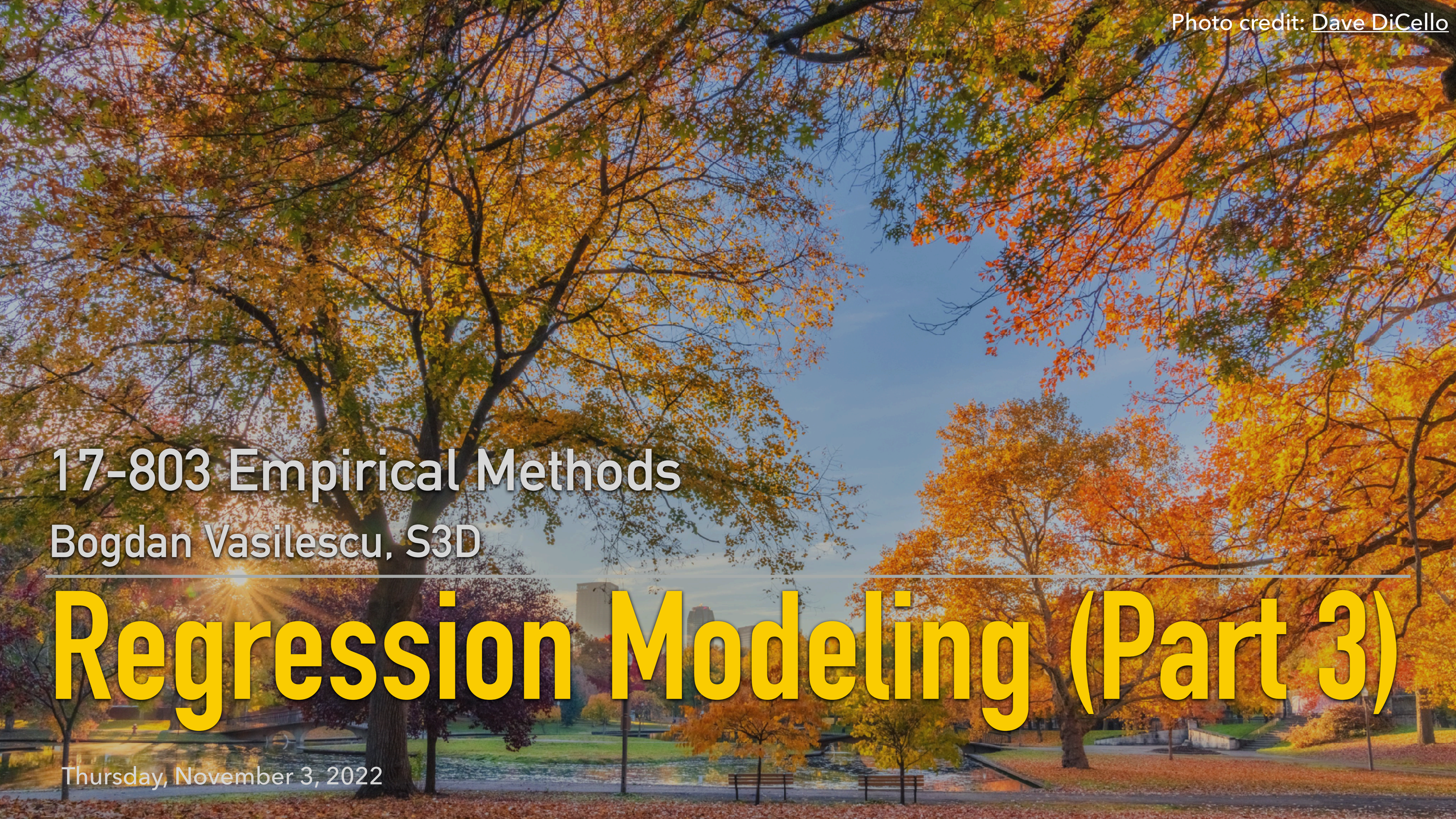


17-803 Empirical Methods

Bogdan Vasilescu, S3D

Regression Modeling (Part 3)

Thursday, November 3, 2022

Plan for Today

- Exemplar papers
- Simpson's paradox
- Hands-on activity

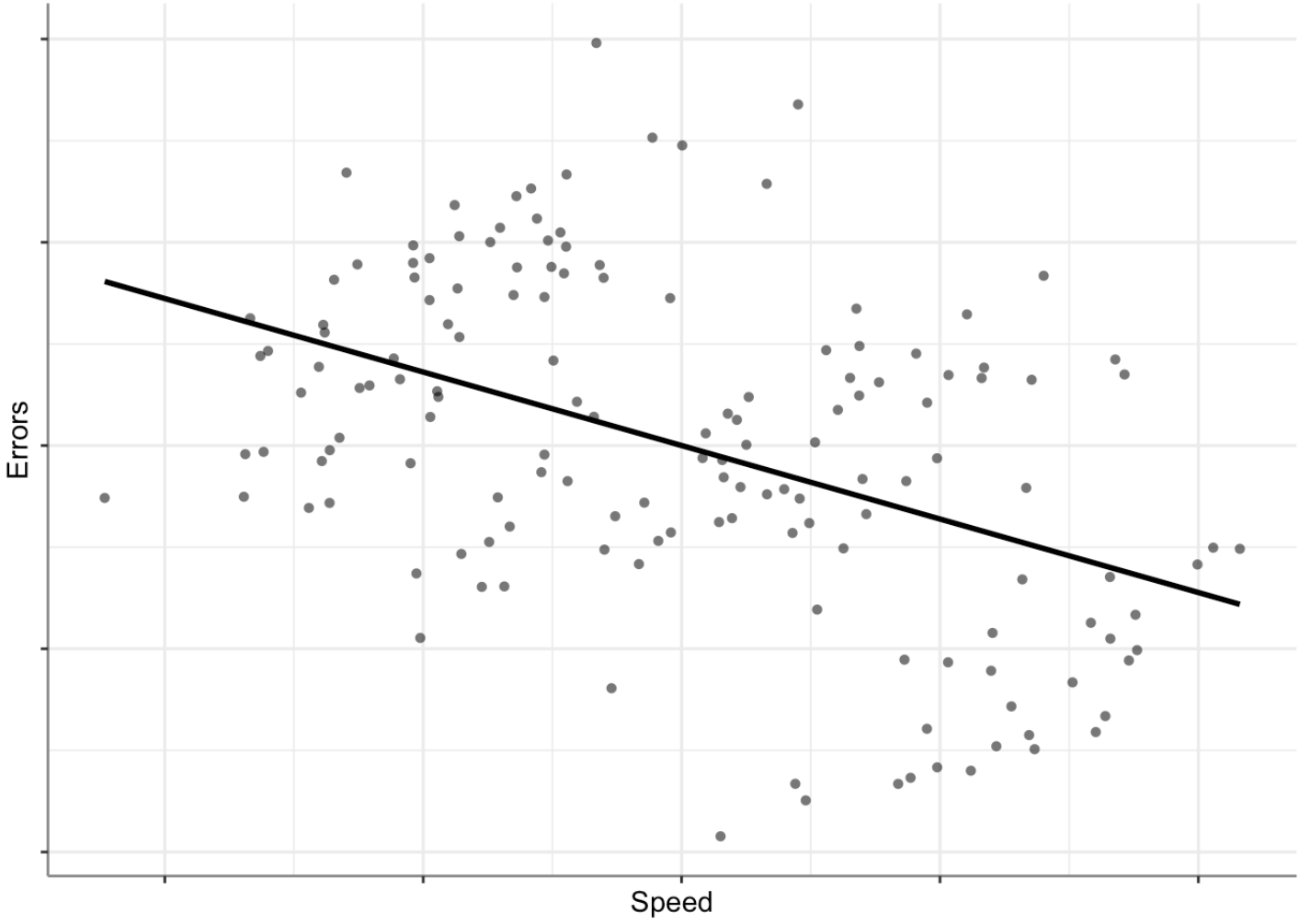
Exemplar Papers (Students Presented)

- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. Science, 330(6004), 686-688.
- Peoples, B. K., Midway, S. R., Sackett, D., Lynch, A., & Cooney, P. B. (2016). Twitter predicts citation rates of ecological research. PloS One, 11(11), e0166570.
- Lim, S. (2009). How and why do college students use Wikipedia?. Journal of the American Society for Information science and Technology, 60(11), 2189-2202 –> we discussed this on Tuesday, November 8 instead, see that video.

Accounting for Within- AND Between-Subject Effects

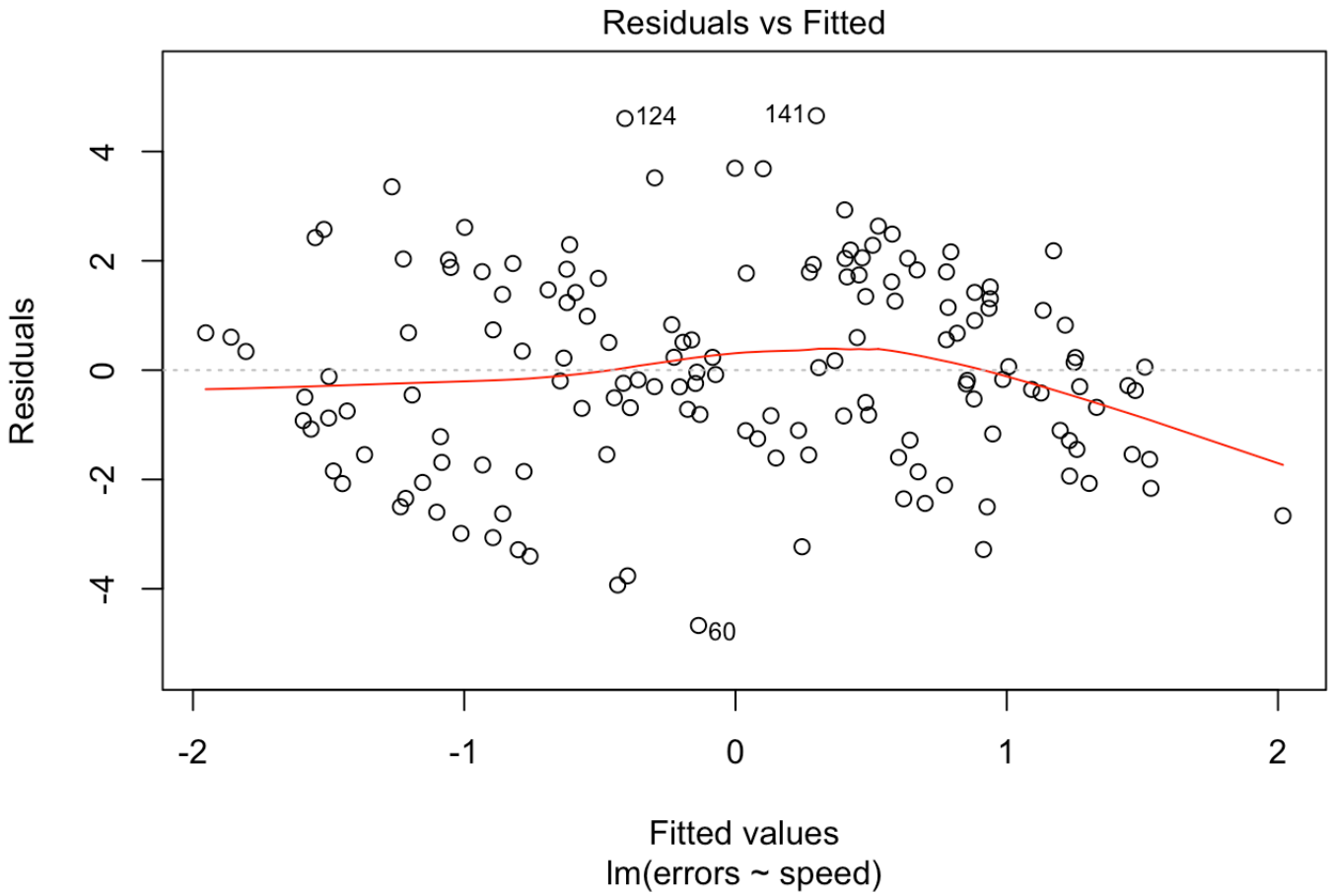
This example follows a blog post by Mattan S. Ben-Shachar: https://shouldbewriting.netlify.app/posts/2019-10-21-accounting-for-within-and-between-subject-effect (https://shouldbewriting.netlify.app/posts/2019-10-21-accounting-for-within-and-between-subject-effect) The key idea is that group-level data does not always reflect individual-level processes.

