GLM Workshop Answers

Crispin Jordan

01/11/2020

NOTE: Datasets and their explanations for Questions 2 - 4 were drawn from Whitlock & Schluter's text, "The analysis of biological data" (a recommended book for undergraduates and graduate students, alike).

Question 1 - GENE EXPRESSION

This experiment aims to test whether gene expression for the gene `CDH1` differs between tumor types (`Phenotype`, which has two levels: `GH` vs.`PRL`); it also aims to test whether gene expression differs between a tumor's `Invasiveness`: non-invasive, `NI` vs. invasive, `I`. The experiment collected data for all 4 possible combinations of these two Factors. We can see these factors, and the response (`CDH1`, i.e., a measure of expression for this gene) in the output, below:

```
pit <- read.table("pitTumours.csv",sep=',',header=TRUE)
str(pit)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Phenotype   : chr  "GH" "GH" "GH" "GH" ...
##  $ Invasiveness: chr  "I" "NI" "NI" "NI" ...
##  $ CDH1        : num  40.7 487 1247.1 1008.3 1016.1 ...
##  $ CDH2        : num  38.61 9.05 121.74 244.62 78.37 ...
```

```
summary(pit)
```

```
##   Phenotype         Invasiveness            CDH1               CDH2
##  Length:50          Length:50          Min.   :   0.9113   Min.   :  0.000
##  Class :character   Class :character   1st Qu.:  96.0359   1st Qu.:  5.995
##  Mode  :character   Mode  :character   Median : 236.6585   Median : 28.586
##                                        Mean   : 388.6692   Mean   : 75.597
##                                        3rd Qu.: 526.5761   3rd Qu.: 63.865
##                                        Max.   :1931.4788   Max.   :667.798
```

```
head(pit)
```

```
##   Phenotype Invasiveness       CDH1      CDH2
## 1        GH            I   40.72132  38.61383
## 2        GH           NI  486.97786   9.04871
## 3        GH           NI 1247.11522 121.74447
## 4        GH           NI 1008.30220 244.61669
## 5        GH           NI 1016.09776  78.36605
## 6        GH            I  276.05721  93.19346
```

We can see from the output of `str()` that the Factors, `Phenotype` and `Invasiveness` are class `chr`, but we want them to be Factors. Let's convert them to Factors and check out efforts:

```
pit$Invasiveness <- factor(pit$Invasiveness)
pit$Phenotype <- factor(pit$Phenotype)
str(pit)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Phenotype   : Factor w/ 2 levels "GH","PRL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Invasiveness: Factor w/ 2 levels "I","NI": 1 2 2 2 2 1 1 2 2 1 ...
##  $ CDH1        : num  40.7 487 1247.1 1008.3 1016.1 ...
##  $ CDH2        : num  38.61 9.05 121.74 244.62 78.37 ...
```

Excellent! They are both Factors now.

We're told that the data are independent, and we'll assume that they're randomly sampled. As we have two factors (`Phenotype` and `Invasiveness`), we will try to analyze these data using a 2-factor GLM.

First, let's look at the sample size for each treatment combination:

```
table(pit$Phenotype, pit$Invasiveness)
```

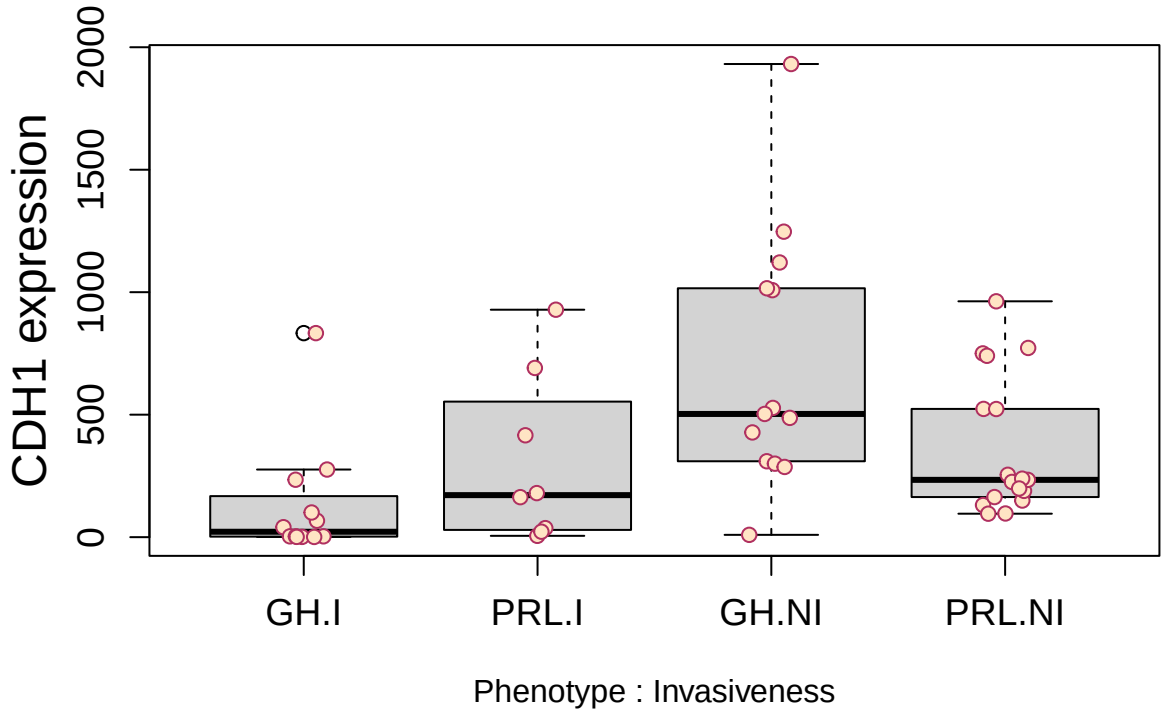
```
##
##        I NI
##   GH  12 13
##   PRL  8 17
```

Note:

- the experiment has all 4 treatment combinations;
- we have at least 2 data points for each treatment combination; this is required to model an interaction between the factors;
- as noted in the question, the data are **unbalanced**.

Now, let's plot the data. Our hypothesis is that gene expression (`CDH1`) will be affected by the tumor's invasiveness status (`Invasiveness`) and type (`Phenotype`). Therefore, we're saying that `CDH1` (expressions) will *depend* on `Invasiveness` and `Phenotype`; i.e., `CDH1` is the *dependent* variable:

```
boxplot(CDH1 ~ Phenotype*Invasiveness, data=pit,cex.axis = 1.2, ylab = "")
stripchart(CDH1 ~ Phenotype*Invasiveness, data=pit,
vertical = TRUE, method = "jitter",
pch = 21, col = "maroon", bg = "bisque",
add = TRUE)
mtext("CDH1 expression",2,line=2.5,cex=1.5)
```



What do we notice from this plot?

