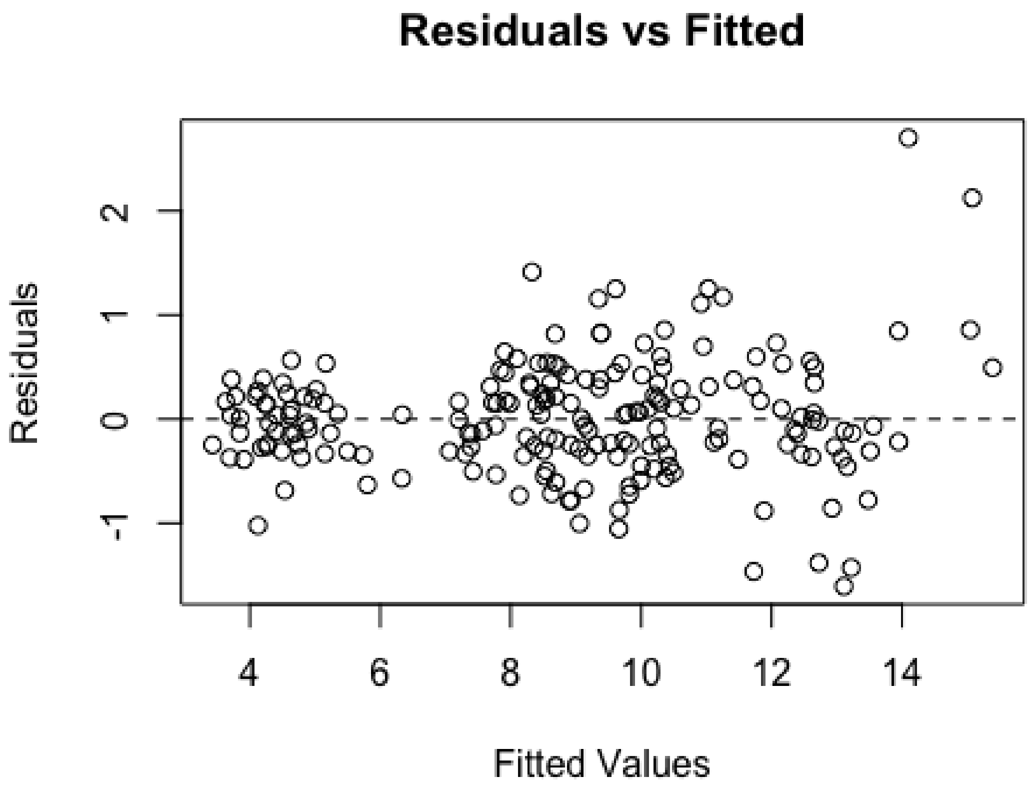
```
Weight_boy_2 = 3.673643 + 15.799952 - 14.002524
Weight_boy_2

## [1] 5.471071
```

For a 2-year-old girl: Weight_girl_2 = β0 + β1 * 2 + β2 * 2^2 + β3 * 1 Weight_girl_2 = 3.673643 + 7.899976 * 2 - 1.750631 * 2^2 - 0.495114