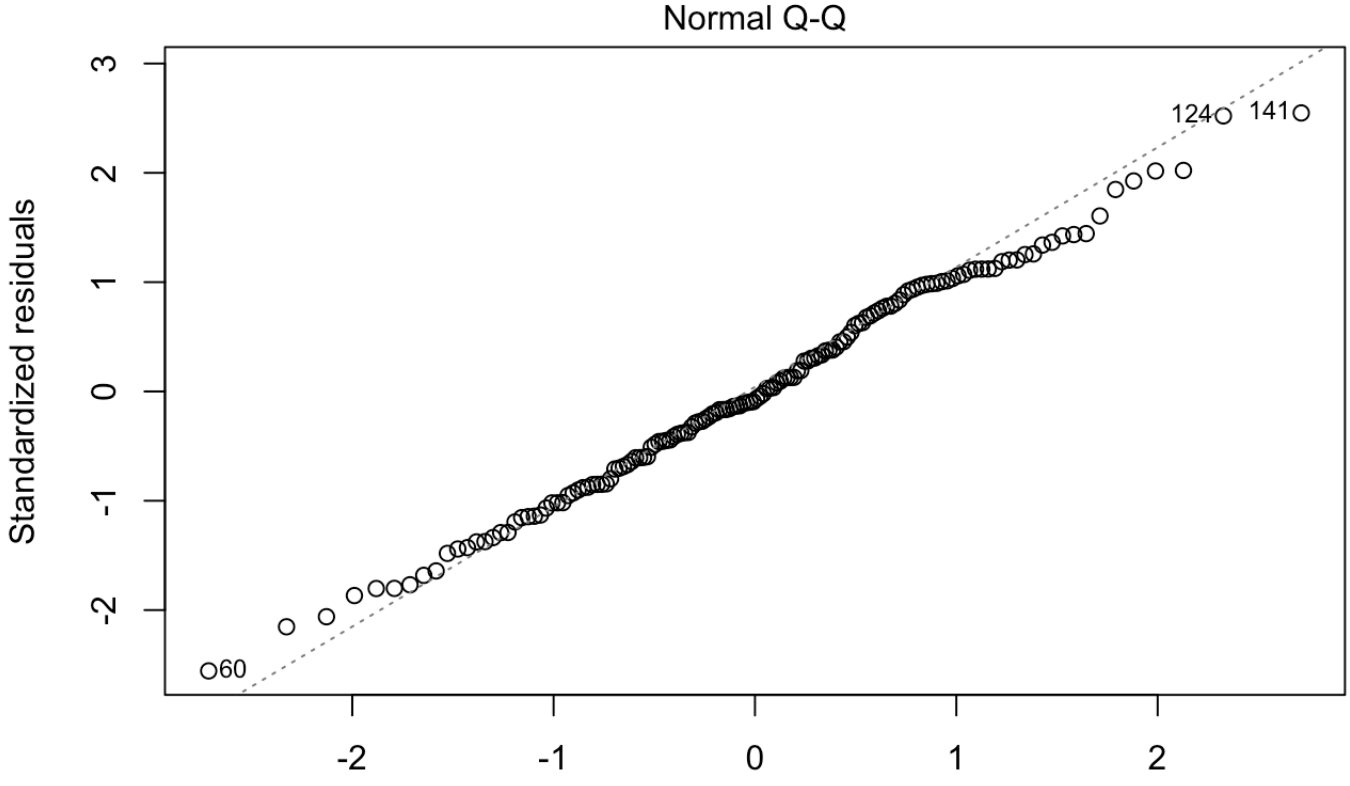
Let's work with the now-classic typing speed example. We take a group of 5 typists, and measure the speed of their typing (words per minute), and the rate of typing errors (errors per 100-words). Looking at the data we might get something like this:



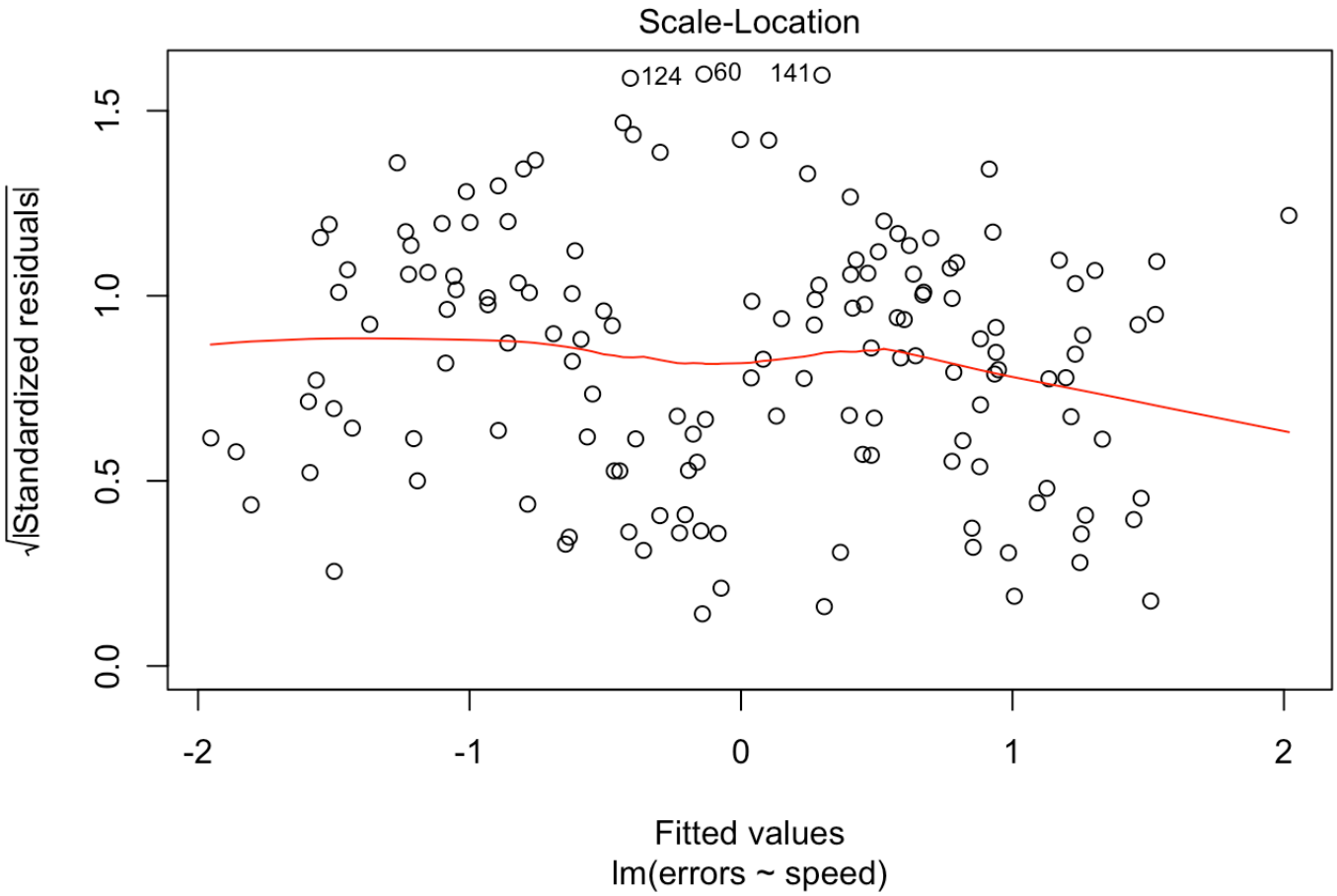
Let's estimate a simple linear model:

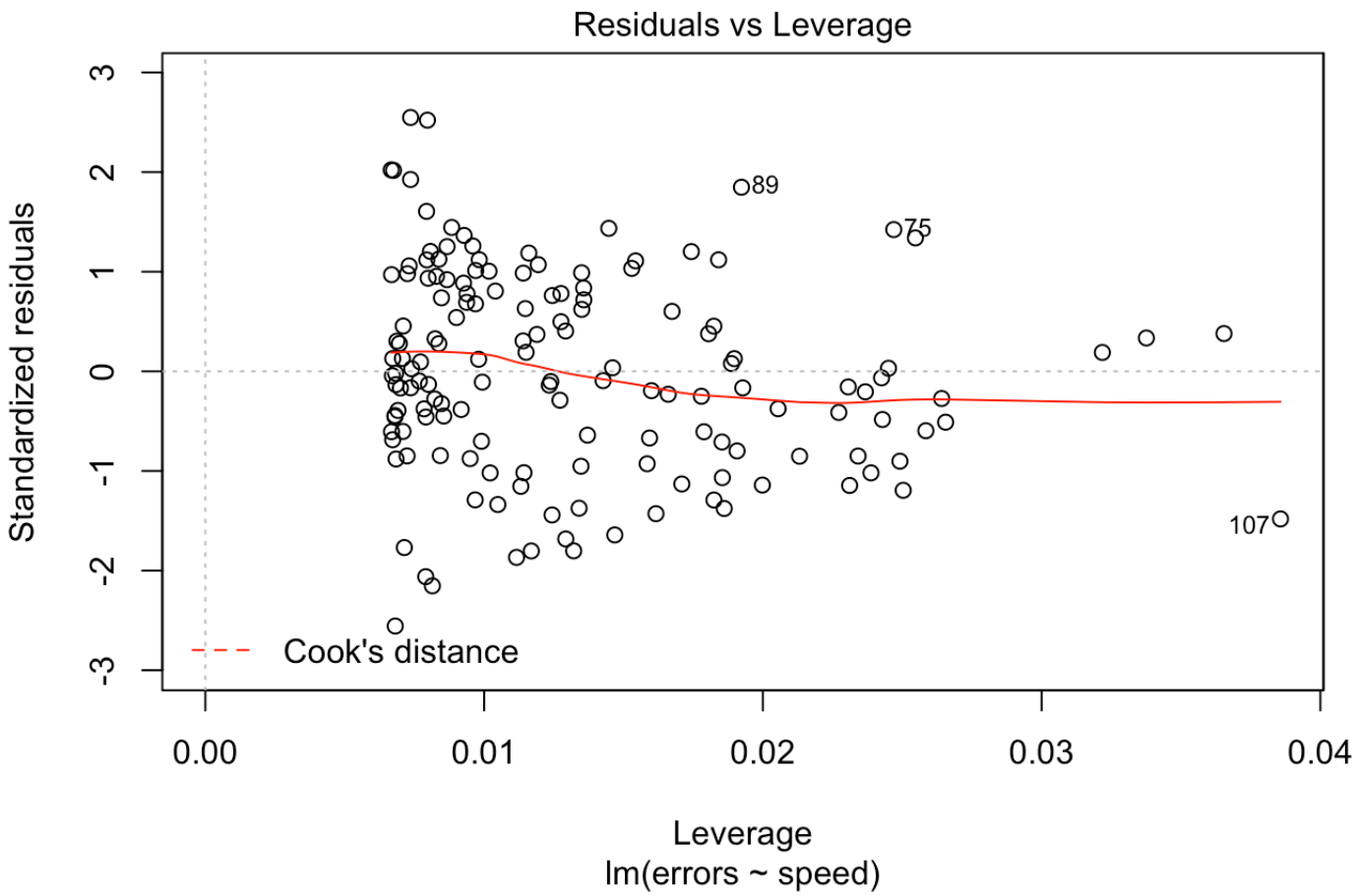
```
##
## Call:
## lm(formula = errors ~ speed, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6704 -1.2742 -0.1443  1.4222  4.6534
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.490e-16  1.497e-01   0.000        1
## speed       -9.040e-01  1.466e-01  -6.164 6.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.833 on 148 degrees of freedom
## Multiple R-squared:  0.2043, Adjusted R-squared:  0.1989
## F-statistic:    38 on 1 and 148 DF,  p-value: 6.391e-09
```


Residuals vs Fitted
Residuals
Fitted values
lm(errors ~ speed)
124
141
60
-2
-1
0
1
2
-4
-2
0
2
4


