The Dutch Diminutive: Between Inflection and Derivation

Analyis Script

Daniil Bondarenko

2023-01-15

Import the experimental dataset, perform some final cleanup

```
trialdata <- read_csv("../corpus/trialdata.csv", show_col_types = FALSE)
# Move all relevant columns to the left for better readability
trialdata <- select(trialdata,
                    "spelling":"dec_criterion",
                    "rtC":"rateC",
                    "Length":"Colt_Nphon",
                    "item",
                    "participant",
                    "lower":"rateR",
                    "rtI":"zrateI",
                    "acc.mean":"zrateI.sd"
)
# Filter out the irrelevant columns based on theory-supported decisions
# NOTE: Done here so variables could be reintroduced at will if need be
trialdata <- select(trialdata,
                    # OLD20 chosen as the neighbourhood var: filter out the rest
                    -"Colt_N",
                    -"PLD30",
                    -"Colt_Nphon",
                    # RTs chosen as the dependent var; filter out the rest,
                    # do the centering and z-Transforming within the script;
                    # filter out the pre-existing variables
                    -"lower",
                    -"upper",
                    -"rtR",
                    -"rateR",
                    -"rtI",
                    -"rateI",
                    -"zrtC",
                    -"zrateC",
                    -"zrtI",
                    -"zrateI",
                    -"acc.mean",
                    -"acc.sd",
                    -"rtC.mean",
                    -"rtC.sd",
```

```
                    -"rateC.mean",
                    -"rateC.sd",
                    -"zrtC.mean",
                    -"zrtC.sd",
                    -"zrateC.mean",
                    -"zrateC.sd",
                    -"rtI.mean",
                    -"rtI.sd",
                    -"rateI.mean",
                    -"rateI.sd",
                    -"zrtI.mean",
                    -"zrtI.sd",
                    -"zrateI.mean",
                    -"zrateI.sd"
)
```

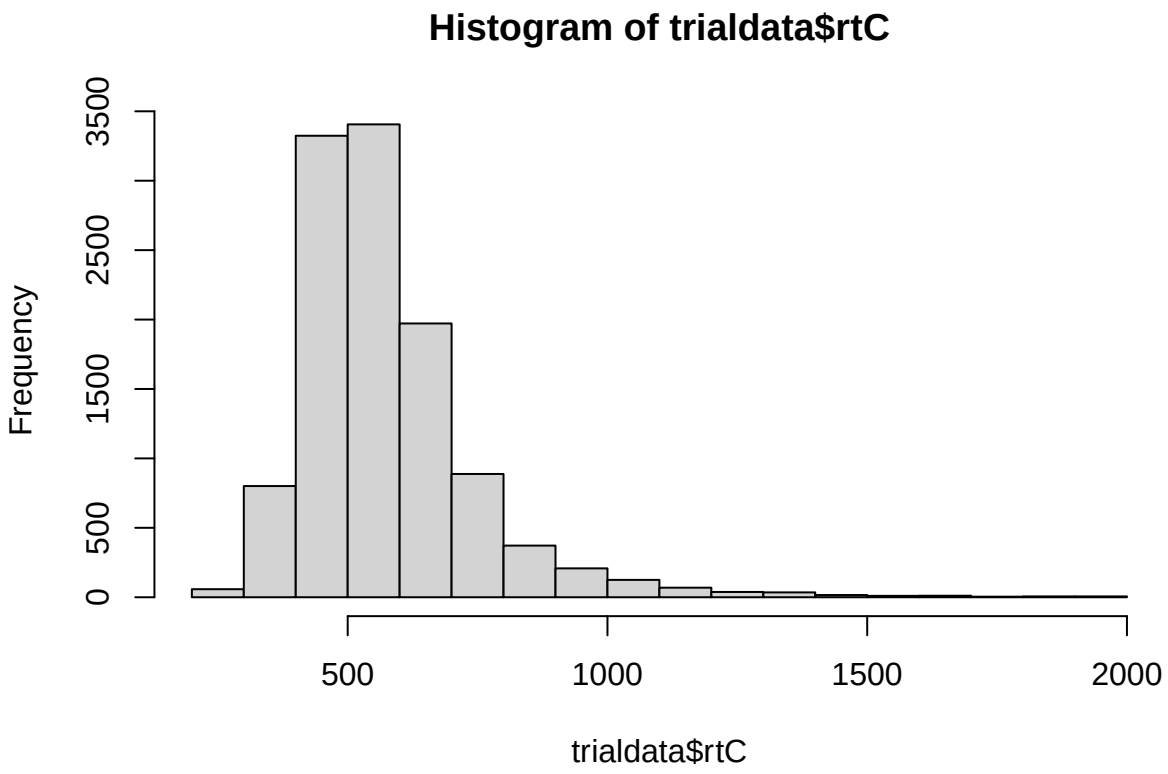
Run sanity checks: plots, correlations, etc.

```
# Check for correlations using Pearson's R
summary(trialdata$SUBTLEX2)
```