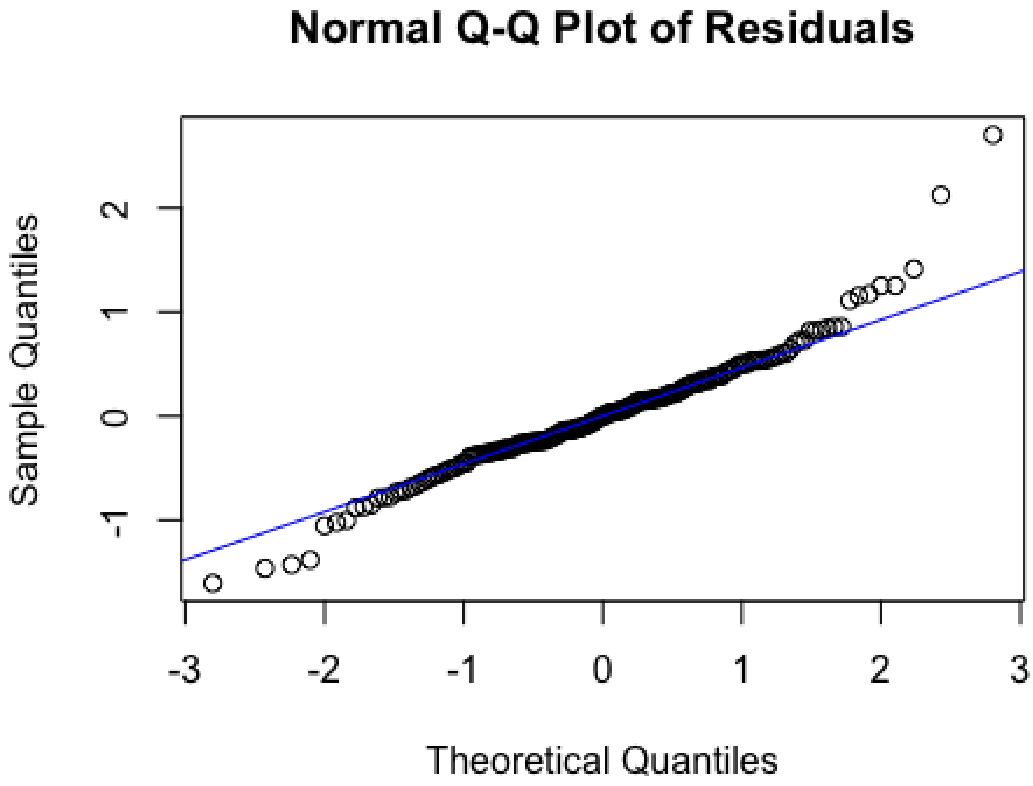
```
Weight_girl_2 = 3.673643 + 15.799952 - 14.002524 - 0.495114
Weight_girl_2

## [1] 4.975957
```

Weight difference at 2 years old:

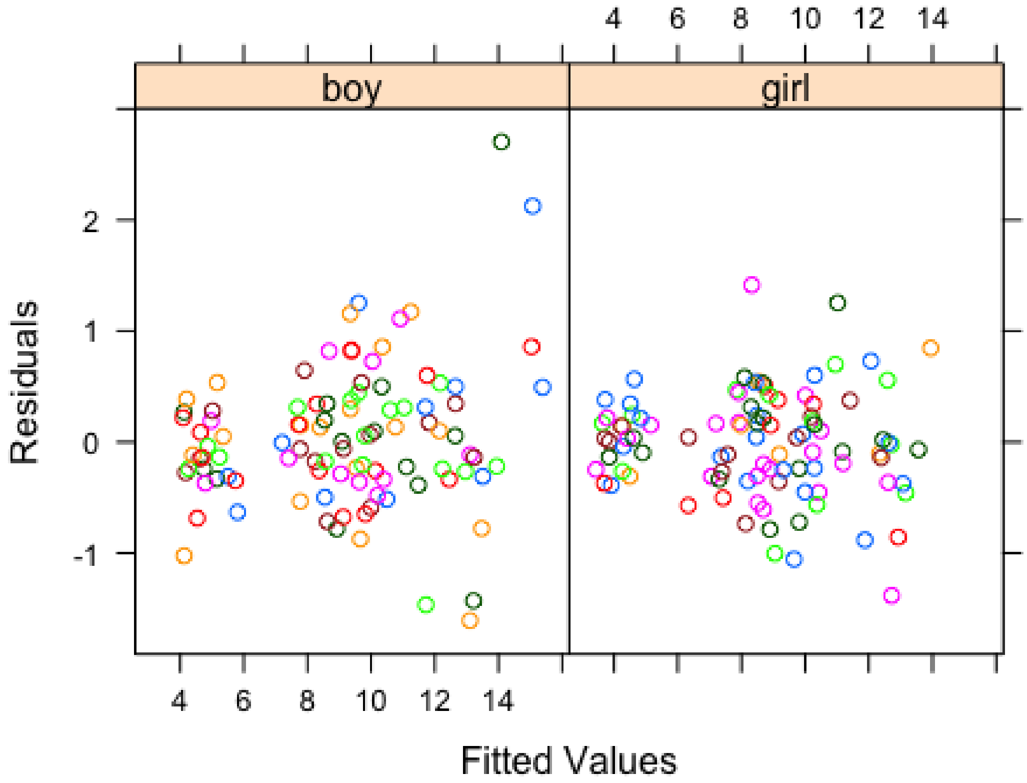
```
Weight_boy_2 - Weight_girl_2

## [1] 0.495114

fixed_coeffs <- fixef(final_model)
weight_diff_age1 <- fixed_coeffs["(Intercept)"] + fixed_coeffs["age"] * 1 + fixe
d_coeffs["I(age^2)"] * (1^2) + fixed_coeffs["gendergirl"] - (fixed_coeffs["(Inte
rcept)"] + fixed_coeffs["age"] * 1 + fixed_coeffs["I(age^2)"] * (1^2))
weight_diff_age2 <- fixed_coeffs["(Intercept)"] + fixed_coeffs["age"] * 2 + fixe
d_coeffs["I(age^2)"] * (2^2) + fixed_coeffs["gendergirl"] - (fixed_coeffs["(Inte
rcept)"] + fixed_coeffs["age"] * 2 + fixed_coeffs["I(age^2)"] * (2^2))
weight_diff_age1

## (Intercept)
##  -0.4951143

weight_diff_age2

## (Intercept)
##  -0.4951143

summary(final_model)

## Linear mixed-effects model fit by maximum likelihood
##   Data: asian
##        AIC      BIC    logLik
##   514.7966 547.6793 -247.3983
##
## Random effects:
##  Formula: ~1 + age | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr  
## (Intercept) 0.5213937 (Intr)
## age         0.3932396 0.674 
## Residual    0.4011110       
## 
## Variance function:
##  Structure: Exponential of variance covariate, different strata
##  Formula: ~age | gender 
##  Parameter estimates:
##       boy      girl 
## 0.4562713 0.3010983 
## Fixed effects:  weight ~ age + I(age^2) + gender 
##                 Value  Std.Error  DF   t-value p-value
## (Intercept)  3.673643 0.13988165 128  26.26251  0.0000
## age          7.899976 0.21450150 128  36.82947  0.0000
## I(age^2)    -1.750631 0.08964916 128 -19.52758  0.0000
## gendergirl  -0.495114 0.17912036  66  -2.76414  0.0074
##  Correlation: 
##            (Intr) age    I(g^2)
## age        -0.408              
## I(age^2)    0.388 -0.919       
## gendergirl -0.633 -0.038  0.056
## 
## Standardized Within-Group Residuals:
##           Min            Q1           Med            Q3           Max 
## -2.416387e+00 -5.562761e-01 -4.588499e-05  5.220651e-01  2.865226e+00 
## 
## Number of Observations: 198
## Number of Groups: 68

# 1. Residuals vs Fitted values plot
plot(fitted(final_model), residuals(final_model), main = "Residuals vs Fitted",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, lty = 2)
```