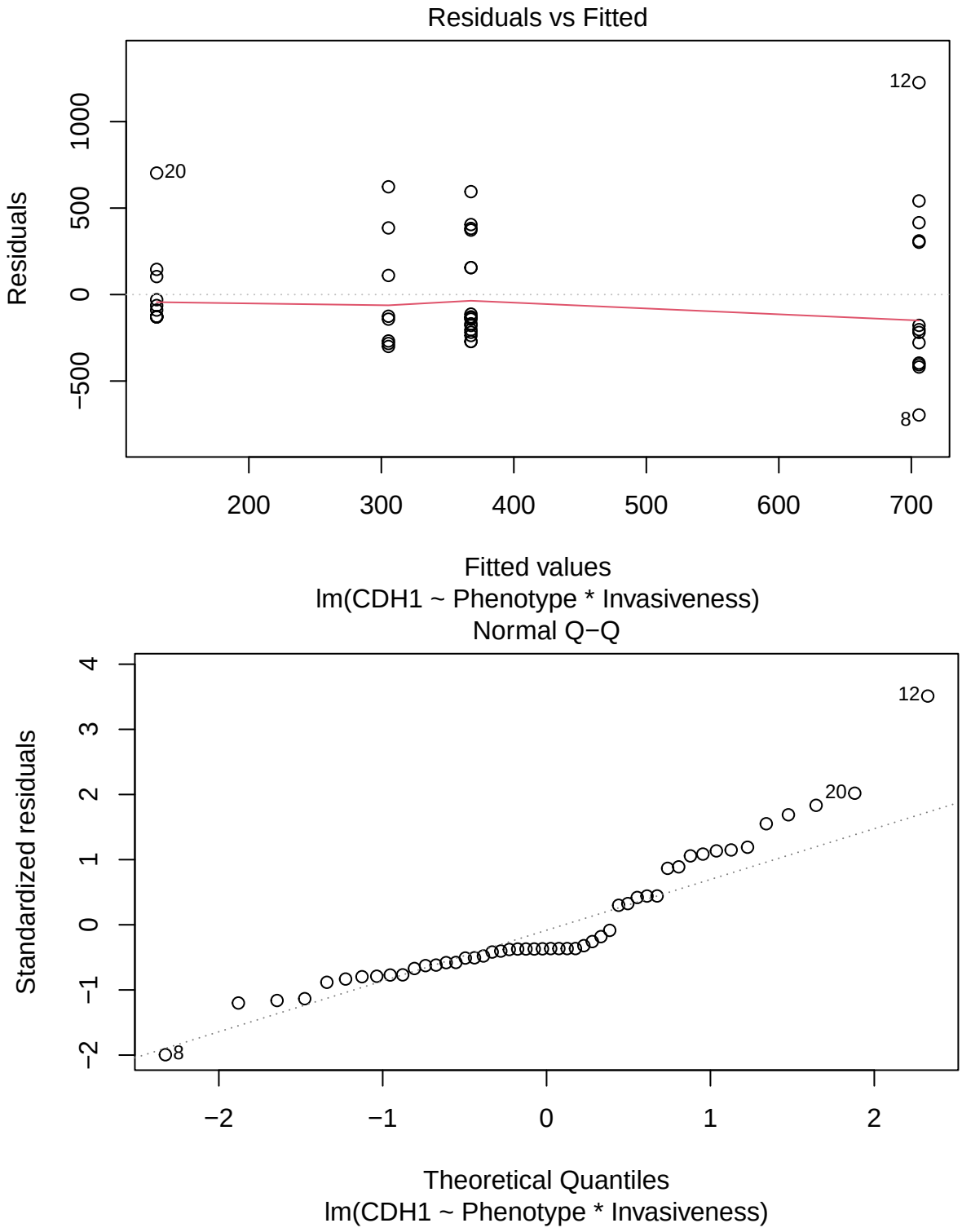
- The data likely violate the assumptions of equal variance; the boxplots have different breadths (also, notice that the breadth increases with the mean - often a sign of needing to transform the data).
- No outliers
- The boxplots are fairly asymmetrical; residuals are likely not normally distributed
- It looks like there may be in interaction between `Phenotype` and `Invasiveness`: notice that when `Invasiveness` is `I` ('invasive'; the boxplots on the left), the `PRL` phenotype tends to have higher expression than the `GH` phenotype; but (!) the opposite appears to be true when `Invasiveness` is `NI` (non-invasive).

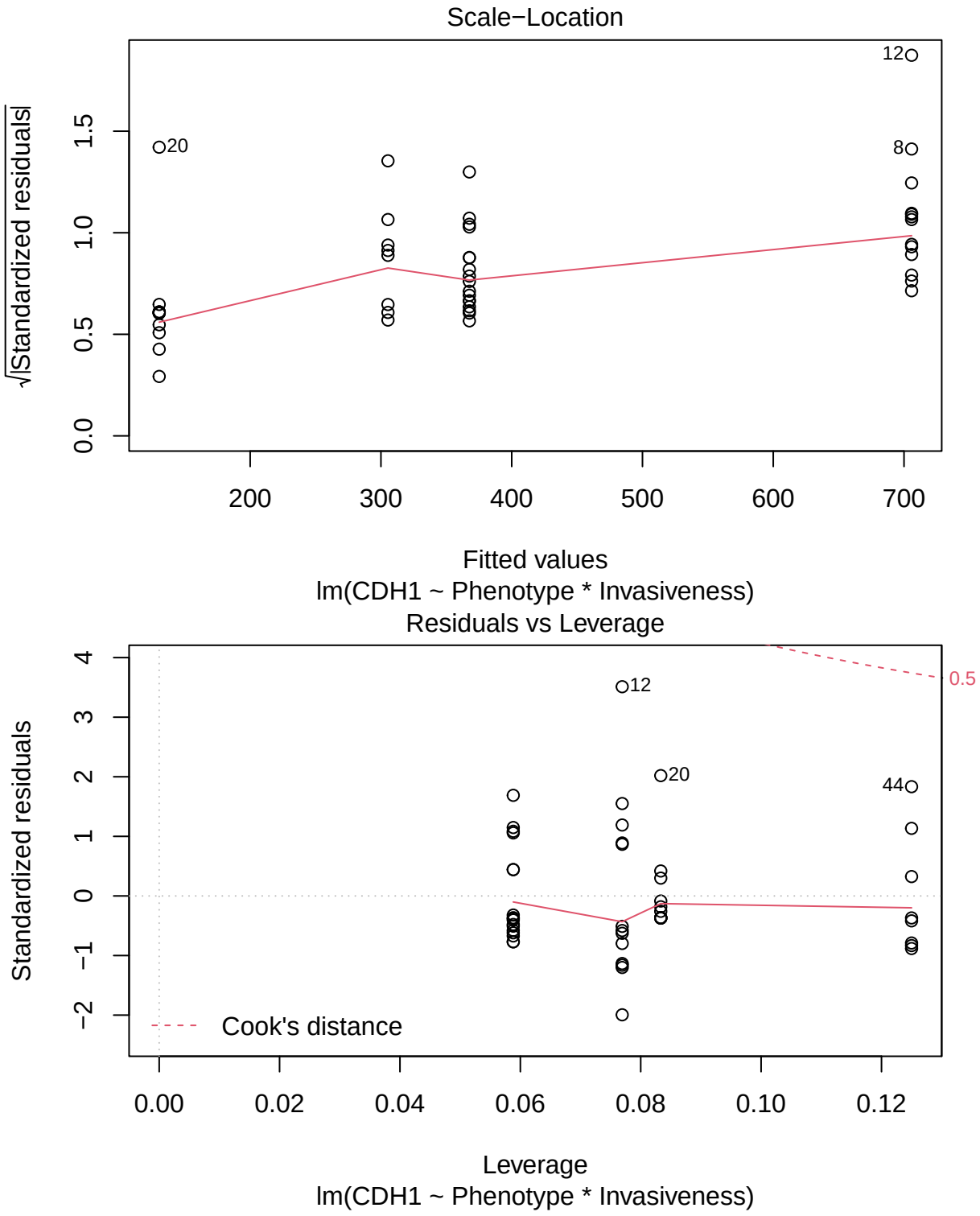
Let's model the data to determine whether our hunches about the assumptions are correct (modeling the two factors and their interaction); remember that `CDH1` is the *dependent* variable (and goes to the left of the tilda, `~`):

```
pit.lm <- lm(CDH1 ~ Phenotype*Invasiveness, data=pit)
```