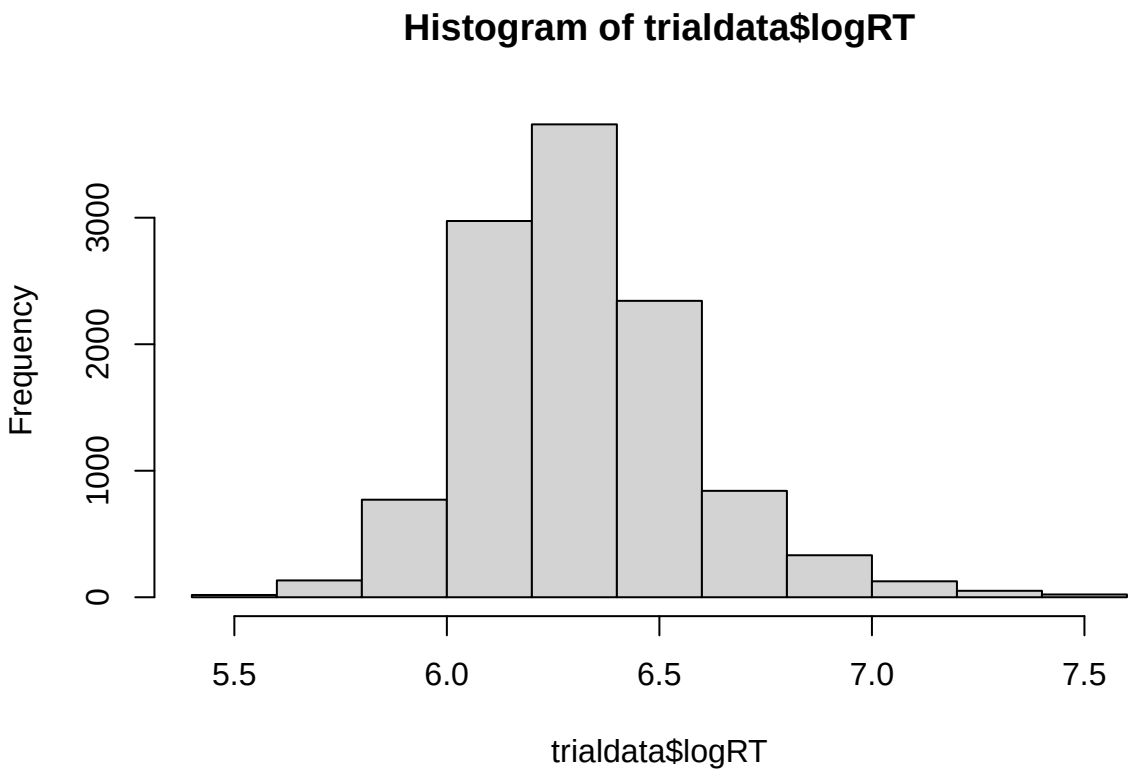
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    31.0   104.0   739.9   268.0 59247.0
```

Finalise fine preprocessing: contrasts, centering, etc.

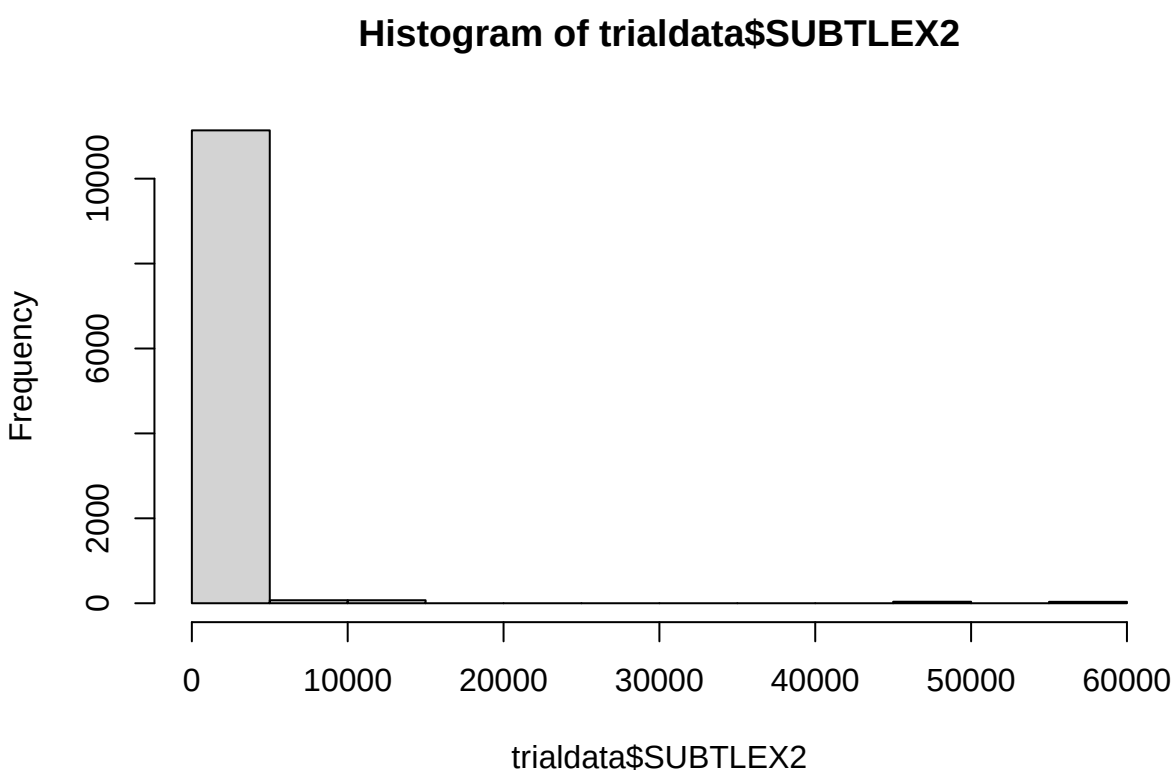
```
# Motivate the log-transformation with reference to Winter 2020 and
# Smith and Levy 2013 therein
hist(trialdata$rtC)
```



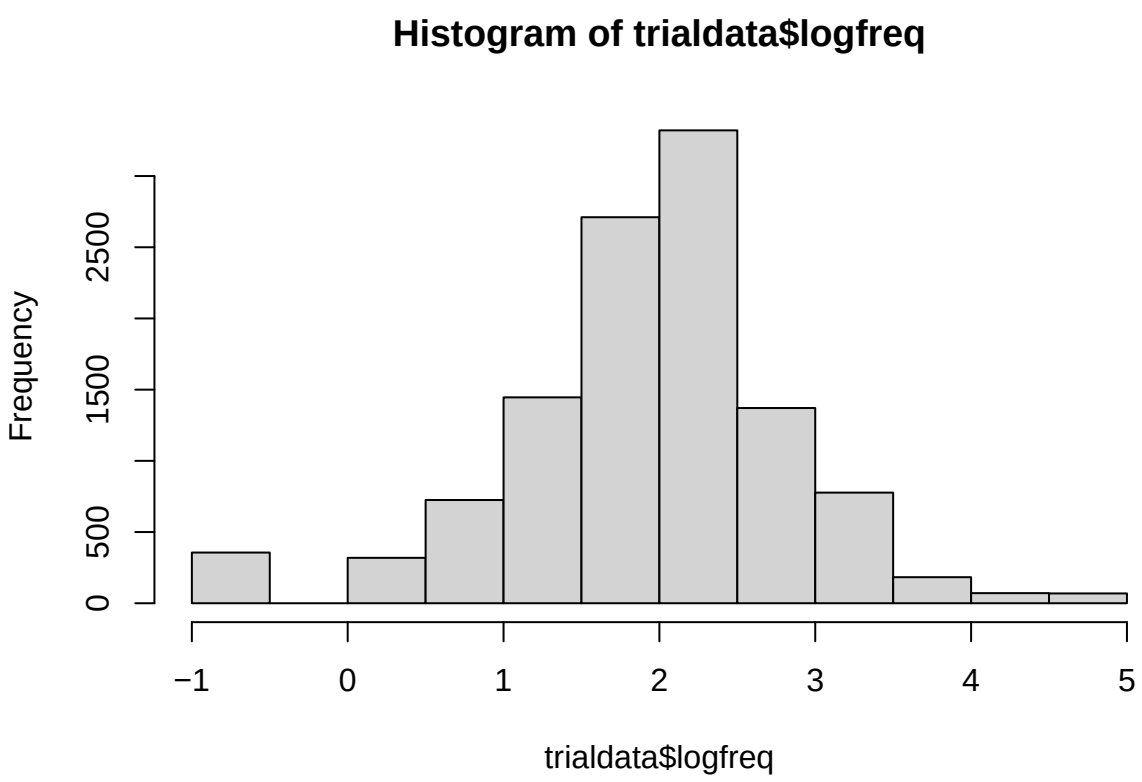