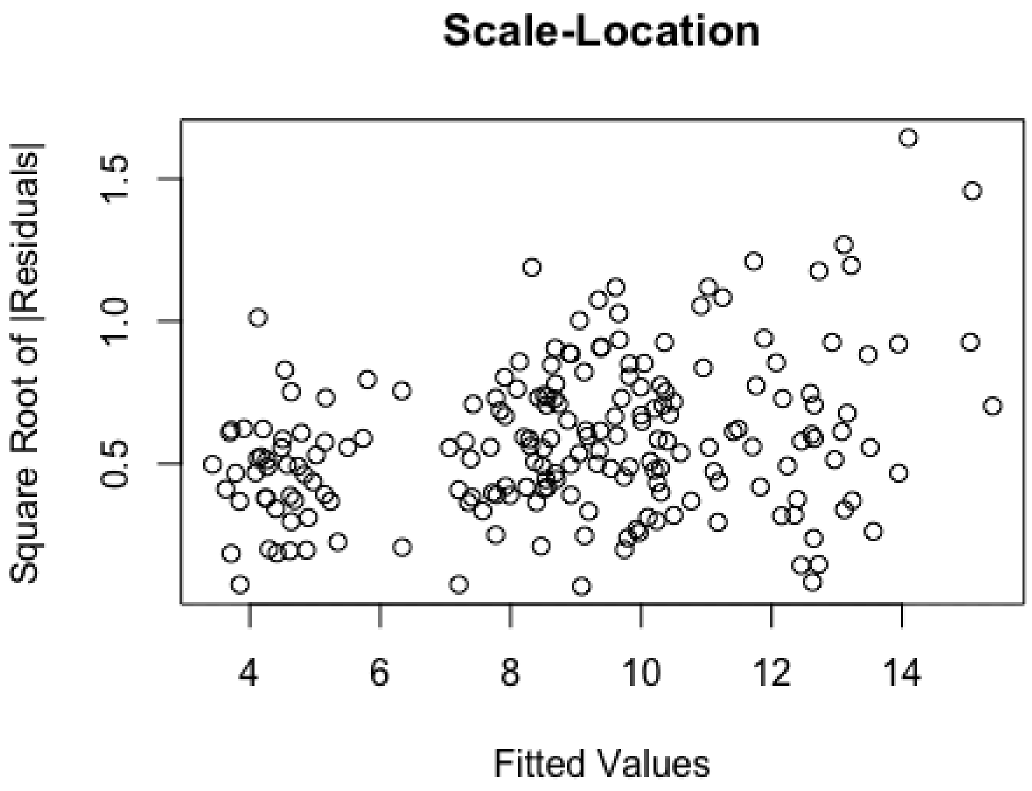
```
# 2. Normal Q-Q plot of residuals
qqnorm(residuals(final_model), main = "Normal Q-Q Plot of Residuals")
qqline(residuals(final_model), col = "blue")
```