Let's check the assumptions:

```
plot(pit.lm)
```

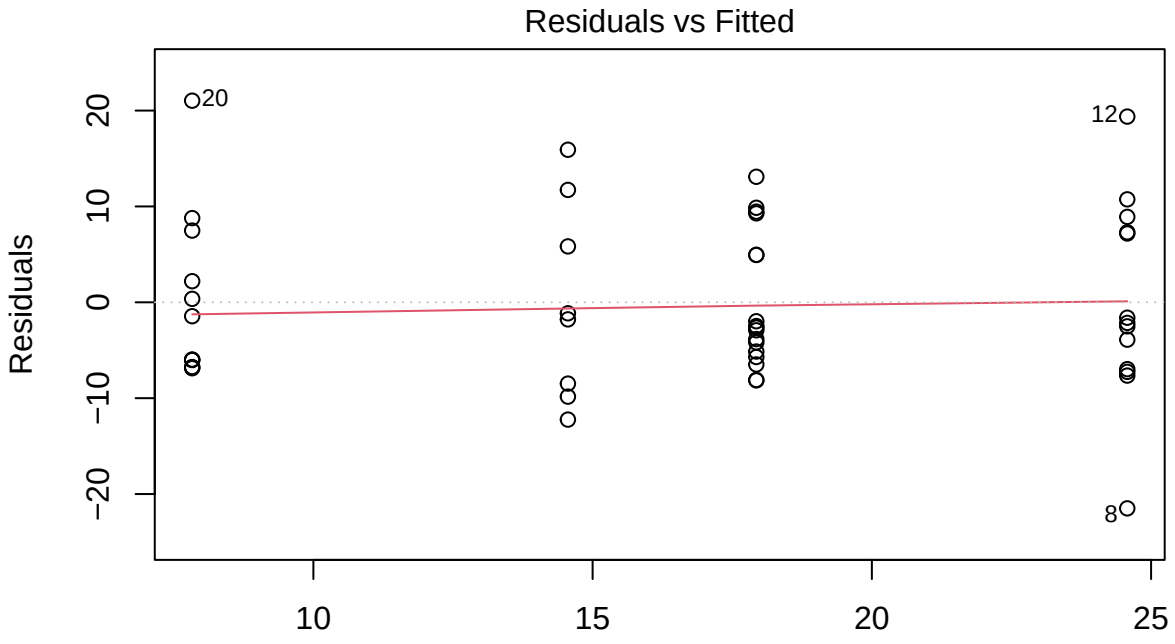

Residuals vs Fitted
Residuals
Fitted values
lm(CDH1 ~ Phenotype * Invasiveness)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(CDH1 ~ Phenotype * Invasiveness)

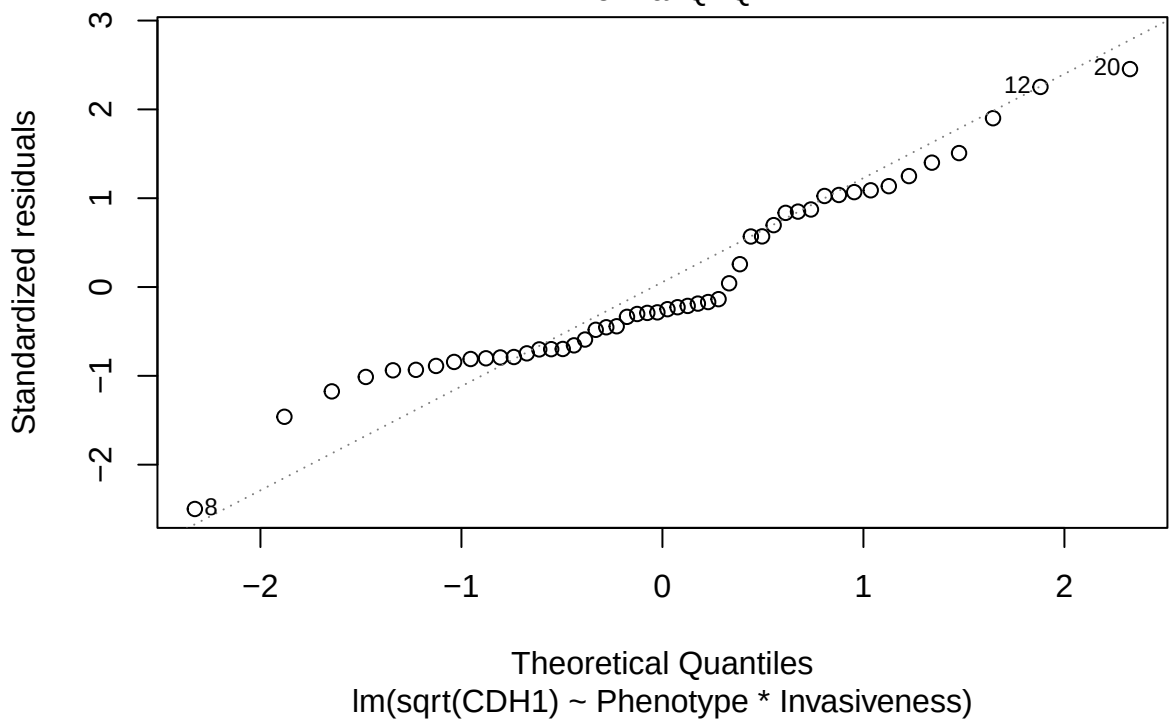


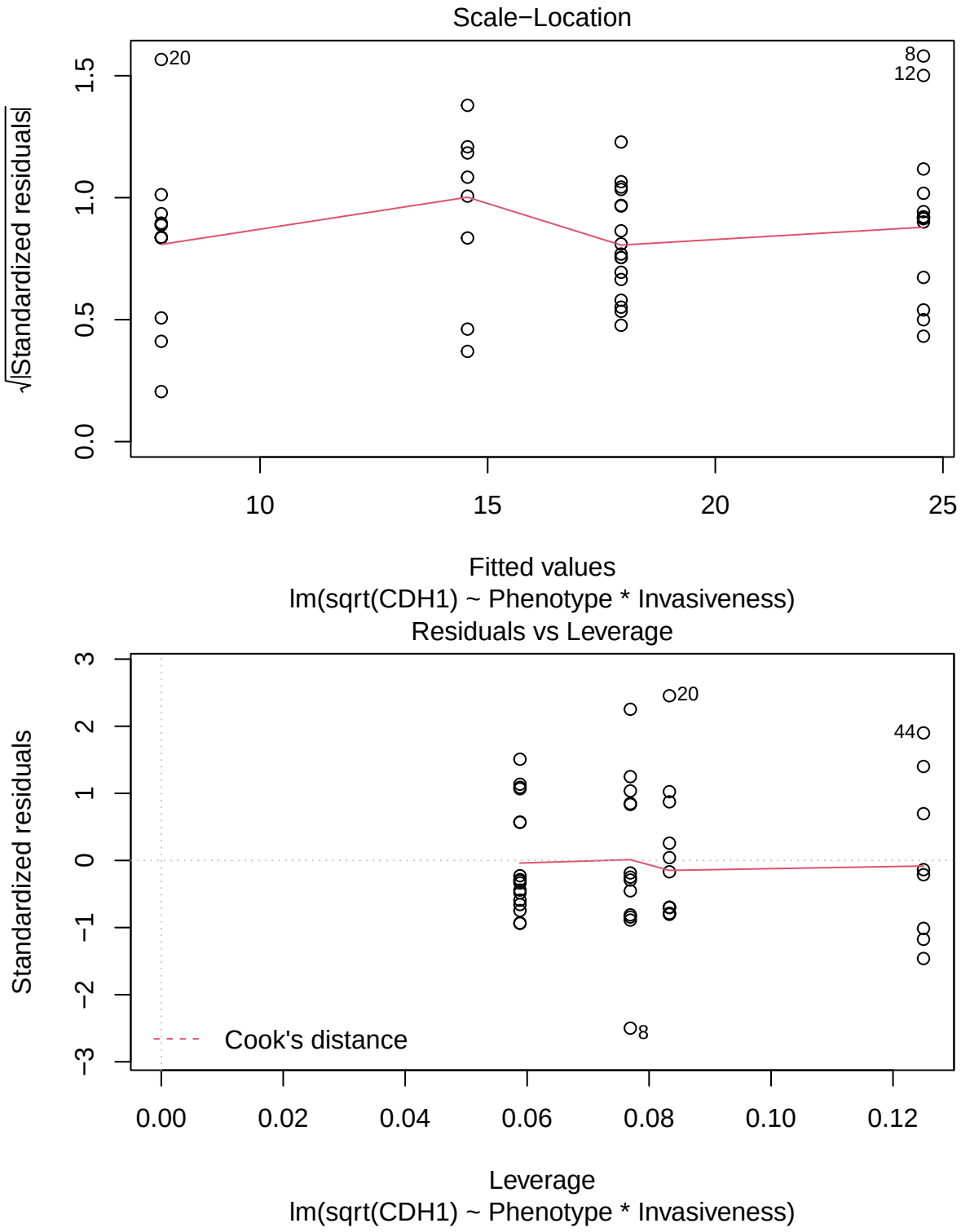
The assumptions are not met well at all: the first plot shows that the variance increases with the fitted values (violates assumption of similar variance among groups); the second plot shows strong deviation from a normal distribution (ideally, we want all points to fall along the dotted line). The third plot also indicates the data violate the assumption os equal variance (notice the red line is not flat).