Normal Q-Q
Standardized residuals
124
141
60
-2
-1
0
1
2
-2
-1
0
1
2
3

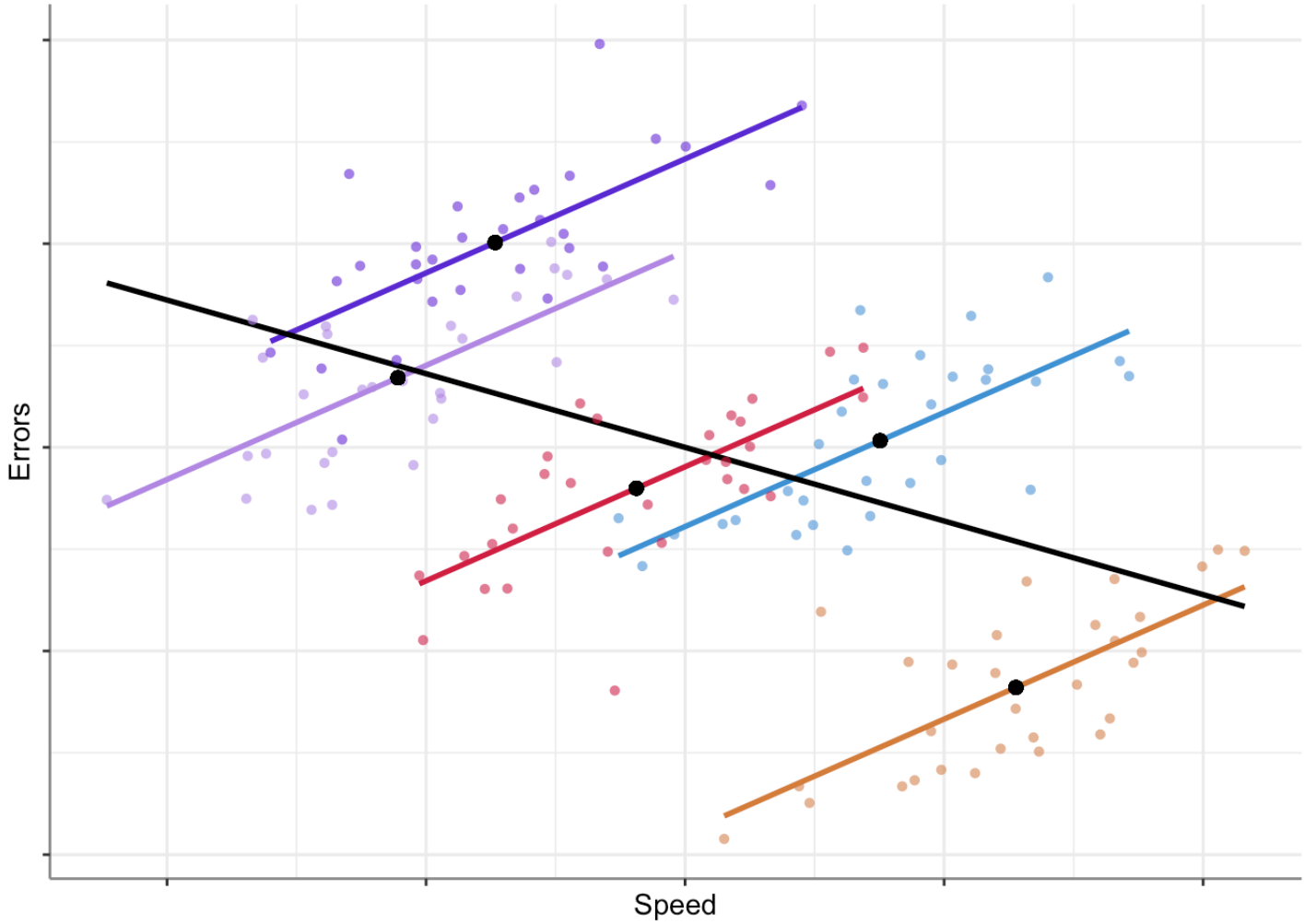
Theoretical Quantiles
lm(errors ~ speed)





Note how the model suggests a negative relationship between speed and errors: the faster people type, the fewer errors they make.

Now let's break down the data by typist:



As we can see, we have two sources of variation that can be used to explain or predict the rate of errors:

1. Overall, faster typists make less mistakes (group-level pattern).
2. When typing faster, typists make more mistakes (individual-level pattern).

Let's model the typist as a fixed effect.

```
##
## Call:
## lm(formula = errors ~ speed + as.factor(ID), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36716 -0.53360 -0.00212  0.42014  1.98377
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.2384     0.1327  -1.796 0.074583 .
## speed            1.4000     0.1190  11.762  < 2e-16 ***
## as.factor(ID)2  -4.4978     0.2543 -17.688  < 2e-16 ***
## as.factor(ID)3  -0.7338     0.2163  -3.393 0.000892 ***
## as.factor(ID)4   2.6441     0.2150  12.296  < 2e-16 ***
## as.factor(ID)5   3.7795     0.1961  19.276  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7166 on 144 degrees of freedom
## Multiple R-squared:  0.8817, Adjusted R-squared:  0.8776
## F-statistic: 214.6 on 5 and 144 DF,  p-value: < 2.2e-16
```

Now we see that typing speed is positively correlated with errors, controlling for typist.

Aside: We can also model these using liner mixed-effects models.

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: errors ~ speed + (1 | ID)
##    Data: data
##
## REML criterion at convergence: 355.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3055 -0.7421  0.0012  0.5969  2.7401
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 10.3393  3.2155
##  Residual              0.5136  0.7167
## Number of obs: 150, groups:  ID, 5
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept) -1.948e-13  1.439e+00  3.963e+00    0.00        1
## speed        1.384e+00  1.187e-01  1.454e+02   11.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr)
## speed 0.000
```

What if we want to capture both the within- and between-group patterns?

We can model these using liner mixed models, but first we need to *split* our predictor (*speed*) into two variables, each representing a different source of variance - each typist's average typing speed, and the deviation of each measurement from the typist's overall mean:[1]

```
library(dplyr)
data <- data %>%
  group_by(ID) %>%
  mutate(speed_M = mean(speed),
         speed_E = speed - speed_M) %>%
  ungroup()

head(data)
```

```
## # A tibble: 6 x 5
##      ID  speed errors speed_M speed_E
##   <int>  <dbl>  <dbl>   <dbl>   <dbl>
## 1     1 -0.773 -1.74   -0.188 -0.585
## 2     1 -0.144 -0.703  -0.188  0.0438
## 3     1 -0.686 -1.73   -0.188 -0.498
## 4     1  0.560  1.17   -0.188  0.748
## 5     1  0.214  0.316  -0.188  0.402
## 6     1  0.179  0.392  -0.188  0.367
```

Let's fit a liner mixed model and see how we can detect both patterns correctly.

```
fit <- lmer(errors ~ speed_M + speed_E + (1 | ID), data = data)
summary(fit)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: errors ~ speed_M + speed_E + (1 | ID)
##    Data: data
##
## REML criterion at convergence: 346.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3132 -0.7550  0.0006  0.6020  2.7570
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 1.9029   1.3794
##  Residual             0.5135   0.7166
## Number of obs: 150, groups:  ID, 5
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  3.515e-15  6.197e-01  3.000e+00   0.000    1.000
## speed_M     -1.600e+00  6.928e-01  3.000e+00  -2.309    0.104
## speed_E      1.400e+00  1.190e-01  1.440e+02  11.762   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) sped_M
## speed_M 0.000
## speed_E 0.000  0.000
```

As we can see, the slope for `speed_M` is negative (-1.6), reflecting the group-level pattern where typists who are overall faster have fewer errors; whereas the slope for `speed_E` is positive (1.4), reflecting the individual-level pattern where faster typing leads to more errors.

We can access the estimated deviation between each subject average typing speed and the overall average:

```
## $ID
##   (Intercept)
## 1  -0.7955464
## 2  -0.8960398
## 3   1.2740378
## 4  -0.9118556
## 5   1.3294040
##
## with conditional variances for "ID"
```

Question: Do we need a random slope?

```
## boundary (singular) fit: see ?isSingular
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: errors ~ speed + (1 + speed | ID)
##    Data: data
##
## REML criterion at convergence: 355.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3056 -0.7419  0.0023  0.5954  2.7289
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  ID       (Intercept) 1.032e+01 3.213124
##           speed       5.225e-05 0.007229 1.00
##  Residual             5.136e-01 0.716659
## Number of obs: 150, groups:  ID, 5
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 5.357e-03  1.438e+00 3.957e+00   0.004    0.997
## speed       1.384e+00  1.187e-01 1.378e+02  11.652   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr)
## speed 0.027
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

Answer: probably not.

Question: Could we capture the between-effect with a fixed-effects model?

```
##
## Call:
## lm(formula = errors ~ speed_M, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2864 -1.0748  0.0028  0.9154  3.7784
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.702e-16  1.196e-01    0.00        1
## speed_M     -1.600e+00  1.338e-01  -11.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.465 on 148 degrees of freedom
## Multiple R-squared:  0.4915, Adjusted R-squared:  0.4881
## F-statistic: 143.1 on 1 and 148 DF,  p-value: < 2.2e-16
```

Answer: It is different from the -0.9 estimated in the model m1. The above model 1 (m1) returns a biased estimate, which is a “weighted average” of the within- and between-effects.

Note that standard errors differ, because the variance in the grouping structure is more accurately taken into account by the mixed-effects model.

When are individual-level the same as group-level patterns?

Experiments!
Or to be more precise, when *we* control the values of the independent variable. Why is this so? Because we control the values of the independent variable, the independent variable cannot be split into different sources of variance: there is either variance between subjects (the variable is manipulated in a between-subjects design) or there is variance within subjects (the variable is manipulated in a within-subjects design), but never both. Thus, although there can be huge heterogeneity in the way subjects present an effect, the average individual-level effect will be the same as the group-level effect (depending on the design).[2]

---

1. Read more in: Hoffman, L. (2015). Time-varying predictors in models of within-person fluctuation. In *Longitudinal analysis: Modeling within-person fluctuation and change* (pp. 327-392). Routledge.↩
2. Ignoring any differences or artifacts that may arise from the differences in the design itself, such as order effects, etc.↩

In-Class Activity

- Galton families data

In-class R activity

```
# Load the Galton height data from the `HistData` library.
install.packages("HistData")
library("HistData")
data(GaltonFamilies)
str(GaltonFamilies)
```

Questions:

- How many children are there in the data?
- How many unique families are there?
- What are the overall mean & median height of the children?
- What are the overall mean & median height of the male children? Idem female children?
- Can you visualize the distributions of children's height by gender?
- Use a linear model to estimate the mean height of children.
- Use a linear model to estimate the mean height of each gender. What do the diagnostic plots tell you about this model?
- Fit a multiple regression model predicting children's height from father's height, mother's height, and gender.
- Change reference level for the gender variable to "female" and re-estimate the model. Inspect the residuals.
- Scale the parents' height, re-estimate the model, and compare interpretations.
- We would expect siblings to be somewhat similar in height as they share genetic factors through their parents and environmental factors through their shared upbringing. Fit a new model to estimate both the population means as well as how average family heights vary around these population means. Hint: mixed-effects models can be used for this.
- Fit a fixed-effects-only model to control for family and compare interpretations.