```
trialdata$logRT <- log(trialdata$rtC)
hist(trialdata$logRT)
```




```
# Center and standardize vars like word length and freq
hist(trialdata$SUBTLEX2)
```




```
trialdata$logfreq <- log10(trialdata$SUBTLEX2 + 0.1)
hist(trialdata$logfreq)
```




```
trialdata <- mutate(trialdata,
                    logfreq_z = scale(logfreq),
                    len_z = scale(Length),
                    nsyl_z = scale(Nsyl),
                    old_z = scale(OLD20),
                    conc_z = scale(Concreteness),
                    aoa_z = scale(AoA),
                    wp_z = scale(Word_prevalence1))
```

Descriptive Statistics

```
# Get means for the diminutives
trialdata %>% group_by(dim_type) %>%
  summarize(M = mean(rtC), SD = sd(rtC))
```

```
## # A tibble: 4 x 3
##   dim_type      M    SD
##   <chr>     <dbl> <dbl>
## 1 both       552.  162.
## 2 deriv      582.  177.
## 3 infl       572.  171.
## 4 undecided  573.  170.
```

```r
# Run paired t-tests, see if the differences between those means are significant
df1 <- trialdata %>%
    filter(dim_type == "both" | dim_type == "infl") %>%
    select(dim_type, rtC)
df2 <- trialdata %>%
    filter(dim_type == "infl" | dim_type == "deriv") %>%
    select(dim_type, rtC)
df3 <- trialdata %>%
    filter(dim_type == "deriv" | dim_type == "undecided") %>%
    select(dim_type, rtC)
t.test(rtC ~ dim_type, data = df1)
```

```
##
##  Welch Two Sample t-test
##
## data:  rtC by dim_type
## t = -3.0162, df = 1017.6, p-value = 0.002624
## alternative hypothesis: true difference in means between group both and group infl is not equal to 0
## 95 percent confidence interval:
##  -33.799143  -7.154739
## sample estimates:
## mean in group both mean in group infl
##           551.7251           572.2020
```

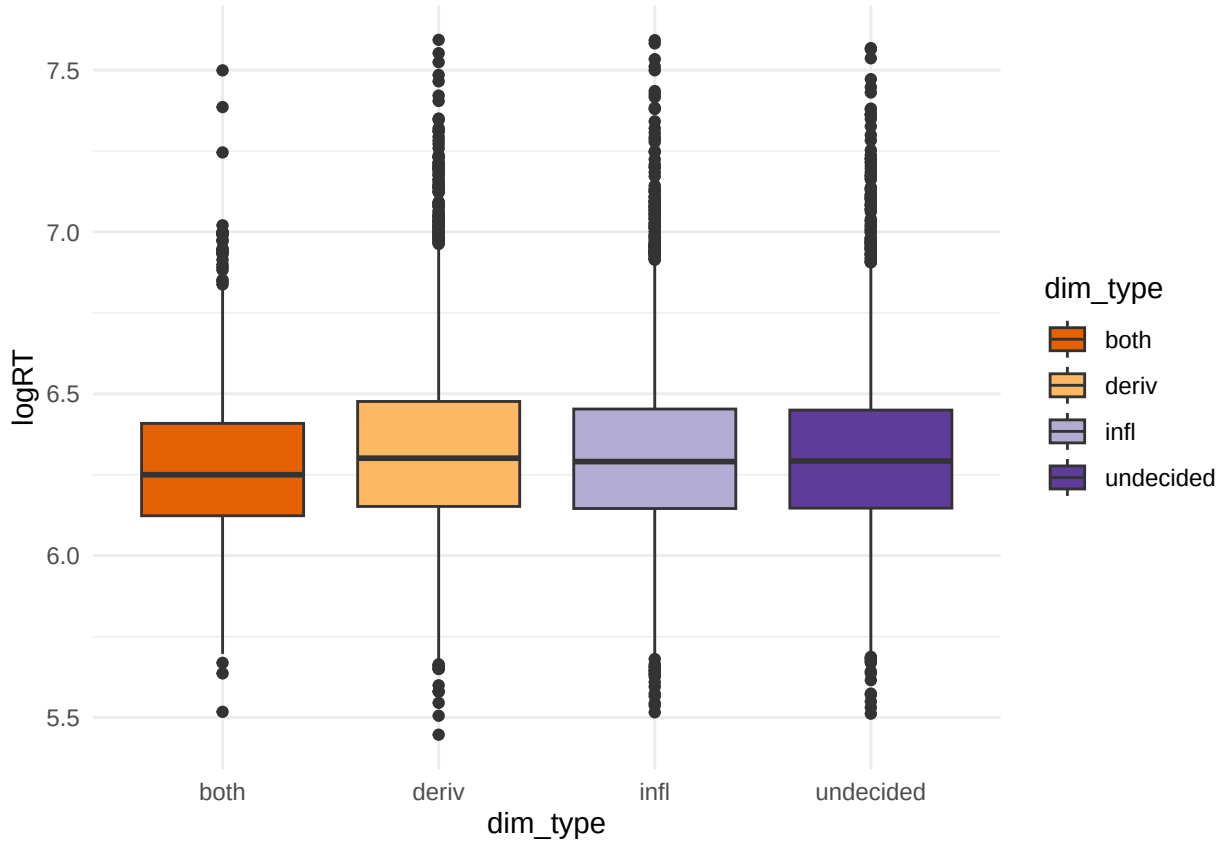
```r
t.test(rtC ~ dim_type, data = df2)
```