Let's try transforming the data. We'll try a `sqrt()` (square-root) transformation:

```
p1.s <- lm(sqrt(CDH1) ~ Phenotype*Invasiveness, data=pit)
plot(p1.s)
```







The assumptions are now OK: The first plot shows that variance is relatively similar among the four treatment groups (i.e., the vertical spread is similar among the 4 vertical lines of residuals); the second plot shows that normality is not ideal, but not too terrible, either. The third plot is OK - the red line is not quite as flat as we'd like, but it is probably OK with respect to the assumption of equal variance.

Remember that we have an unbalanced design. As a result, we will use the following code to use for Type 3 Sum-of Square to calculate our p-values:

```
pit.lm.sq <- lm(sqrt(CDH1) ~ Phenotype*Invasiveness, data=pit, contrasts =
                  list(Phenotype = contr.sum, Invasiveness = contr.sum))
```

```
library(car)
```

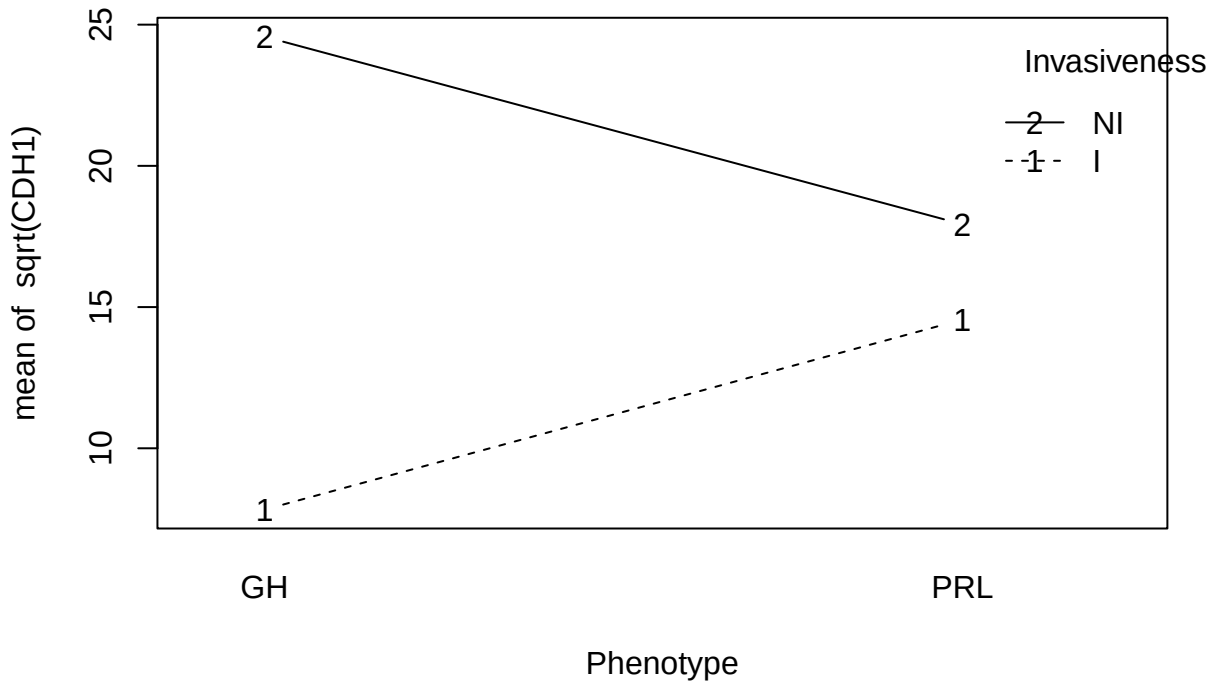
```
## Loading required package: carData
```

```
Anova(pit.lm.sq, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: sqrt(CDH1)
##                          Sum Sq Df  F value    Pr(>F)
## (Intercept)             12238.6  1 152.7300 3.207e-16 ***
## Phenotype                   0.0  1   0.0003 0.9870611
## Invasiveness             1176.2  1  14.6782 0.0003847 ***
## Phenotype:Invasiveness    519.6  1   6.4839 0.0142981 *
## Residuals                3686.1 46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results provide moderate evidence for an interaction between Phenotype and Invasiveness ($p = 0.0142981$). Let's look at in interaction plot to understand why (note that we `sqrt` transform the data in the interaction plot, too):

```
attach(pit)
interaction.plot(Phenotype,Invasiveness,sqrt(CDH1),type='b',legend=TRUE)
```




```
detach(pit)
```

This plot suggests that the interaction arises because `Invasiveness` more strongly impacts gene expression of the `GH` phenotype than the `PRL` phenotype. (We can see this in the boxplot, above, too: Notice the large difference between `GH.I` vs. `GH.NI`, whereas the difference between `PRL.I` and `PRL.NI` is less pronounced.)

With a general understanding of the results in hand, we should now obtain estimates of our effect sizes; exactly

which effect sizes we calculate, however, will depend on the focus of our biological interests / hypotheses. Let's imagine that we were primarily interested in the differences between `Invasiveness` levels; in this case we could determine effect sizes for differences between `I` and `NI` separately for each level of `Phenotype` (as we saw in the interaction plot). We will investigate this perspective, below. (Note that we could swap `Invasiveness` and `Phenotype` to investigate the effect sizes for `Phenotype`; we'll not perform this analysis simply to save space.)

```
library(emmeans)
pit.emmeans.in <- emmeans(pit.lm.sq, "Invasiveness", by = "Phenotype")
pit.emmeans.in
```

```
## Phenotype = GH:
##  Invasiveness emmean   SE df lower.CL upper.CL
##  I              7.83 2.58 46     2.63     13.0
##  NI            24.57 2.48 46    19.58     29.6
##
## Phenotype = PRL:
##  Invasiveness emmean   SE df lower.CL upper.CL
##  I             14.56 3.16 46     8.19     20.9
##  NI            17.93 2.17 46    13.56     22.3
##
## Results are given on the sqrt (not the response) scale.
## Confidence level used: 0.95
```

These results provide the `emmean` gene expressions (and SE) for `CDH1` for each treatment combination. Note that these means are on the transformed (`sqrt`) scale. *We can back-transforming square-root transformed data in emmeans, but this involves 're-gridding', which we will not demonstrate; see here:* https://cran.r-project.org/web/packages/emmeans/vignettes/transformations.html

Now, we wish to obtain estimates of the differences between the two levels of `Invasiveness` (i.e., 'contrasts', or effect size) with their SE's, as well as p-values:

```
pit.pairs.in <- pairs(pit.emmeans.in)
```

```
## Note: Use 'contrast(regrid(object), ...)' to obtain contrasts of back-transformed estimates
```