R hints:

- lm: built-in
- lmer: `library(lme4); library(lmerTest)`
    - Recall typing speed example: `lmer(errors ~ speed + (1 | ID), data = data)`
- vif: `library(car)`

Beyond Linear Regression

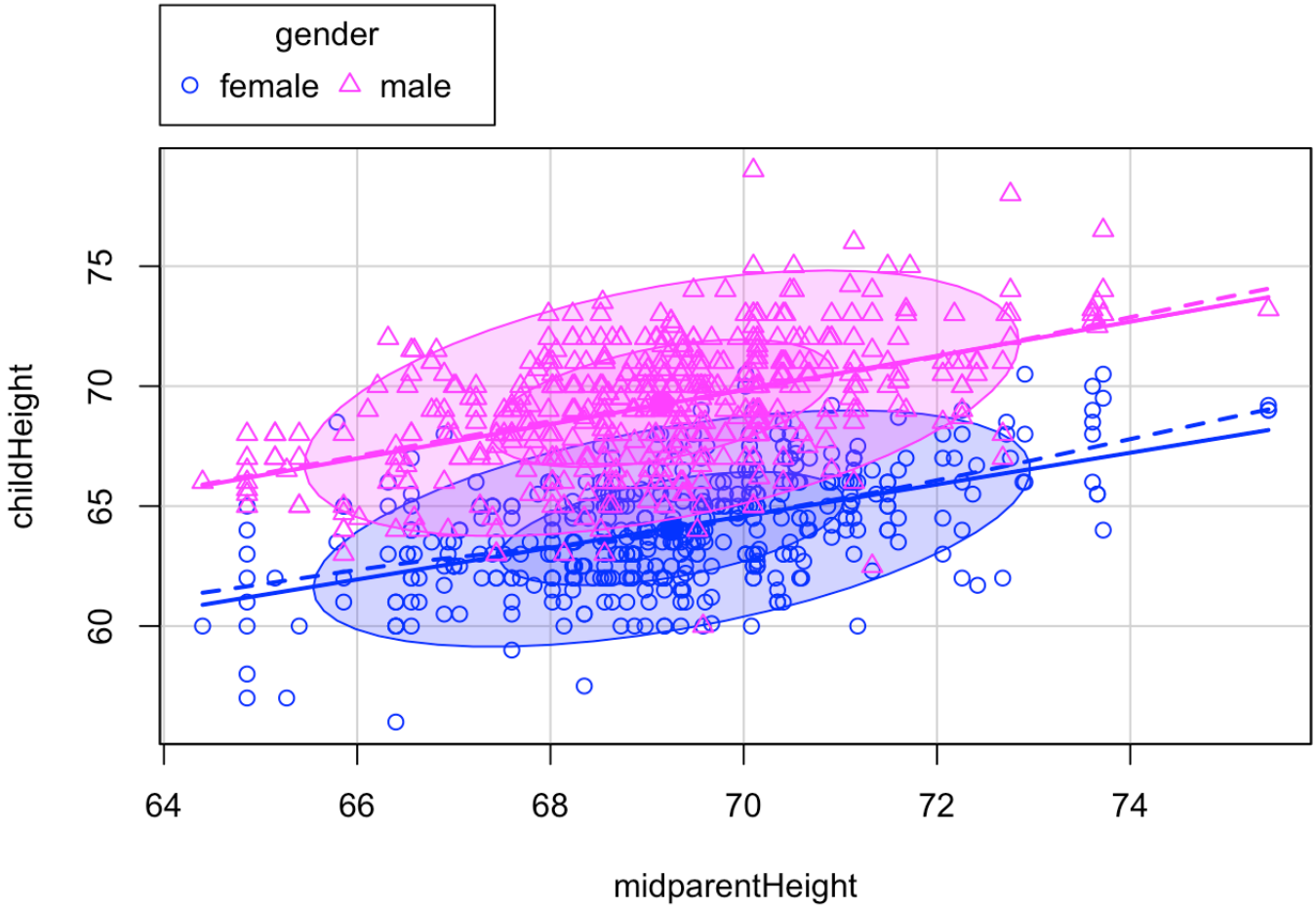
Load the Galton height data from the `HistData` library.

```
data(GaltonFamilies)
str(GaltonFamilies)
```

```
## 'data.frame':    934 obs. of  8 variables:
##  $ family         : Factor w/ 205 levels "001","002","003",..: 1 1 1 1 2 2 2 2 3 3
...
##  $ father         : num  78.5 78.5 78.5 78.5 75.5 75.5 75.5 75.5 75 75 ...
##  $ mother         : num  67 67 67 67 66.5 66.5 66.5 66.5 64 64 ...
##  $ midparentHeight: num  75.4 75.4 75.4 75.4 73.7 ...
##  $ children       : int  4 4 4 4 4 4 4 4 2 2 ...
##  $ childNum       : int  1 2 3 4 1 2 3 4 1 2 ...
##  $ gender         : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 1 1 2 1 ...
##  $ childHeight    : num  73.2 69.2 69 69 73.5 72.5 65.5 65.5 71 68 ...
```

Reproduce Fig 2 in Hanley (2004).

```
scatterplot(childHeight ~ midparentHeight | gender, data=GaltonFamilies,
    ellipse=TRUE, levels=0.68, legend.coords=list(x=64, y=78))
```



Data exploration

```
nrow(GaltonFamilies)
```

```
## [1] 934
```

```
length(unique(GaltonFamilies$family))
```

```
## [1] 205
```

```
summary(GaltonFamilies$children)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   6.000   6.171   8.000  15.000
```

```
table(GaltonFamilies$gender)
```

```
##
## female   male
##    453    481
```

```
summary(GaltonFfamilies$childHeight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.00   64.00   66.50   66.75   69.70   79.00
```

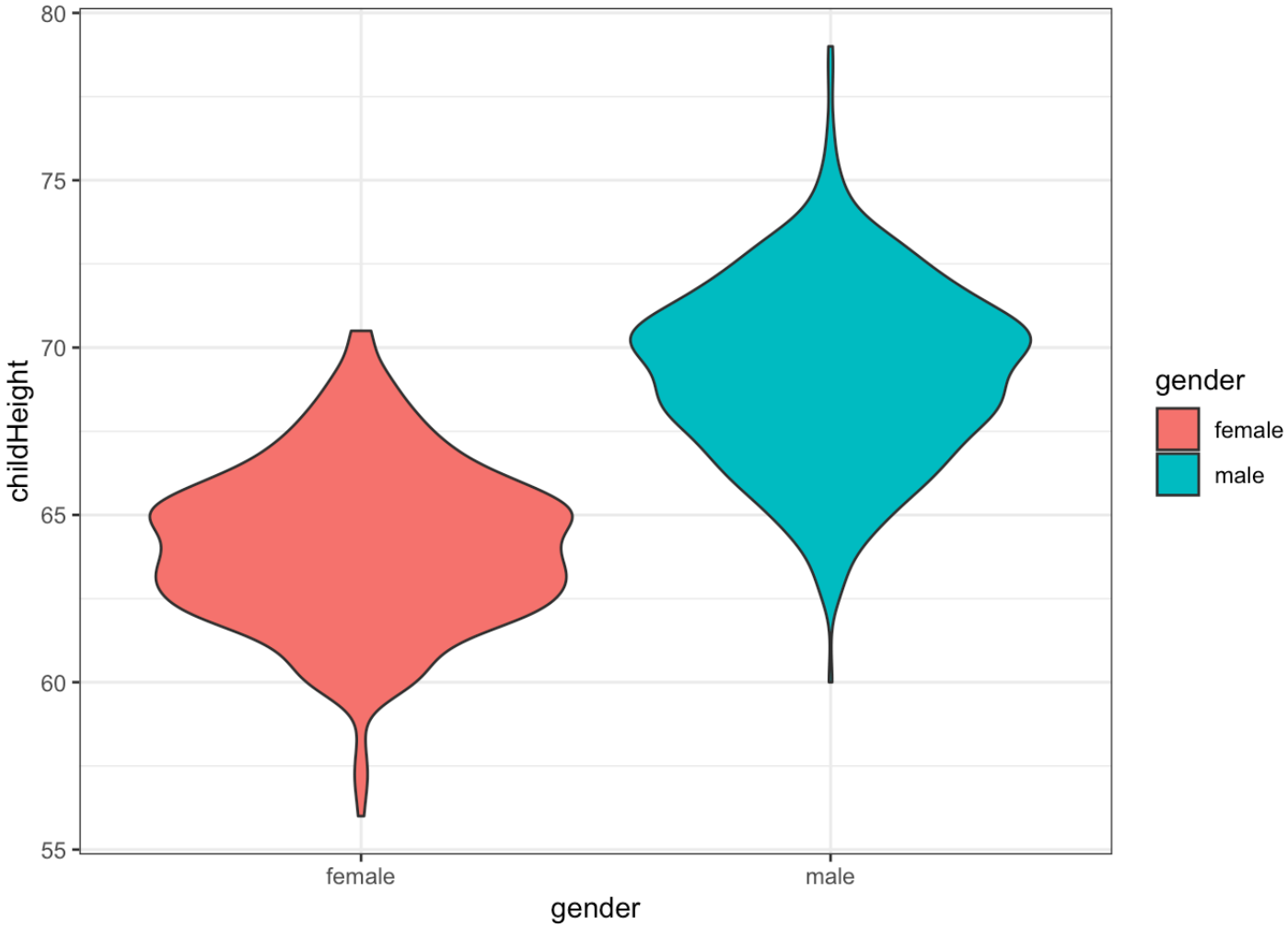
```
sd(GaltonFamilies$father)
```

```
## [1] 2.476479
```

```
sd(GaltonFamilies$mother)
```

```
## [1] 2.290886
```

```
ggplot(GaltonFamilies, aes(x=gender, y=childHeight, fill=gender)) + geom_violin() + t
heme_bw()
```



What's the simplest possible model of this data you could imagine? Probably one that just computes the mean height.

```
summary(lm(childHeight ~ 1, data=GaltonFamilies))
```

```
##
## Call:
## lm(formula = childHeight ~ 1, data = GaltonFamilies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7459  -2.7459  -0.2459   2.9541  12.2541
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.7459     0.1171   569.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.579 on 933 degrees of freedom
```

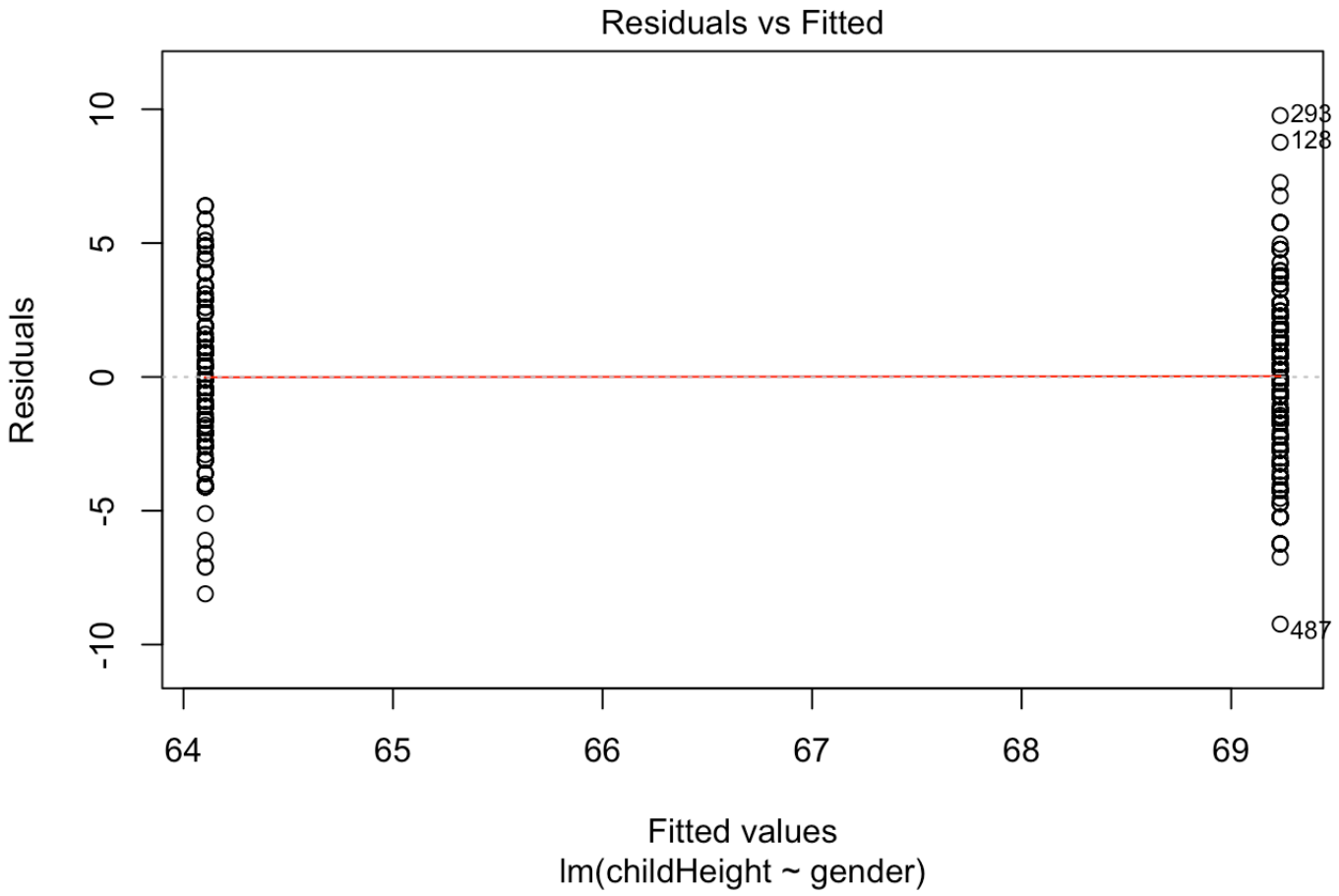
Let's make it more realistic. Fit a simple regression model predicting children's height from gender.