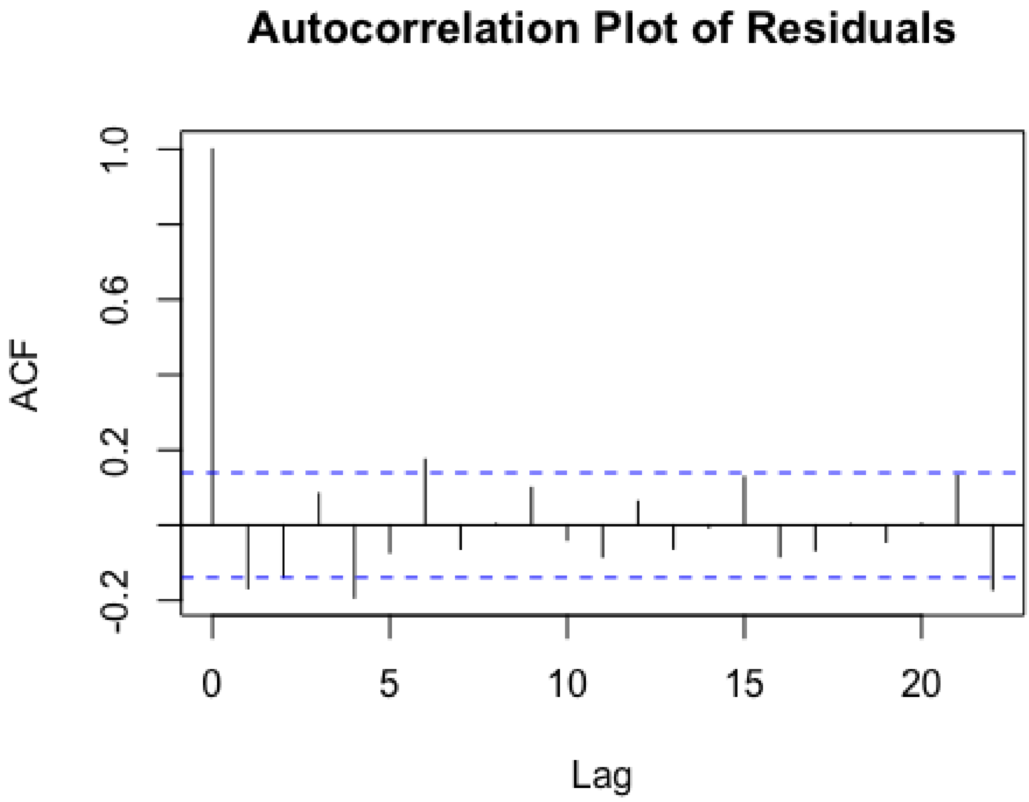
```r
xyplot(resid(final_model)~fitted(final_model)|gender, groups=id, data=asian, xlab = "Fitted Values", ylab = "Residuals")
```



```
# 3. Scale-location plot
plot(fitted(final_model), sqrt(abs(residuals(final_model))), main = "Scale-Locat
ion",
     xlab = "Fitted Values", ylab = "Square Root of |Residuals|")
abline(h = 0, lty = 2)
```



```r
# 4. Autocorrelation plot of residuals
acf(residuals(final_model), main = "Autocorrelation Plot of Residuals")
```


Autocorrelation Plot of Residuals
ACF
1.0
0.6
0.2
-0.2
0
5
10
15
20
Lag