```
pit.pairs.in
```

```
## Phenotype = GH:
##  contrast estimate   SE df t.ratio p.value
##  I - NI     -16.74 3.58 46 -4.672  <.0001
##
## Phenotype = PRL:
##  contrast estimate   SE df t.ratio p.value
##  I - NI      -3.37 3.84 46 -0.879  0.3840
##
## Note: contrasts are still on the sqrt scale
```

These results provide the effect size (and SE) for the difference between the `I` and `NI` levels of `Invasiveness`, calculated separately for each type of `Phenotype`. Note that, as we saw, above, the difference between `I` and `NI` is much greater for `Phenotype GH`; in fact, formally, we do have any evidence for a difference between `NI` and `I` for `Phenotype` at level, `PRL`. We can also obtain the 95% CI's for these effect sizes:

```
confint(pit.pairs.in)
```

```
## Phenotype = GH:
##  contrast estimate   SE df lower.CL upper.CL
##  I - NI     -16.74 3.58 46    -24.0    -9.53
```

```
##
## Phenotype = PRL:
##  contrast estimate   SE df lower.CL upper.CL
##  I - NI      -3.37 3.84 46    -11.1     4.35
##
## Note: contrasts are still on the sqrt scale
## Confidence level used: 0.95
```

The difference of (square-root transformed) gene expression for gene `CDH1` between `NI` and `I` ranges from -24.0 to -9.53 when `Phenotype` is GH; but the comparable 95% CI for the effect size in `Phenotype PRL` is less pronounced, and includes 0 as a possible difference: -11.1 to 4.35. Are these 'important' effects? That's hard for me to say, as I know very little about tumors; but presumably another biologist could use these effect size estimates (and 95% CI's) to gain a sense of the biological impact of `Phenotype` and `Invasiveness` in this experiment.

Question 2

In this experiment, Cook et al. (1993) wished to compare a hippocampal volume loss (expressed as a percent) in patients with drug-resistant epilepsy among subjects with different histories of seizures: a history of childhood febrile seizures (`CFS`), childhood non-febrile seizures (`No CFS`), and `No seizures`.

Let's look at the data:

```
seiz <- read.table("seizuresCFS.csv",sep=',',header=TRUE)
str(seiz)
```

```
## 'data.frame':    107 obs. of  3 variables:
##  $ subject         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ group           : chr  "CFS" "CFS" "CFS" "CFS" ...
##  $ hippoVolumeRatio: num  47 51.2 55.3 56.5 56.6 56.6 59.4 60.7 64.7 59.4 ...
```

```
summary(seiz)
```

```
##     subject         group           hippoVolumeRatio
##  Min.   :  1.0   Length:107         Min.   : 47.00
##  1st Qu.: 27.5   Class :character   1st Qu.: 88.00
##  Median : 54.0   Mode  :character   Median : 96.40
##  Mean   : 54.0                      Mean   : 89.61
##  3rd Qu.: 80.5                      3rd Qu.: 98.45
##  Max.   :107.0                      Max.   :100.00
```

```
head(seiz)
```

```
##   subject group hippoVolumeRatio
## 1       1   CFS             47.0
## 2       2   CFS             51.2
## 3       3   CFS             55.3
## 4       4   CFS             56.5
## 5       5   CFS             56.6
## 6       6   CFS             56.6
```

This output reveals three columns: `subject` (which we will ignore as it is not useful to us), `group`, which designates a subject's history of seizures, and `hippoVolumeRatio`, the measure of hippocampal volume loss. Note that smaller values of `hippoVolumeRatio` denote greater volume **loss**.

Note that the `str()` function indicates that `group` is currently type `chr`; let's change it to a Factor:

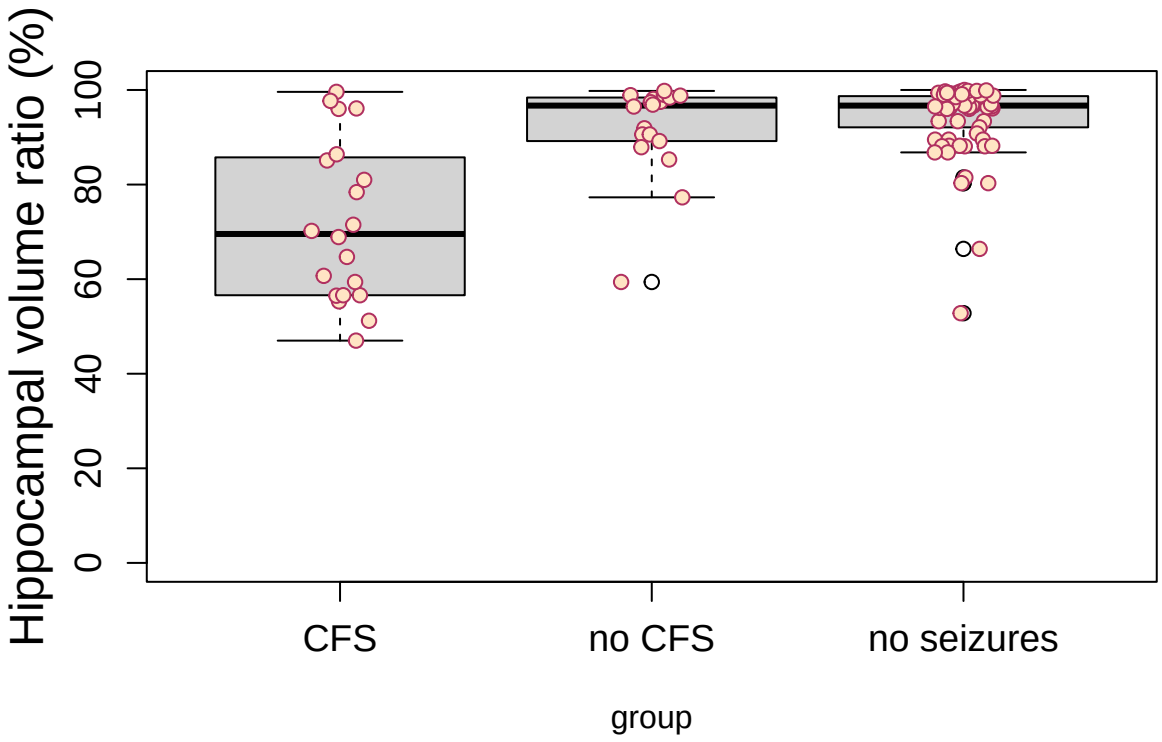
```
seiz$group <- factor(seiz$group)
str(seiz)
```