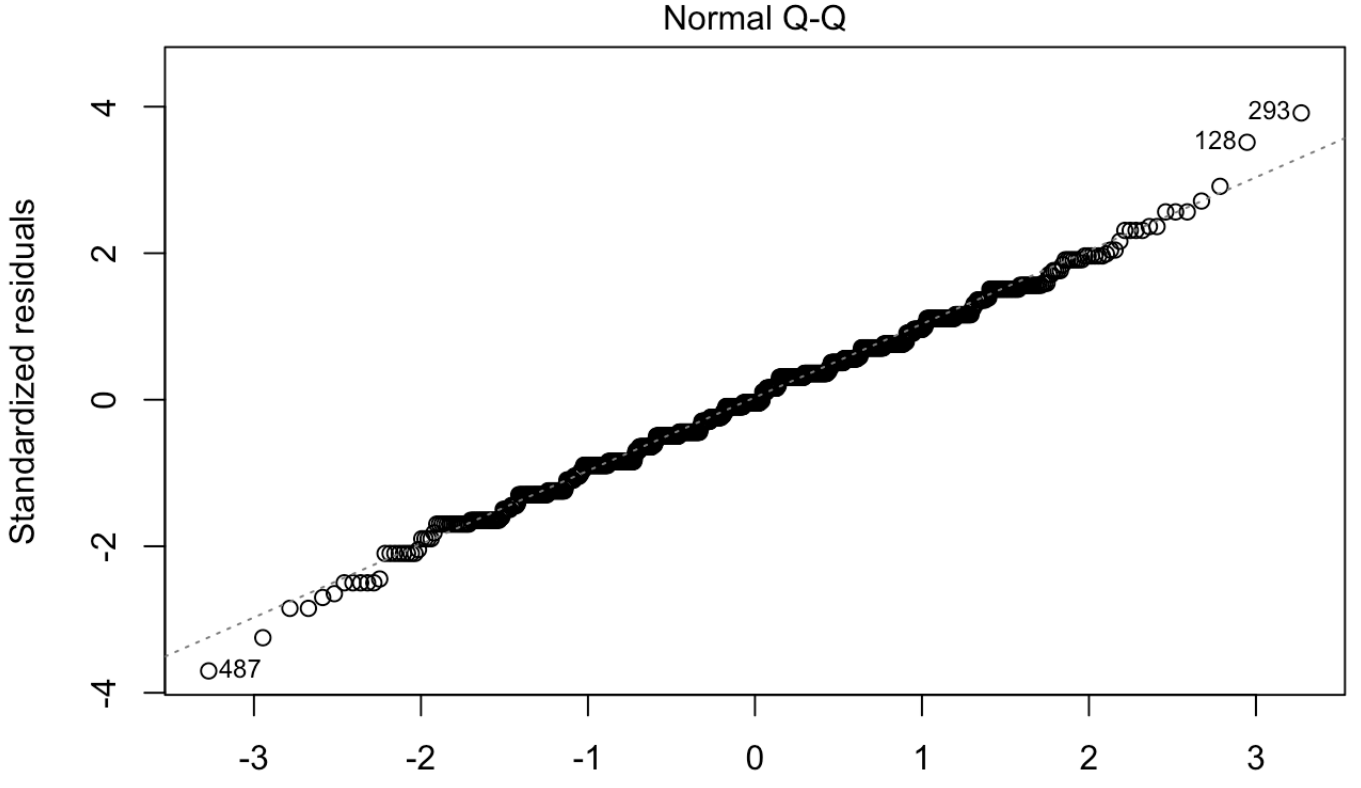
```
m0 = lm(childHeight ~ gender, data=GaltonFamilies)
summary(m0)
```

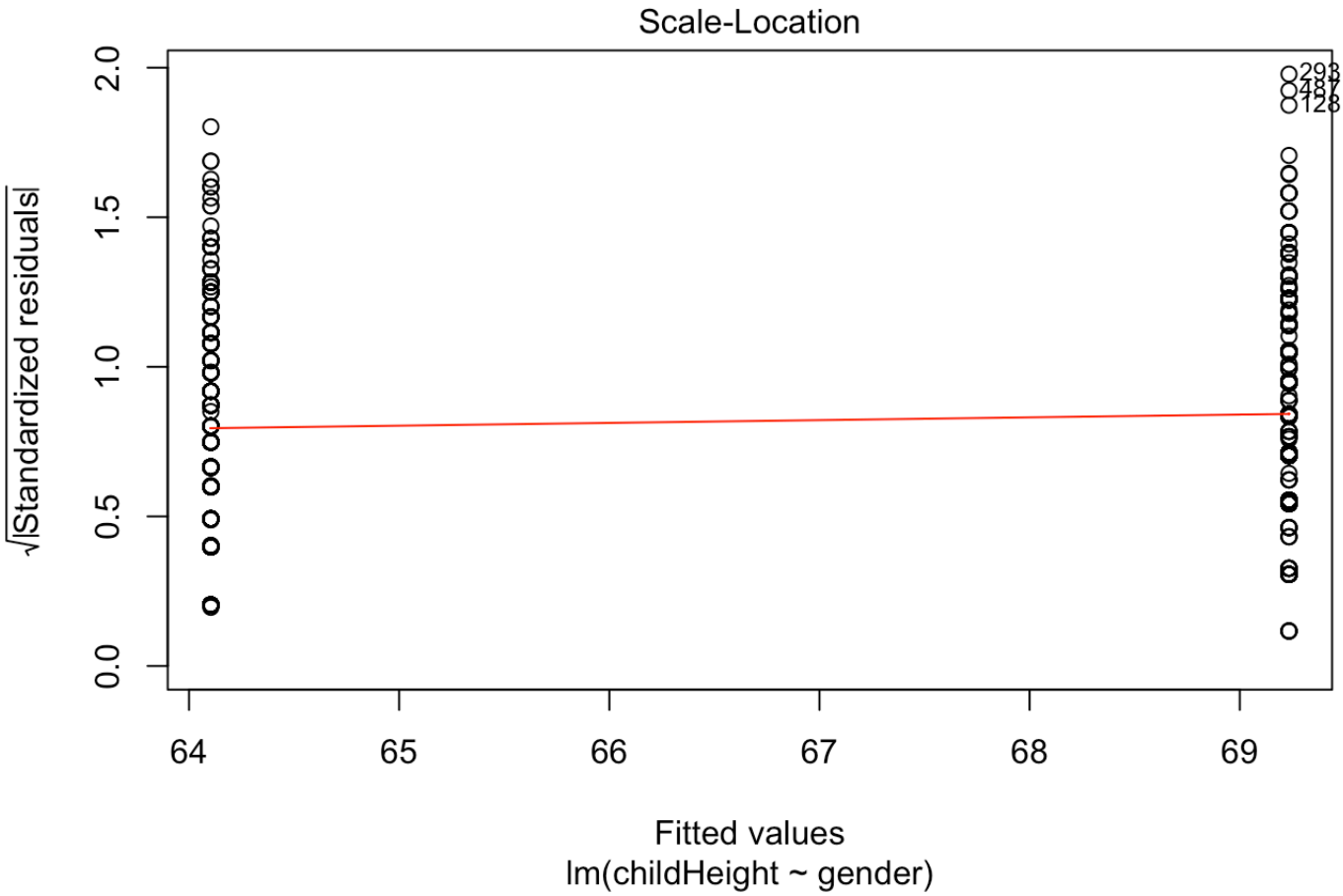
```
##
## Call:
## lm(formula = childHeight ~ gender, data = GaltonFamilies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.234 -1.604 -0.104  1.766  9.766
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.1040     0.1173  546.32   <2e-16 ***
## gendermale    5.1301     0.1635   31.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.497 on 932 degrees of freedom
## Multiple R-squared:  0.5137, Adjusted R-squared:  0.5132
## F-statistic: 984.4 on 1 and 932 DF,  p-value: < 2.2e-16
```

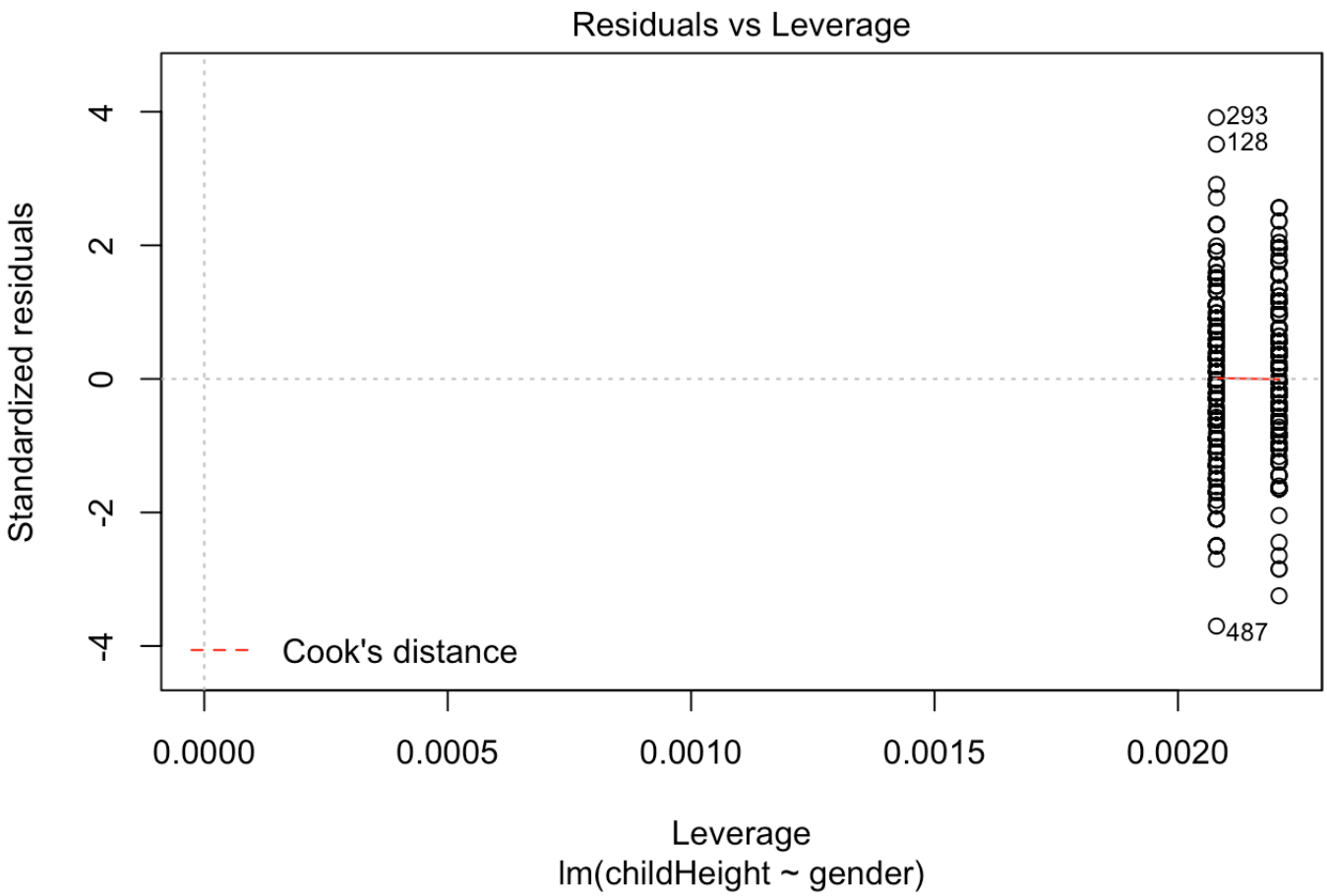
Look at some model diagnostics to confirm that this is an appropriate model. Since this model only produces two different predictions (one for males and one for females), that isn't very helpful here.

```
plot(m0)
```