```
##
##  Welch Two Sample t-test
##
## data:  rtC by dim_type
## t = 2.3554, df = 7144.8, p-value = 0.01853
## alternative hypothesis: true difference in means between group deriv and group infl is not equal to 0
## 95 percent confidence interval:
##   1.625703 17.759122
## sample estimates:
## mean in group deriv  mean in group infl
##            581.8944            572.2020
```

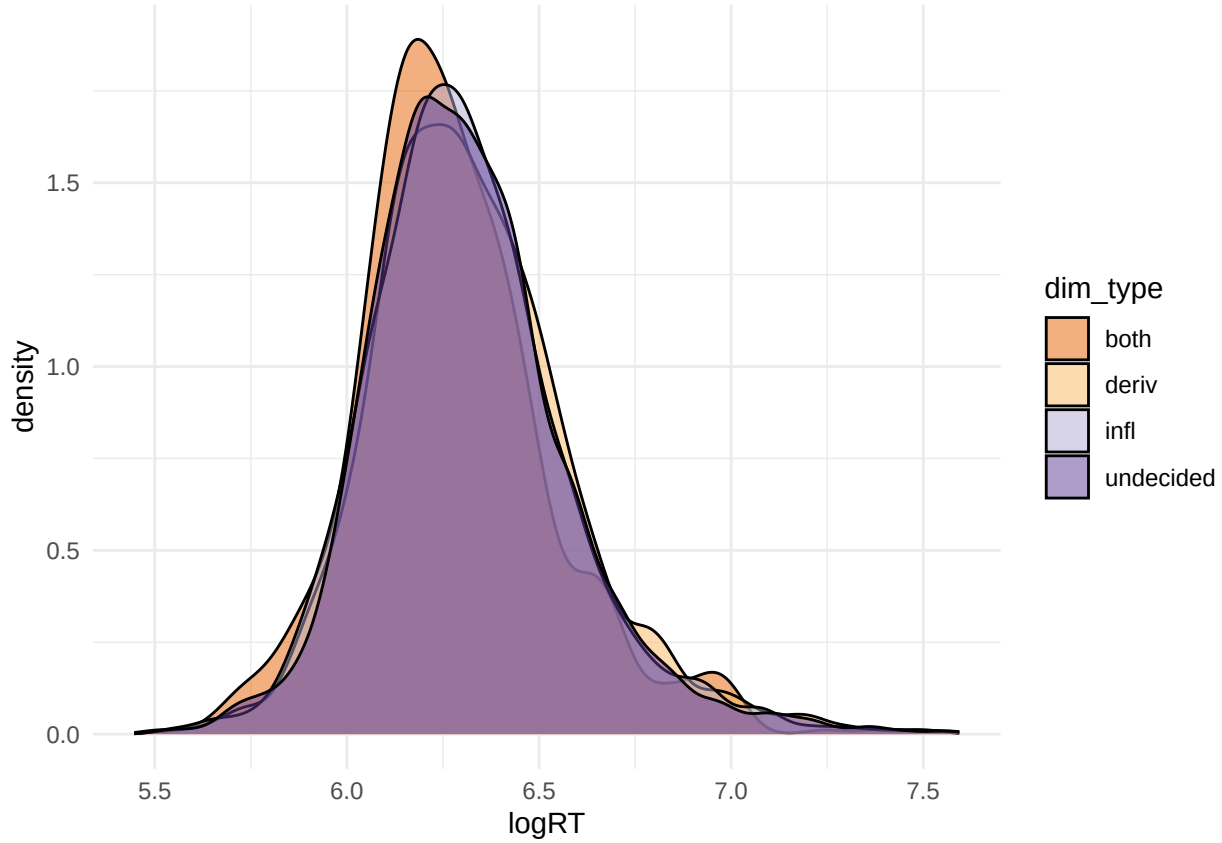
```r
t.test(rtC ~ dim_type, data = df3)
```

```
##
##  Welch Two Sample t-test
##
## data:  rtC by dim_type
## t = 2.1302, df = 7125.8, p-value = 0.03319
## alternative hypothesis: true difference in means between group deriv and group undecided is not equal
## 95 percent confidence interval:
##   0.6983034 16.8110405
## sample estimates:
##     mean in group deriv mean in group undecided
##                581.8944                573.1397
```

```
trialdata %>% ggplot(aes(x = dim_type, y = logRT, fill = dim_type)) +
  geom_boxplot() + theme_minimal() +
  scale_fill_brewer(palette = "PuOr")
```




```
ggsave("../figures/dim_box.png", width = 8, height = 6)

trialdata %>% ggplot(aes(x = logRT, fill = dim_type)) +
  geom_density(alpha = 0.5) + theme_minimal() +
  scale_fill_brewer(palette = "PuOr")
```




```
ggsave("../figures/dim_density.png", width = 8, height = 6)
```

Inferential Statistics

```
mod.base <- lmer(logRT ~logfreq_z + (1|participant), data = trialdata)
summary(mod.base)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: logRT ~ logfreq_z + (1 | participant)
##    Data: trialdata
##
## REML criterion at convergence: -3179.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9657 -0.6679 -0.1424  0.4903  6.0710
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  participant (Intercept) 0.02286  0.1512
##  Residual                0.04284  0.2070
## Number of obs: 11349, groups:  participant, 81
```

```
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  6.31641    0.01691  373.50
## logfreq_z   -0.05540    0.00195  -28.41
##
## Correlation of Fixed Effects:
##           (Intr)
## logfreq_z 0.000
```