```
## 'data.frame':    107 obs. of  3 variables:
##  $ subject         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ group           : Factor w/ 3 levels "CFS","no CFS",..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ hippoVolumeRatio: num  47 51.2 55.3 56.5 56.6 56.6 59.4 60.7 64.7 59.4 ...
```

OK - `group` is now a Factor, which is what we want.

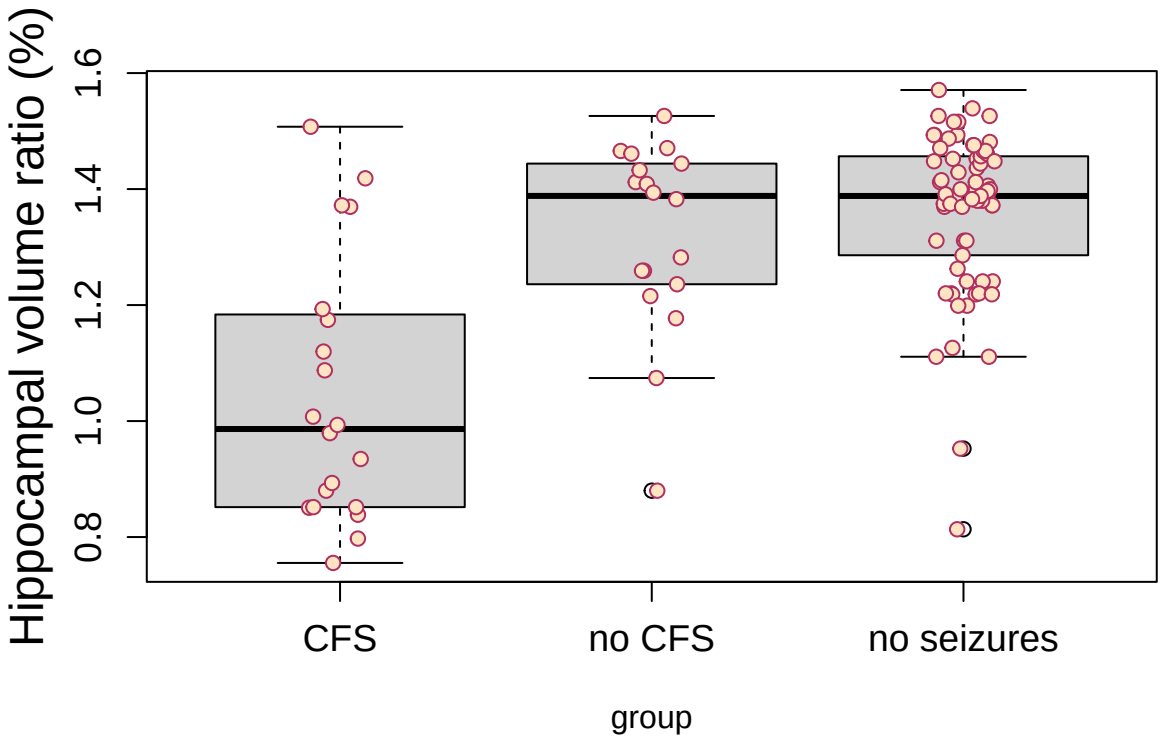
Let's plot the data. The hypothesis is that the volume loss will depend on the history of seizures (`group`); therefore, `hippoVolumeRatio` is the *dependent* variable:

```
boxplot(hippoVolumeRatio ~ group, data=seiz,cex.axis = 1.2, ylab = "", ylim = c(0,100))
stripchart(hippoVolumeRatio ~ group, data=seiz,
vertical = TRUE, method = "jitter",
pch = 21, col = "maroon", bg = "bisque",
add = TRUE)
mtext("Hippocampal volume ratio (%)",2,line=2.5,cex=1.5)
```



Recall that the data are expressed as a percent. As a result, we do not expect the data to be normally distributed, and the boxplot, above, matches this expectation (boxplots loox asymmetrical). Let's try transforming the data and inspecting the plot, again, before we run our model. An arcsine-square root transformation can be helpful for proportions; to use this transformation, we need to remember to change the data from a percent to a proportion by dividing all data by 100 before applying the transformation:

```
boxplot(asin(sqrt(hippoVolumeRatio/100)) ~ group, data=seiz,cex.axis = 1.2, ylab = "")
stripchart(asin(sqrt(hippoVolumeRatio/100)) ~ group, data=seiz,
vertical = TRUE, method = "jitter",
pch = 21, col = "maroon", bg = "bisque",
add = TRUE)
mtext("Hippocampal volume ratio (%)",2,line=2.5,cex=1.5)
```




```
seiz$Asin <- asin(sqrt(seiz$hippoVolumeRatio/100))
```

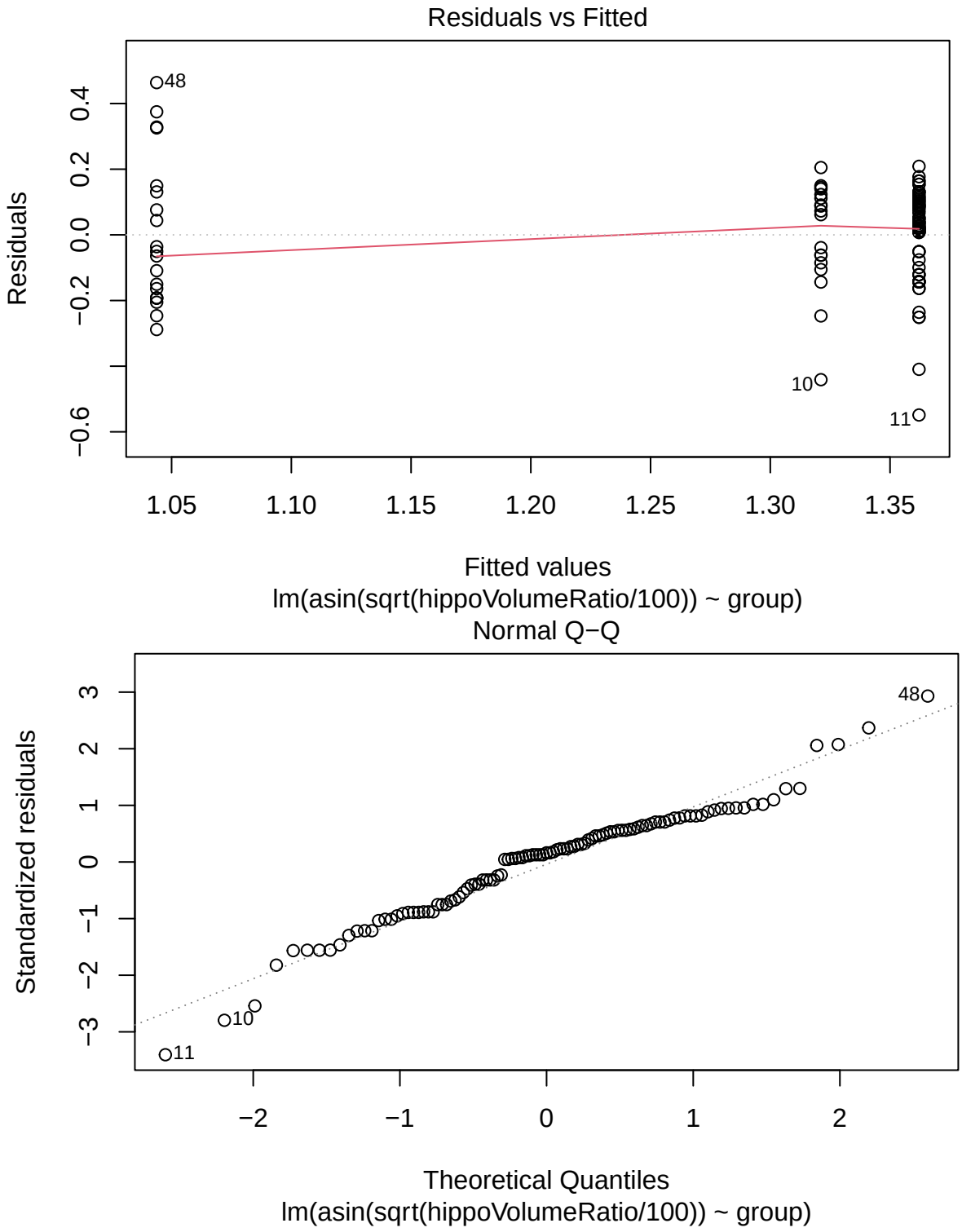
Before continuing, please note that we assess assumption by looking at residual plots, *not* from simply looking at plots of the data, like those, above. Indeed, we will formally check the assumptions using residual plots, below. Why did we try transforming the data this time, after examining the boxplots? Well, it has to do with the fact that we knew at the beginning the the data are expressed as proportions; this knowledge leads us to expect non-normally distributed data, even without plotting the data! We also know that arcsine square-root transformation can be an effective way to deal with data such as these. So it is this *a priori* knowledge that the data were proportions that led us to transform the data at this stage, not the plots, per se. That said, we could have tried to model the data without transforming before we used arcsine square-root.

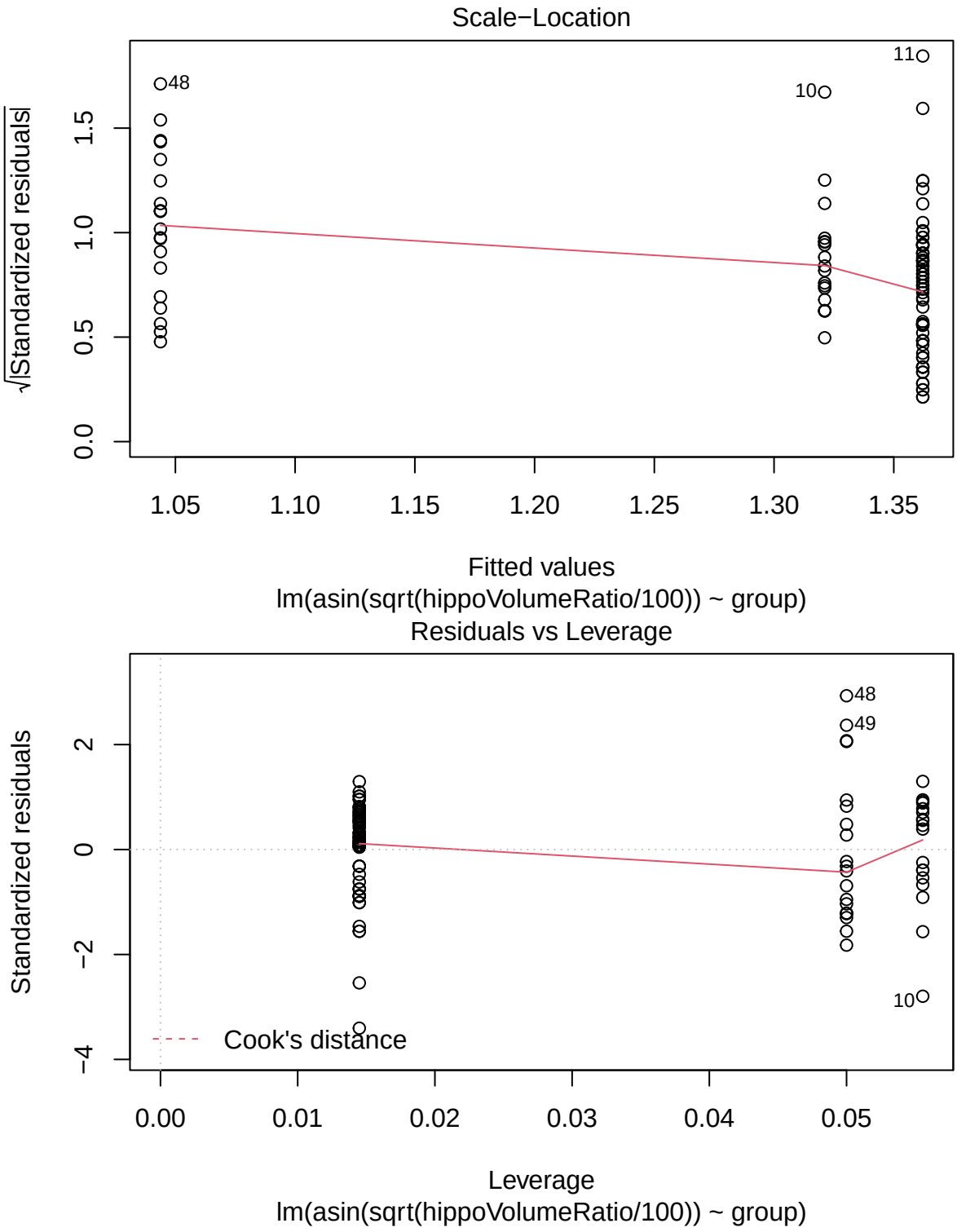
This plot suggests that the data will now meet our assumptions. Here's what I notice:

- The breadth of the boxplots is roughly similar, suggesting that the data will meet the assumption of equal variance;
- The boxplots are roughly symmetrical, suggesting that the transformed data will meet the assumption of normality;
- There are a few unusual data-points (exceptionally large values for `CFS`, and some unusually small values for `No CFS` and `No seizures`), but these are not too great
- We might predict that hippocampal volume loss will be great for `CSF` (has history of febrile seizures) compared to the remaining groups, which lacked pre-febrile seizures; the latter two groups likely differ little.

Let's check our predictions when we model our data. Again, remember that `hippoVolumeRatio` (transformed) is the dependent variable:

```
seiz.lm <- lm(asin(sqrt(hippoVolumeRatio/100)) ~ group, data=seiz)
plot(seiz.lm)
```


Residuals vs Fitted
Residuals
48
10
11
0.4
0.2
0.0
−0.2
−0.6
1.05
1.10
1.15
1.20
1.25
1.30
1.35
Fitted values
lm(asin(sqrt(hippoVolumeRatio/100)) ~ group)
Normal Q−Q
Standardized residuals
48
10
11
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
Theoretical Quantiles
lm(asin(sqrt(hippoVolumeRatio/100)) ~ group)



- The first plot suggests that the data likely meet the assumption of equal variance, but, to be honest, I'm not 100% certain. I find this a tough call. If I aimed to publish these data, I would be tempted to use another approach (e.g., a type of randomization test that is suitable for 1-factor experiments) to analyze these data because I don't fully trust that the data meet the assumption of equal variance. But, we've not learned that yet. So, we'll press on, simply to hone our skills with 1-factor GLM's. (Besides, this example has also taught you that sometimes we might need to find an alternate approach

if transformation does not work - that's a worthwhile lesson.)
- The residuals look relatively normally distributed (see the second plot)
- The third plot confirms our worries from the first plot: the red line is not terribly flat, and the data likely violate the assumption of equal variance. But, for reasons given in discussion of plot 1, we'll carry on, remembering that we'd be more cautious if we aimed to actually publish these data.

If we assume that the assumptions are met, we can check our p-values:

```
anova(seiz.lm)
```

```
## Analysis of Variance Table
##
## Response: asin(sqrt(hippoVolumeRatio/100))
##            Df Sum Sq Mean Sq F value    Pr(>F)
## group       2 1.5853 0.79265  30.072 4.943e-11 ***
## Residuals 104 2.7412 0.02636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is very small (p = 4.943e-11), providing very strong evidence that the (transformed) means differ among some levels of `group` (that is, if we trust our assumptions, which is questionable. But anyway...). It is comforting, at least, that we predicted some differences among levels of `group` based on our boxplots. Let's do a post-hoc test to further test our original predictions:

```
library(emmeans)
seiz.emmeans <- emmeans(seiz.lm, "group")
```

```
## Warning in (function (object, at, cov.reduce = mean, cov.keep = get_emm_option("cov.keep"), : There a
## Auto-detection of the response transformation may be incorrect
```

```
seiz.emmeans
```

```
##  group       emmean     SE  df lower.CL upper.CL
##  CFS           1.04 0.0363 104    0.972     1.12
##  no CFS        1.32 0.0383 104    1.245     1.40
##  no seizures   1.36 0.0195 104    1.323     1.40
##
## Results are given on the asin(sqrt(mu)) (not the response) scale.
## Confidence level used: 0.95
```

This output provides the `emmean` for each level of group (and SE), on the transformed data scale. As we saw in our plot, hippocampal volume seems to have decreased most for `CFS` subjects; inspecting the SE's (and remembering that a 95% CI can be roughly equal to twice the size of the SE), we do not expect any strong evidence between `No CFS` and `No seizure`, but we do expect evidence that these `group` levels differ from `CFS`. Let's check this expectation using the `pairs()` function:

```
seiz.pairs <- pairs(seiz.emmeans)
```

```
## Note: Use 'contrast(regrid(object), ...)' to obtain contrasts of back-transformed estimates
```

```
seiz.pairs
```

```
##  contrast             estimate     SE  df t.ratio p.value
##  CFS - no CFS           -0.277 0.0527 104 -5.259  <.0001
##  CFS - no seizures      -0.318 0.0412 104 -7.721  <.0001
##  no CFS - no seizures   -0.041 0.0430 104 -0.953  0.6081
##
## Note: contrasts are still on the asin.sqrt scale
## P value adjustment: tukey method for comparing a family of 3 estimates
```

These results indicate the effect sizes of `group` with their SE's; we should report these values, and remember to tell a reader that these values are on the transformed scale. Note that the size of the difference between `CFS` and the remaining two treatments (`No CFS` and `No seizures`) is much greater than any possible difference between the latter two treatment levels. The p-values also provide strong evidence that `CFS` differs from the remaining two treatments, but no evidence that these latter treatments (`no CFS`, `no seizures`) differ from each other.