Residuals vs Fitted
Residuals
10
5
0
-5
-10
293
128
487
64
65
66
67
68
69
Fitted values
lm(childHeight ~ gender)

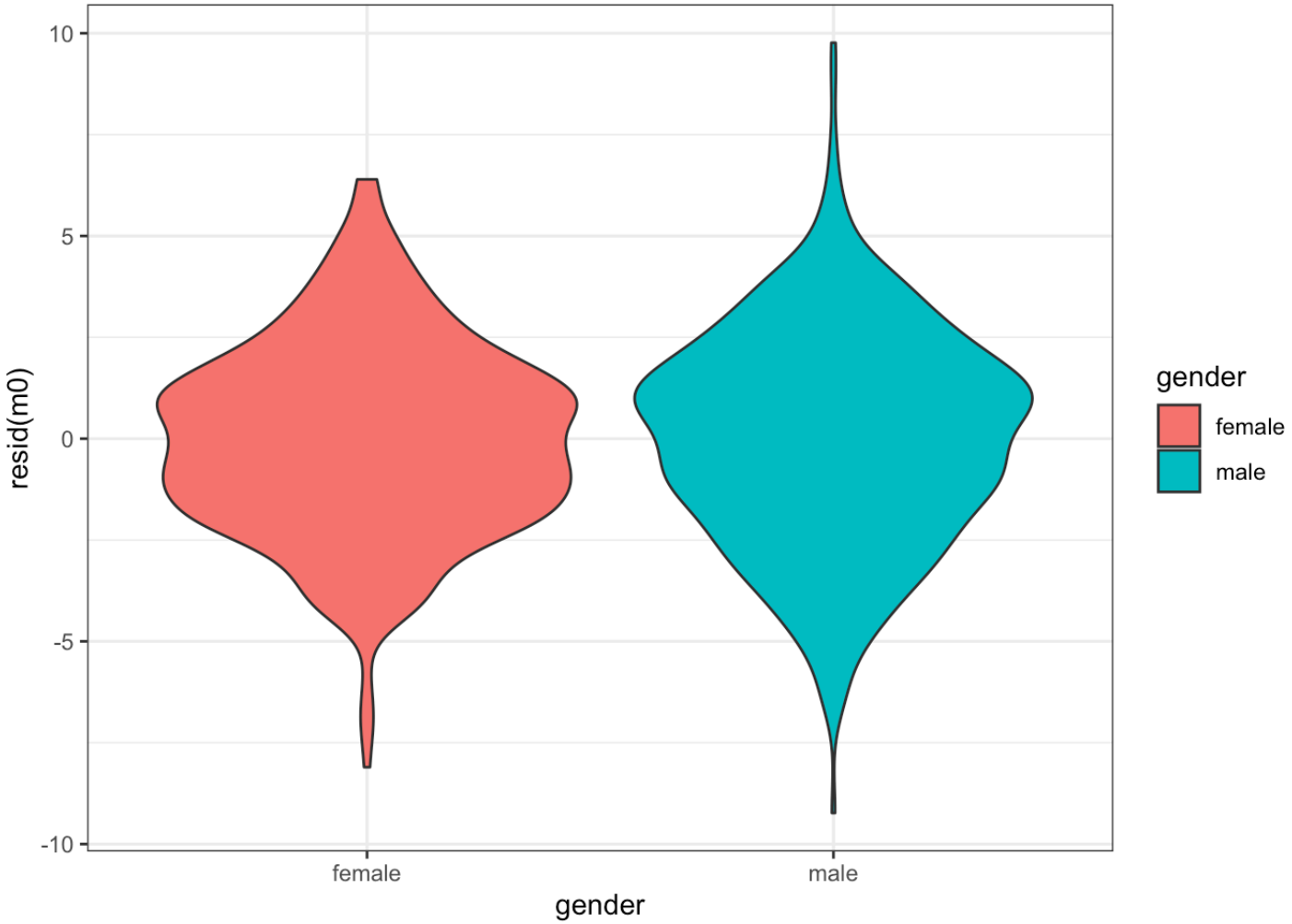

Normal Q-Q
Standardized residuals
4
2
0
-2
-4
293
128
487
-3
-2
-1
0
1
2
3


Theoretical Quantiles
lm(childHeight ~ gender)
Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
293
187
128
64
65
66
67
68
69
Fitted values
lm(childHeight ~ gender)



A boxplot or violin plot can help to summarise the distribution of residuals by group. Since the model simply estimates the mean heights of males and females a violin plot of the residuals should look very similar to the violin plot of heights above, but with the means of both groups aligned at 0.

```
ggplot(GaltonFamilies, aes(x=gender, y=resid(m0), fill=gender)) +
   geom_violin() +
   theme_bw()
```



Fit a multiple regression model predicting children's height from father's height, mother's height, and gender

```
m1 = lm(childHeight ~ gender + father + mother, data=GaltonFamilies)
summary(m1)
```

```
##
## Call:
## lm(formula = childHeight ~ gender + father + mother, data = GaltonFamilies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5247 -1.4653  0.0943  1.4860  9.1201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.52124    2.72720   6.058    2e-09 ***
## gendermale   5.21499    0.14181  36.775   <2e-16 ***
## father       0.39284    0.02868  13.699   <2e-16 ***
## mother       0.31761    0.03100  10.245   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.165 on 930 degrees of freedom
## Multiple R-squared:  0.6354, Adjusted R-squared:  0.6342
## F-statistic: 540.3 on 3 and 930 DF,  p-value: < 2.2e-16
```

Change reference level for gender variable.

```
levels(GaltonFamilies$gender)
```

```
## [1] "female" "male"
```

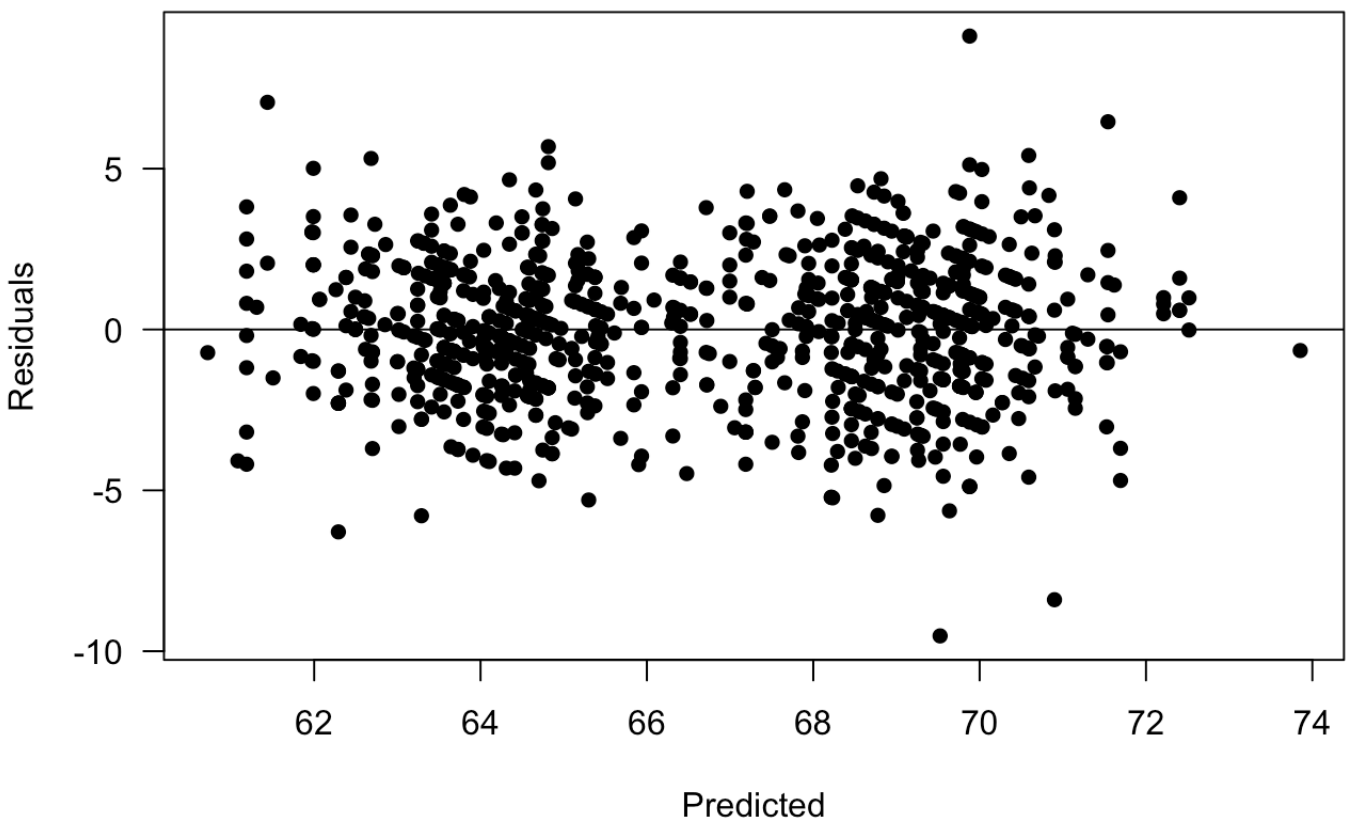
```
GaltonFamilies$gender = factor(GaltonFamilies$gender, levels = c("male", "female"))

# The regression
m2 = lm(childHeight ~ gender + father + mother, data=GaltonFamilies)
summary(m2)
```

```
##
## Call:
## lm(formula = childHeight ~ gender + father + mother, data = GaltonFamilies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5247 -1.4653  0.0943  1.4860  9.1201
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.73623    2.72223   7.985 4.14e-15 ***
## genderfemale -5.21499    0.14181 -36.775  < 2e-16 ***
## father        0.39284    0.02868  13.699  < 2e-16 ***
## mother        0.31761    0.03100  10.245  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.165 on 930 degrees of freedom
## Multiple R-squared:  0.6354, Adjusted R-squared:  0.6342
## F-statistic: 540.3 on 3 and 930 DF,  p-value: < 2.2e-16
```

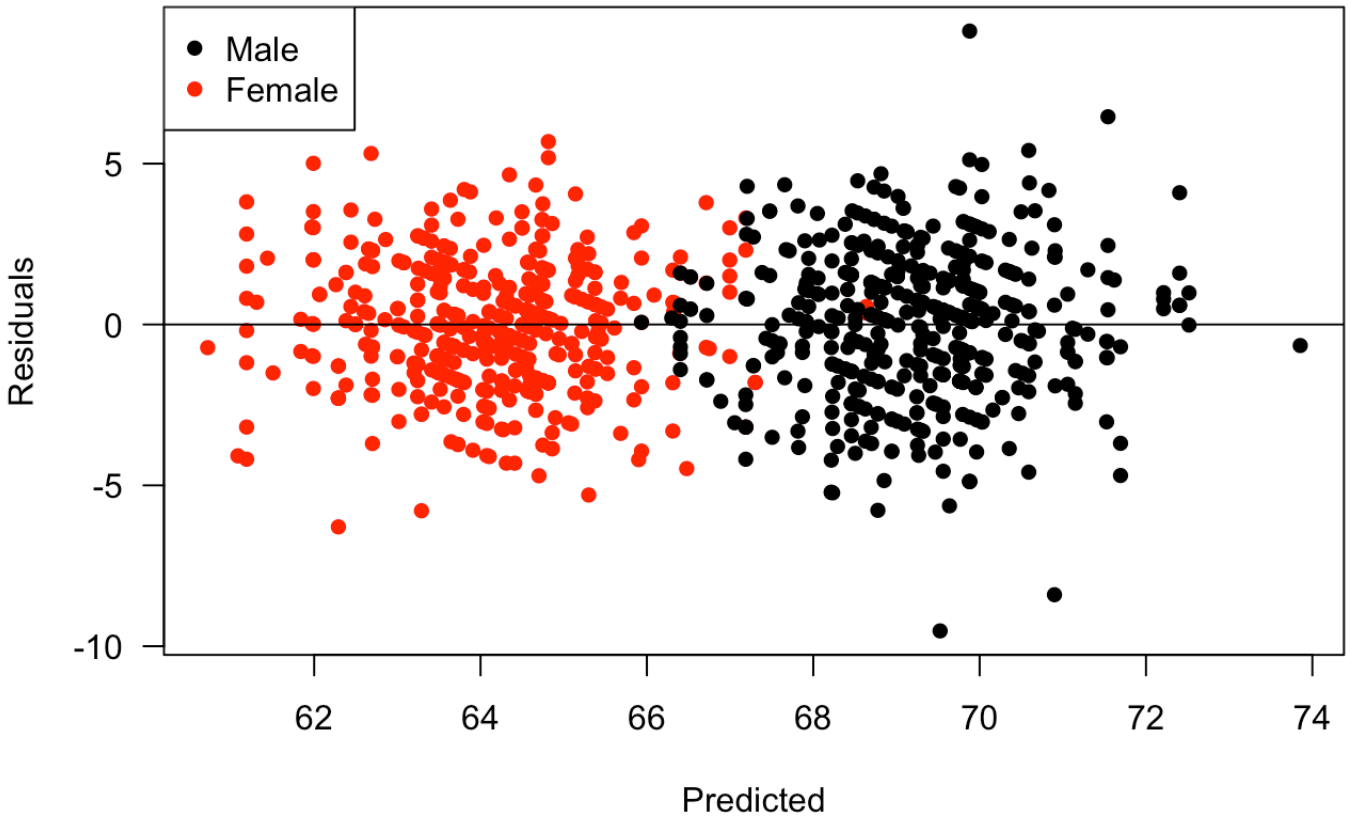
Inspect residuals. We can clearly see two clusters of points here. The cluster of points with smaller predicted heights belong to the female children and the other cluster of points belong to the male children.

```
plot(m1$residuals ~ m1$fitted.values,
     xlab="Predicted", ylab="Residuals",
     pch=16, las=1)
abline(h=0)
```