```
mod.base.nearfull <- lmer(logRT ~
                        logfreq_z +
                        len_z +
                        nsyl_z +
                        old_z +
                        conc_z +
                        aoa_z +
                        wp_z +
                        (1|participant), data = trialdata)
summary(mod.base.nearfull)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: logRT ~ logfreq_z + len_z + nsyl_z + old_z + conc_z + aoa_z +
##     wp_z + (1 | participant)
##    Data: trialdata
##
## REML criterion at convergence: -3412.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3480 -0.6518 -0.1419  0.4797  6.3288
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  participant (Intercept) 0.02279  0.1510
##  Residual                0.04177  0.2044
## Number of obs: 11349, groups:  participant, 81
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  6.316468   0.016884 374.107
## logfreq_z   -0.027684   0.002683 -10.316
## len_z       -0.001042   0.005180  -0.201
## nsyl_z       0.002375   0.003915   0.607
## old_z        0.020328   0.004318   4.708
## conc_z       0.001193   0.002259   0.528
## aoa_z        0.015183   0.002559   5.932
## wp_z        -0.024448   0.002442 -10.011
##
## Correlation of Fixed Effects:
##           (Intr) lgfrq_ len_z  nsyl_z old_z  conc_z aoa_z
## logfreq_z  0.000
## len_z      0.000  0.062
## nsyl_z     0.000  0.059 -0.558
```

```
## old_z      0.000  0.111 -0.628 -0.124
## conc_z     0.000  0.350  0.074  0.019 -0.060
## aoa_z      0.000  0.250 -0.082  0.046 -0.072  0.382
## wp_z       0.000 -0.449 -0.079  0.006 -0.079 -0.217  0.249
```

```
mod.base.full <- lmer(logRT ~
                        logfreq_z +
                        len_z +
                        nsyl_z +
                        old_z +
                        conc_z +
                        aoa_z +
                        wp_z +
                    (1|participant) +
                    (1|spelling), data = trialdata)
summary(mod.base.full)
```

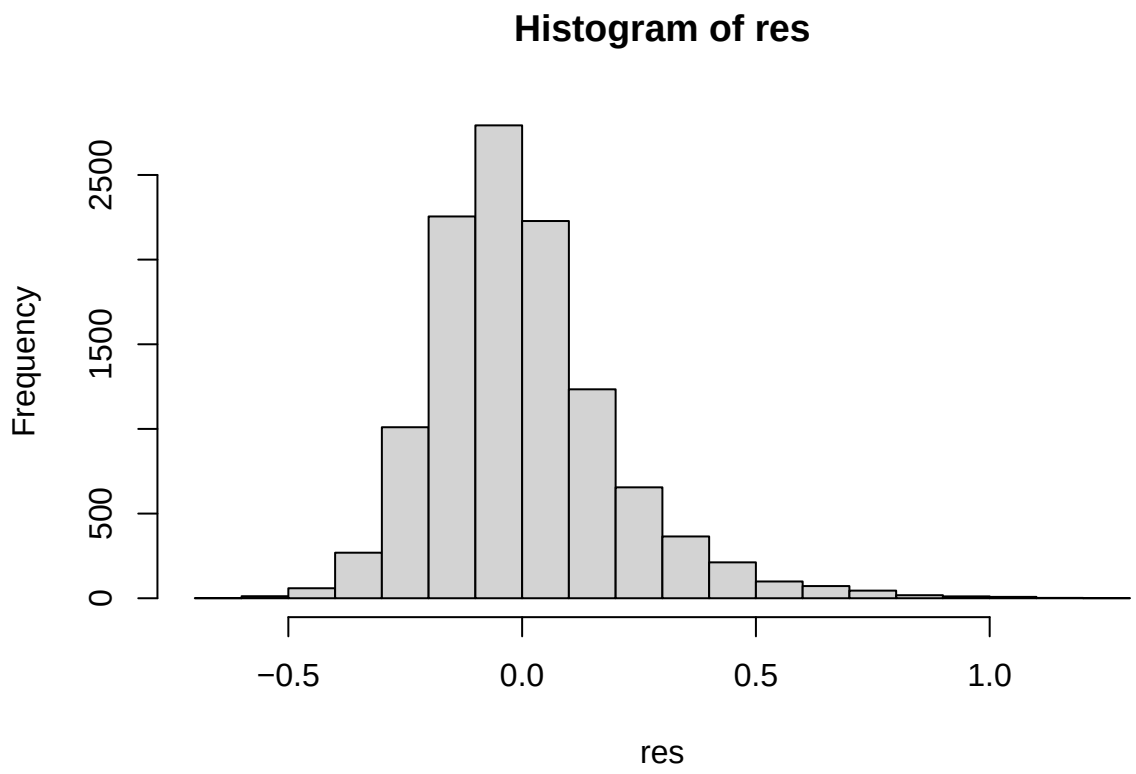
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: logRT ~ logfreq_z + len_z + nsyl_z + old_z + conc_z + aoa_z +
##     wp_z + (1 | participant) + (1 | spelling)
##    Data: trialdata
##
## REML criterion at convergence: -3597.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1805 -0.6457 -0.1366  0.4728  6.3979
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  spelling    (Intercept) 0.001809 0.04254
##  participant (Intercept) 0.022769 0.15089
##  Residual                0.040031 0.20008
## Number of obs: 11349, groups:  spelling, 310; participant, 81
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  6.317143   0.017044 370.627
## logfreq_z   -0.027390   0.004256  -6.436
## len_z       -0.001220   0.008277  -0.147
## nsyl_z       0.002013   0.006265   0.321
## old_z        0.021018   0.006895   3.048
## conc_z       0.001390   0.003610   0.385
## aoa_z        0.015867   0.004100   3.870
## wp_z        -0.024487   0.003854  -6.353
##
## Correlation of Fixed Effects:
##           (Intr) lgfrq_ len_z  nsyl_z old_z  conc_z aoa_z
## logfreq_z  0.001
## len_z      0.002  0.061
## nsyl_z     0.000  0.062 -0.560
## old_z     -0.001  0.113 -0.626 -0.122
## conc_z    -0.001  0.355  0.074  0.021 -0.060
## aoa_z      0.000  0.250 -0.079  0.043 -0.076  0.380
```

```
## wp_z       0.004 -0.457 -0.082  0.004 -0.078 -0.224  0.255
```