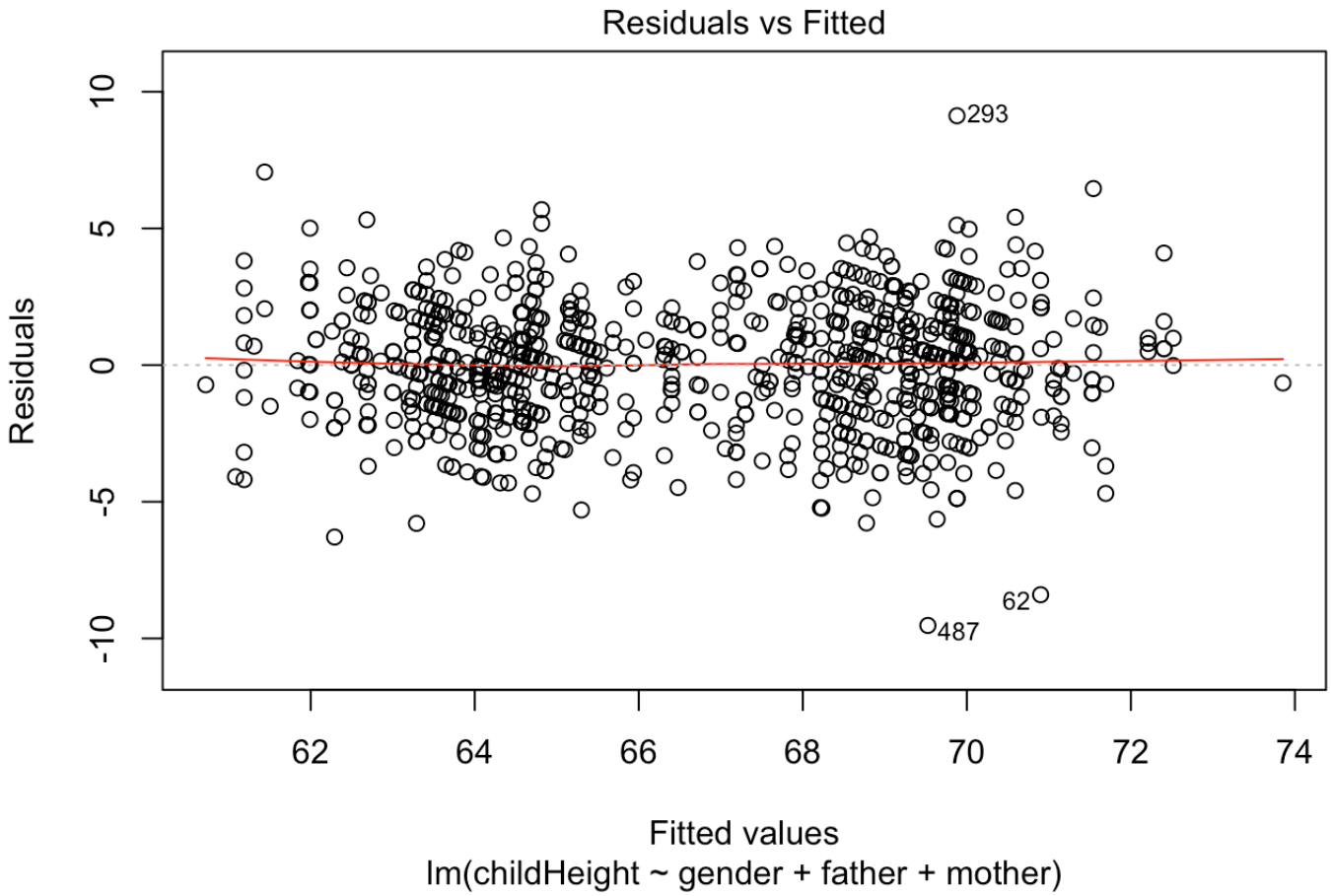
Color-code the residuals by gender

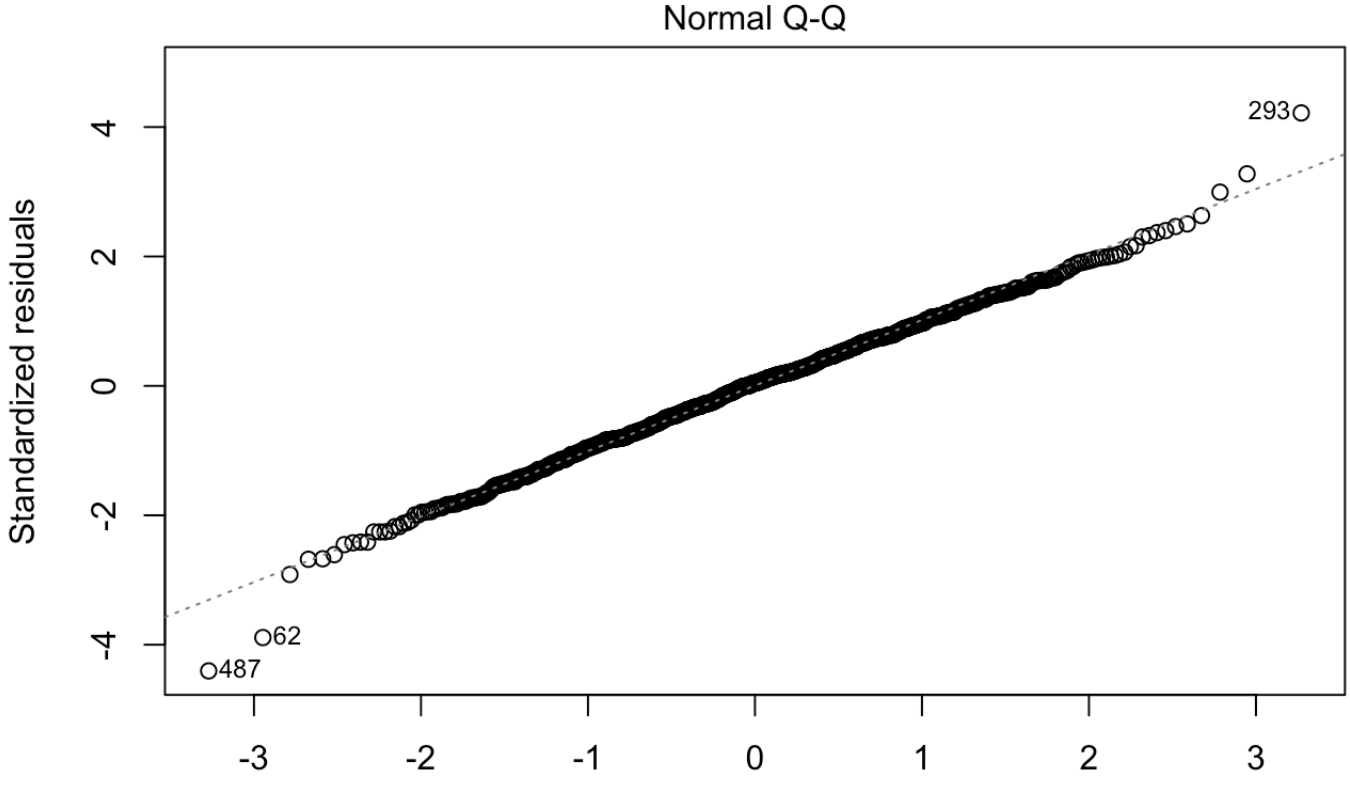
```
plot(m1$residuals ~ m1$fitted.values,
     col=GaltonFamilies$gender,
     xlab="Predicted", ylab="Residuals",
     pch=16, las=1)
abline(h=0)
legend("topleft",c("Male","Female"), pch=16, col=1:2)
```