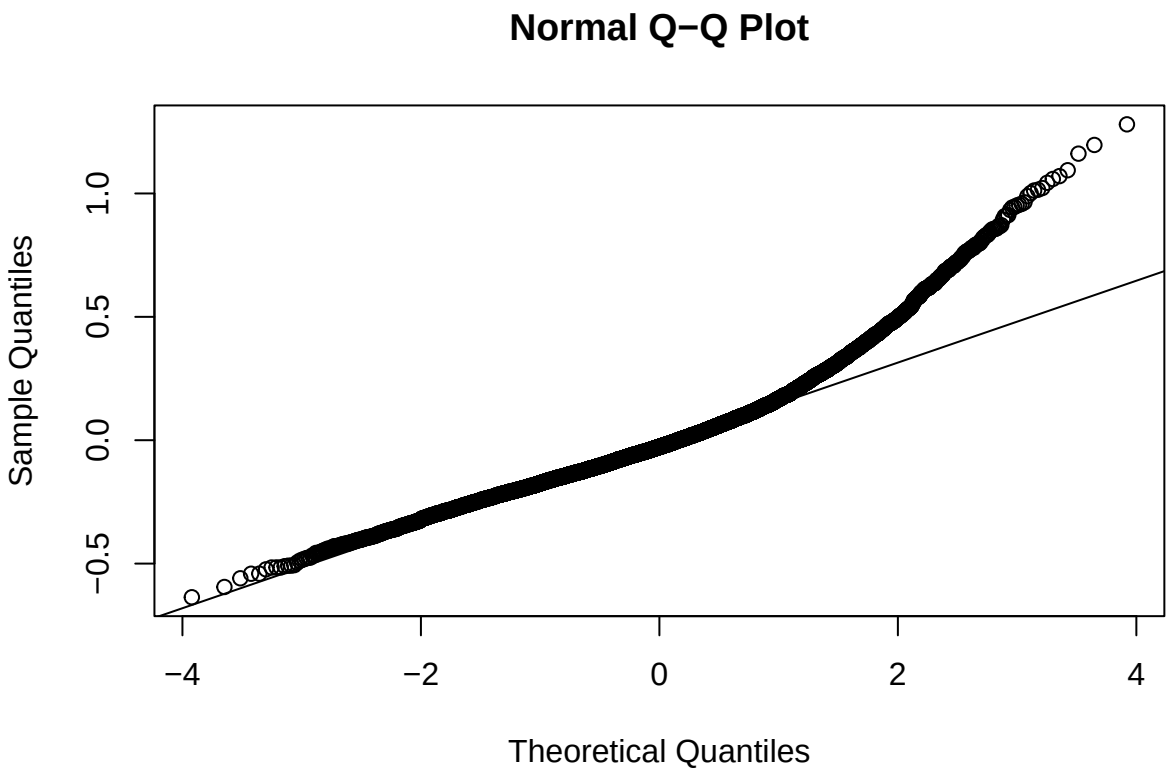
```
anova(mod.base.nearfull, mod.base.full)
```

```
## refitting model(s) with ML (instead of REML)

## Data: trialdata
## Models:
## mod.base.nearfull: logRT ~ logfreq_z + len_z + nsyl_z + old_z + conc_z + aoa_z + wp_z + (1 | partici
## mod.base.full: logRT ~ logfreq_z + len_z + nsyl_z + old_z + conc_z + aoa_z + wp_z + (1 | participant)
##                   npar     AIC     BIC logLik deviance  Chisq Df
## mod.base.nearfull   10 -3469.3 -3396.0 1744.7  -3489.3
## mod.base.full       11 -3645.6 -3564.9 1833.8  -3667.6 178.22  1
##                            Pr(>Chisq)
## mod.base.nearfull
## mod.base.full     < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

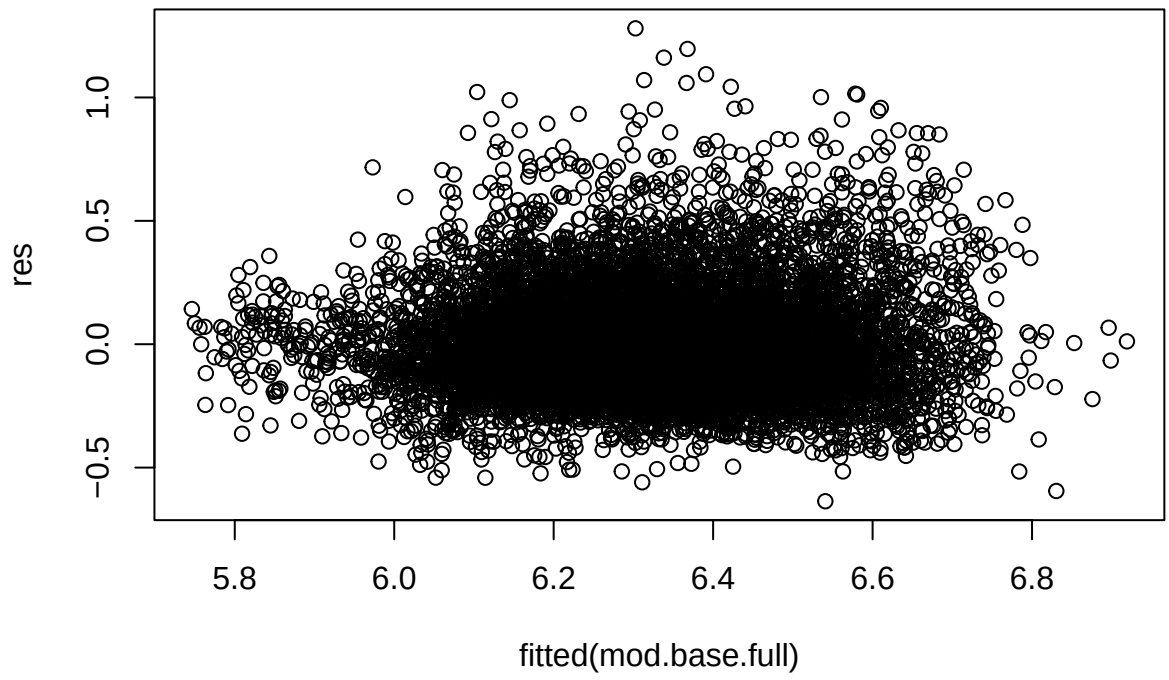
```
res <- residuals(mod.base.full)
hist(res)
```




```
qqnorm(res)
qqline(res)
```




```
plot(fitted(mod.base.full), res)
```




```
mod.full <- lmer(logRT ~
                   dim_type +
                   logfreq_z +
                   len_z +
                   nsyl_z +
                   old_z +
                   conc_z +
                   aoa_z +
                   wp_z +
                   (1|participant) +
                   (1|spelling), data = trialdata)
summary(mod.full)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: logRT ~ dim_type + logfreq_z + len_z + nsyl_z + old_z + conc_z +
##     aoa_z + wp_z + (1 | participant) + (1 | spelling)
##    Data: trialdata
##
## REML criterion at convergence: -3577.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1702 -0.6444 -0.1357  0.4713  6.3886
##
## Random effects:
```

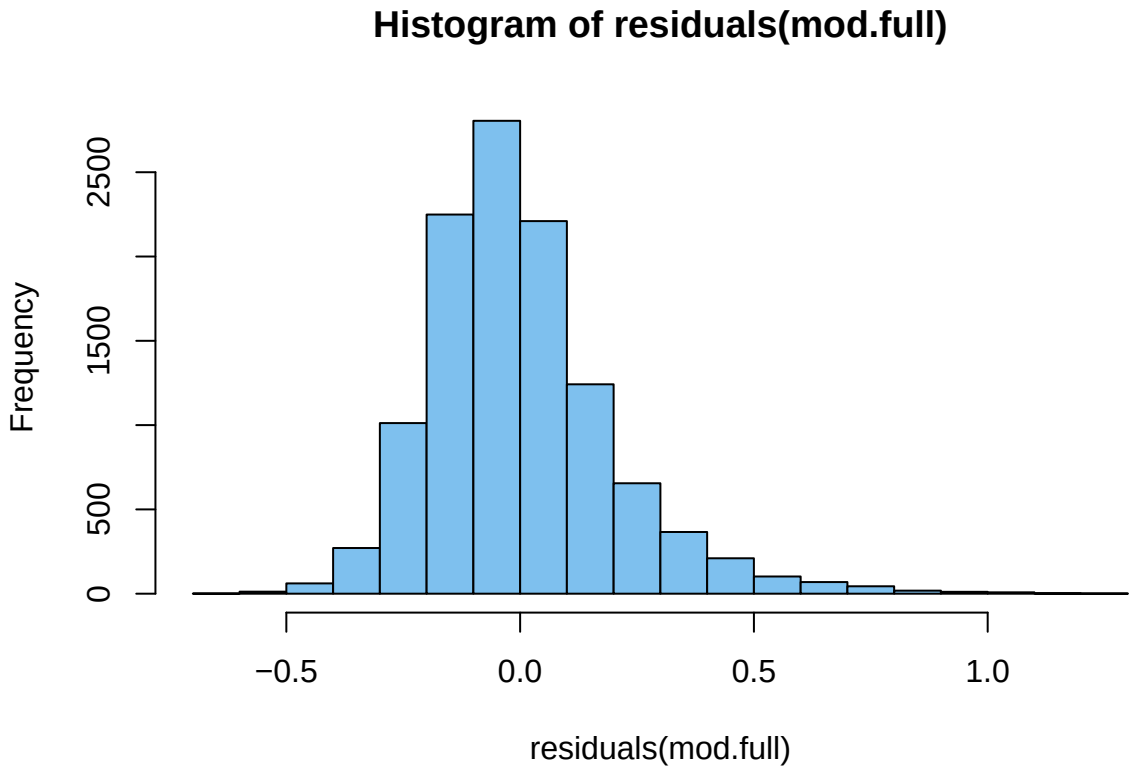
```
##  Groups      Name        Variance Std.Dev.
##  spelling    (Intercept) 0.001812 0.04256
##  participant (Intercept) 0.022780 0.15093
##  Residual                0.040031 0.20008
## Number of obs: 11349, groups:  spelling, 310; participant, 81
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)        6.3195596  0.0210781 299.816
## dim_typederiv     -0.0078104  0.0140926  -0.554
## dim_typeinfl       0.0053562  0.0138358   0.387
## dim_typeundecided -0.0051837  0.0140620  -0.369
## logfreq_z         -0.0271313  0.0042666  -6.359
## len_z              0.0006654  0.0085475   0.078
## nsyl_z             0.0008681  0.0063859   0.136
## old_z              0.0204088  0.0069715   2.927
## conc_z             0.0003027  0.0036734   0.082
## aoa_z              0.0166795  0.0041315   4.037
## wp_z              -0.0246800  0.0038643  -6.387
##
## Correlation of Fixed Effects:
##             (Intr) dm_typd dm_typnf dm_typnd lgfrq_ len_z  nsyl_z old_z  conc_z
## dim_typedrv -0.556
## dim_typenfl -0.552  0.828
## dm_typndcdd -0.555  0.840   0.829
## logfreq_z    0.001 -0.021   0.009    0.013
## len_z        0.072 -0.095  -0.062   -0.184    0.053
## nsyl_z      -0.039  0.047   0.019    0.120    0.065 -0.580
## old_z       -0.010 -0.006  -0.010    0.063    0.117 -0.632 -0.092
## conc_z      -0.026  0.090  -0.014    0.044    0.339  0.054  0.033 -0.056
## aoa_z        0.020 -0.063   0.004   -0.039    0.253 -0.063  0.032 -0.078  0.349
## wp_z        -0.017  0.050   0.027    0.020   -0.458 -0.077  0.002 -0.082 -0.211
##             aoa_z
## dim_typedrv
## dim_typenfl
## dm_typndcdd
## logfreq_z
## len_z
## nsyl_z
## old_z
## conc_z
## aoa_z
## wp_z         0.248
```