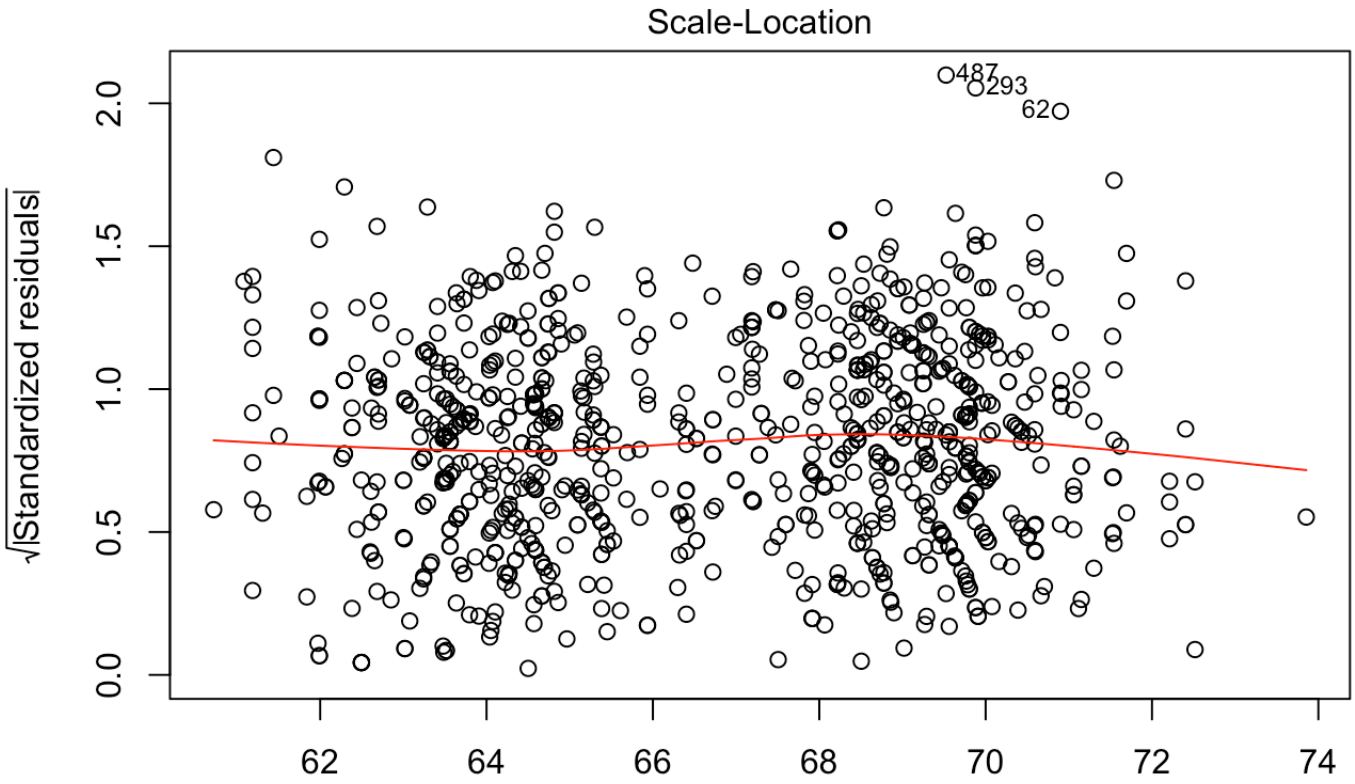
Diagnostics

```
plot(m1)
```

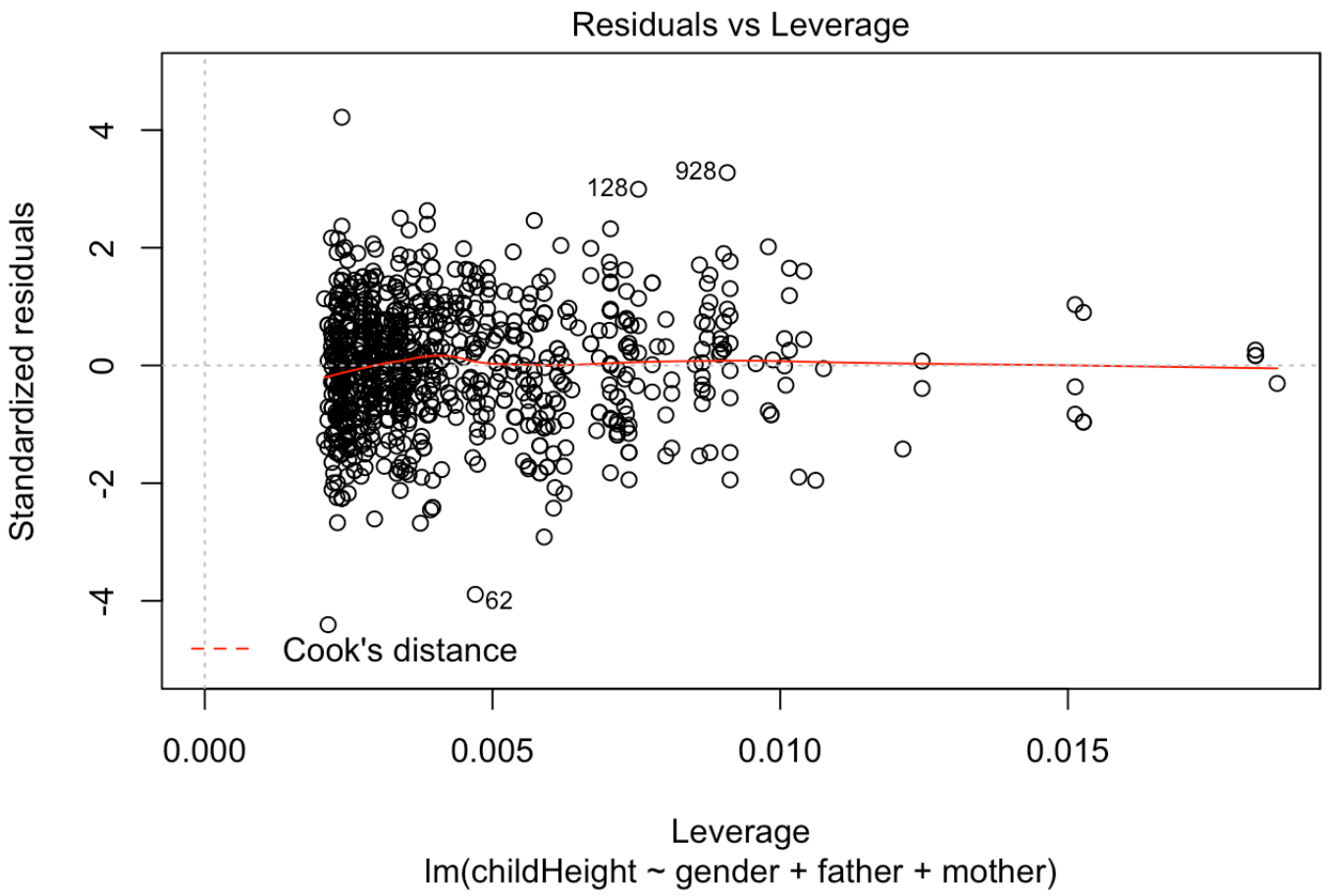

Residuals vs Fitted
Residuals
10
5
0
-5
-10
293
62
487
62
64
66
68
70
72
74
Fitted values
lm(childHeight ~ gender + father + mother)


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
293
62
487
-3
-2
-1
0
1
2
3

Theoretical Quantiles
lm(childHeight ~ gender + father + mother)



Fitted values
lm(childHeight ~ gender + father + mother)



Scale parents. See also https://stats.stackexchange.com/questions/254934/what-is-the-interpretation-of-scaled-regression-coefficients-when-only-the-predi/254982 (https://stats.stackexchange.com/questions/254934/what-is-the-interpretation-of-scaled-regression-coefficients-when-only-the-predi/254982)

```
m3 = lm(scale(childHeight) ~ gender + scale(father) + scale(mother), data=GaltonFamil
ies)
summary(m3)
```

```
##
## Call:
## lm(formula = scale(childHeight) ~ gender + scale(father) + scale(mother),
##     data = GaltonFamilies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.66108 -0.40938  0.02635  0.41517  2.54804
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.70666    0.02758   25.62   <2e-16 ***
## genderfemale  -1.45701    0.03962  -36.77   <2e-16 ***
## scale(father)  0.27181    0.01984   13.70   <2e-16 ***
## scale(mother)  0.20329    0.01984   10.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6048 on 930 degrees of freedom
## Multiple R-squared:  0.6354, Adjusted R-squared:  0.6342
## F-statistic: 540.3 on 3 and 930 DF,  p-value: < 2.2e-16
```

We would expect siblings to be somewhat similar in height as they share genetic factors through their parents and environmental factors through their shared upbringing.

We can model this structure of the data, children clustering in families, using linear mixed effects models. In addition to estimating population means (fixed effects) these models will also allow us to estimate how average family heights vary around these population means (random effects).

```
library(lme4)
library(lmerTest)

# The random effect for family indicates that the mean height of each family may diff
er from the population mean.
fit_me = lmer(childHeight ~ gender + father + mother + (1|family), data=GaltonFamilie
s)

# In addition to the gender fixed effect that we have already seen in the simple line
ar regression model, this model also provides us with an estimate of the variance in
average height between families (0.91) as well as the remaining (residual) variance w
ithin families (3.82).
summary(fit_me)
```

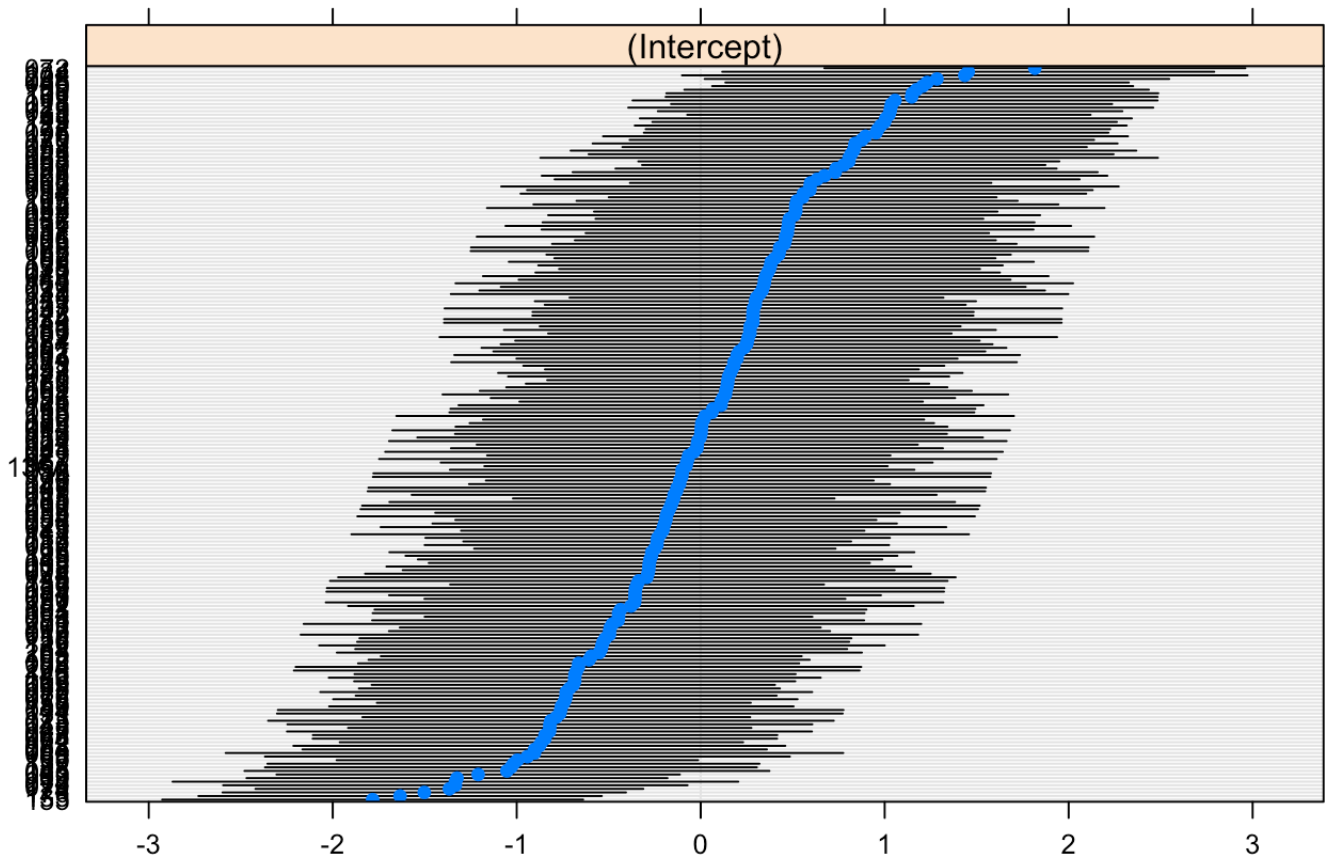
```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: childHeight ~ gender + father + mother + (1 | family)
##    Data: GaltonFamilies
##
## REML criterion at convergence: 4053.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2081 -0.5887  0.0073  0.6202  3.7316
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  family   (Intercept) 0.9073   0.9525
##  Residual             3.8197   1.9544
## Number of obs: 934, groups:  family, 205
##
## Fixed effects:
##               Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)   23.56134    3.65758 193.25285   6.442 9.16e-10 ***
## genderfemale  -5.22364    0.13522 877.55415 -38.631  < 2e-16 ***
## father         0.38161    0.03798 193.65643  10.049  < 2e-16 ***
## mother         0.30180    0.04245 178.66568   7.109 2.70e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) gndrfm father
## genderfemal  0.021
## father      -0.670 -0.031
## mother      -0.696 -0.022 -0.065
```

A dot plot, also known as a caterpillar plot, can help to visualise random effects. This plot shows the deviation from the mean population height for each family, together with standard errors. Note how some families fall clearly below or above the population mean.

```
library(lattice)

randoms = ranef(fit_me)
dotplot(randoms)
```

```
## $family
```



Model comparison with anova() and ranova(). In this case, the inclusion of the family random effect clearly improves model fit.

```
fit_lm = lm(childHeight ~ gender, data=GaltonFamilies)
## Re-fit model using ML, rather than REML
fit_me = lmer(childHeight ~ gender + (1|family), data=GaltonFamilies, REML=FALSE)
anova(fit_me, fit_lm)
```

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
fit_lm	3	4364.269	4378.788	-2179.135	4358.269	*NA*	*NA*	*NA*
fit_me	4	4164.460	4183.817	-2078.230	4156.460	201.8098	1	8.412093e-46

2 rows

```
ranova(fit_me)
```

	npar	logLik	AIC	LRT	Df	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>

<none>	4	-2078.230	4164.460	*NA*	*NA*	*NA*
(1 | family)	3	-2179.135	4364.269	201.8098	1	8.412093e-46

2 rows