```
anova(mod.base.full, mod.full)
```

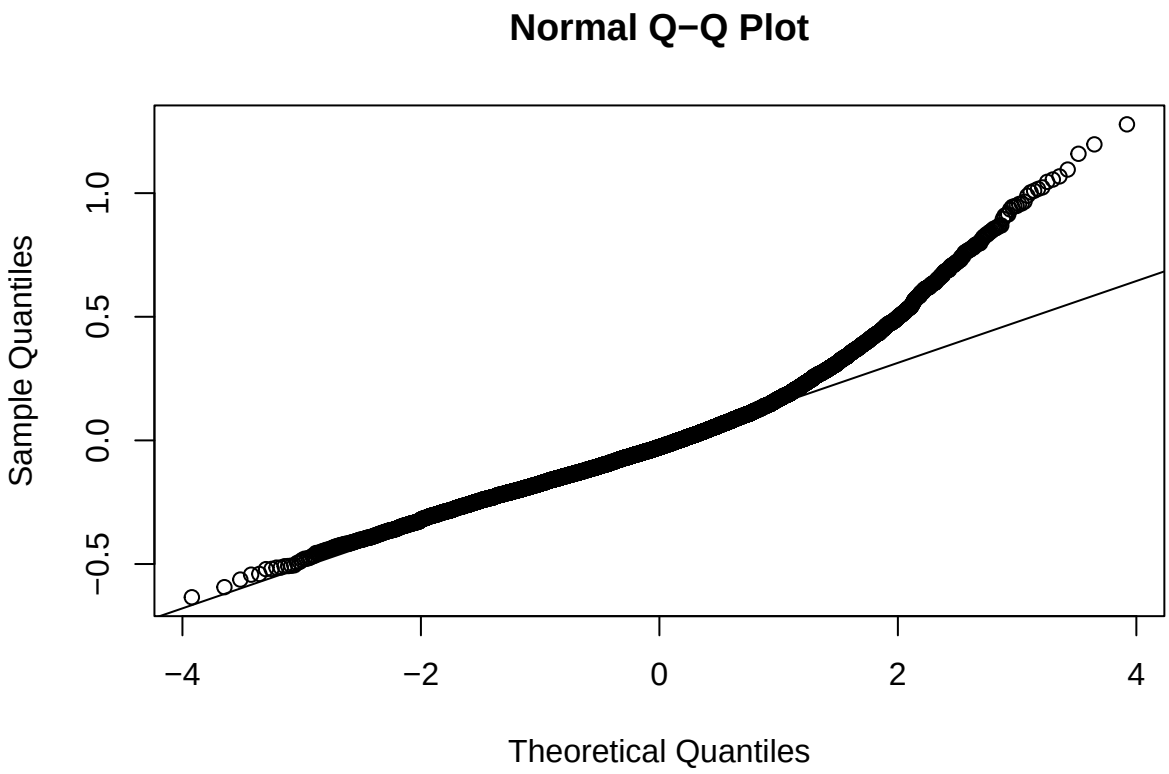
```
## refitting model(s) with ML (instead of REML)

## Data: trialdata
## Models:
## mod.base.full: logRT ~ logfreq_z + len_z + nsyl_z + old_z + conc_z + aoa_z + wp_z + (1 | participant)
## mod.full: logRT ~ dim_type + logfreq_z + len_z + nsyl_z + old_z + conc_z + aoa_z + wp_z + (1 | partic
##               npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
```

```
## mod.base.full   11 -3645.6 -3564.9 1833.8  -3667.6
## mod.full        14 -3642.5 -3539.8 1835.3  -3670.5 2.9547  3     0.3987
```

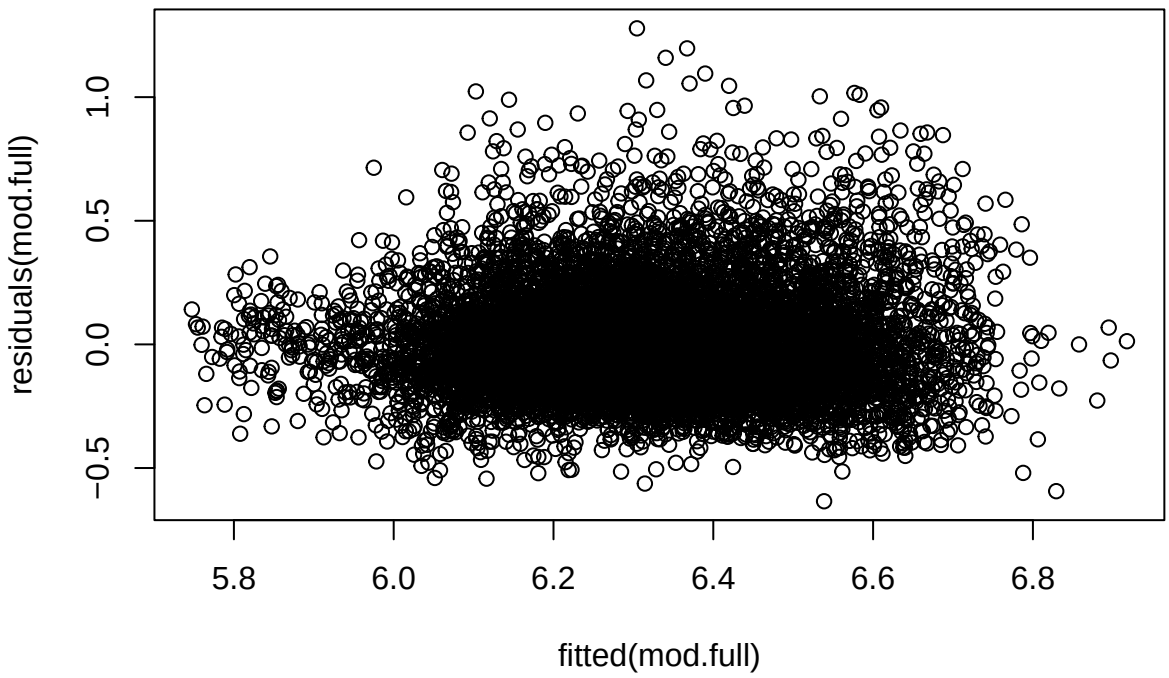
```
# Plot 1, histogram:
hist(residuals(mod.full), col = 'skyblue2')
```




```
# Plot 2, Q-Q plot:
qqnorm(residuals(mod.full))
qqline(residuals(mod.full))
```




```
# Plot 3, residual plot:
plot(fitted(mod.full), residuals(mod.full))
```


residuals(mod.full)
−0.5
0.0
0.5
1.0
5.8
6.0
6.2
6.4
6.6
6.8
fitted(mod.full)