Let's get some 95% CI's for these effect sizes and see what else we can learn:

```
confint(seiz.pairs)
```

```
##  contrast             estimate     SE  df lower.CL upper.CL
##  CFS - no CFS           -0.277 0.0527 104   -0.403  -0.1520
##  CFS - no seizures      -0.318 0.0412 104   -0.416  -0.2203
##  no CFS - no seizures   -0.041 0.0430 104   -0.143   0.0612
##
## Note: contrasts are still on the asin.sqrt scale
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
```

The range of 95% CI's for the difference between `CFS` and `no CFS` (-0.403 to -0.1520) is very similar to the interval for `CFS` and `no seizures` (-0.416 to -0.2203); this is consistent with a conclusion that the `no CFS` and `no seizures` are similarly different from `CFS`. Note that the 95% CI's for the difference between `No CFS` and `No seizures` is range from -0.143 to 0.0612, which includes zero; therefore, it is plausible that there is no difference in hippocampal volume loss for these two groups. However, this 95% CI mostly includes negative values (i.e., this 95% CI is not centered on zero.) As this 95% CI was based on the contrast, 'no CFS - no seizures', a tendency towards a negative value would suggest a tendency towards less hippocampal loss for individuals in the `no seizures` group. This tendency might be worth remembering for future research.

Let's end by considering the experimental design in a bit more detail. We had three levels of `group`:

- `CFS`, which had febrile seizues
- `no CFS`, which has seizures but not febrile seizures
- `no seizures`, which had no seizures (febrile or otherwise).

The `no CFS` serves as a Control treatment in this experiment: if we only compared `CFS` to `no seizures` there could be two differences between them: 1) the presence of seizures, and 2) the febrile nature of the seizures in `CFS`. With only these two treatments we cannot infer whether seizures, per se, decrease hippocampal volume, or whether *febrile* seizures decrease volume. This is where the group, `no CFS` is so useful: this group had seizures, but they were not febrile. The fact that hippocampal volume loss was very similar between `no CFS` and `no seizures` suggests that **seizures, per se, have little effect on hippocampal volume loss** (see the 95% CI's for the contrast between `no CFS` and `no seizures`). In contrast, we did see larger losses in hippocampal value when *febrile* seizures were present (in the `CFS` group), which would imply that it is the febrile nature of seizures that especially leads to hippocampal volume loss.

Question 3 - MUTATIONS

This experiment aims to test whether the number of mutations that are passed on to human offspring is related to the age of the offspring's father at the time of fertilization. Why would we be so interested in the father's age, and perhaps less interested in the mother's age in this respect? The answer has to do with meiosis: males continually produce new gametes (sperm) throughout their lives, and this repeated cell-division increases the opportunity for mutations to accumulate in the cells that produce sperm. As males get older, there's more time for mutations to arise in these male cells. On the other hand, females are born with all of their eggs already produced, so opportunity for these types of mutations (single-nucleotide mutations, I assume) does not increase with age in females. Cool, eh?

Let's get to know the data:

```
mut <- read.table("mutations.csv",sep=',',header=TRUE)
str(mut)
```

```
## 'data.frame':    21 obs. of  2 variables:
##  $ AgeOfFather         : int  16 18 20 19 22 24 24 24 25 28 ...
##  $ numberOfNewMutations: int  39 41 39 49 50 54 55 61 57 52 ...
```

```
summary(mut)
```

```
##   AgeOfFather   numberOfNewMutations
##  Min.   :16.0   Min.   :39.0
##  1st Qu.:24.0   1st Qu.:52.0
##  Median :28.0   Median :57.0
##  Mean   :27.1   Mean   :58.1
##  3rd Qu.:32.0   3rd Qu.:67.0
##  Max.   :37.0   Max.   :83.0
```

```
head(mut)
```

```
##   AgeOfFather numberOfNewMutations
## 1          16                   39
## 2          18                   41
## 3          20                   39
## 4          19                   49
## 5          22                   50
## 6          24                   54
```

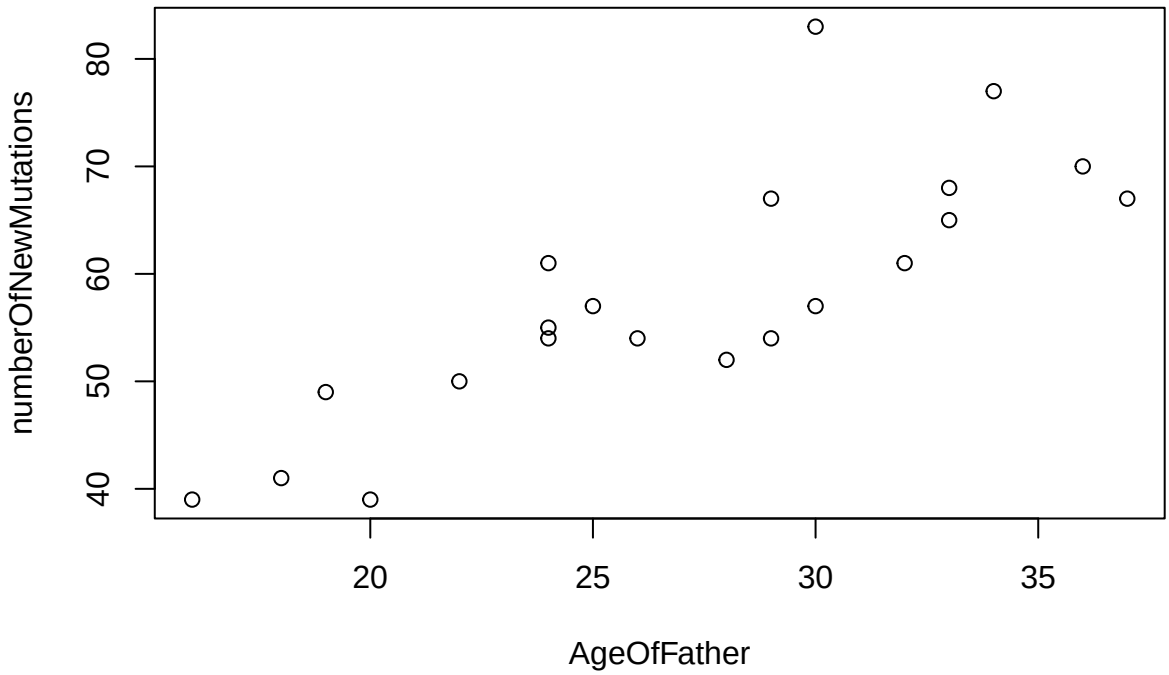
This output reveals that the dataframe, `mut`, has only two variables, whose names are self-explanatory: `AgeOfFather` and `numberOfNewMutations`.

With this is mind, we should start by plotting our data. However, when plotting these data we need to consider which variable is the *dependent* variable? Are we hypothesizing that:

1) the father's age will affect the number of mutations? Or...
2) the number of mutations will determine the father's age?

The first option makes more sense; therefore, `numberOfNewMutations` is the dependent variable and will be listed first in out plot commands and our models:

```
plot(numberOfNewMutations~AgeOfFather,data=mut)
```

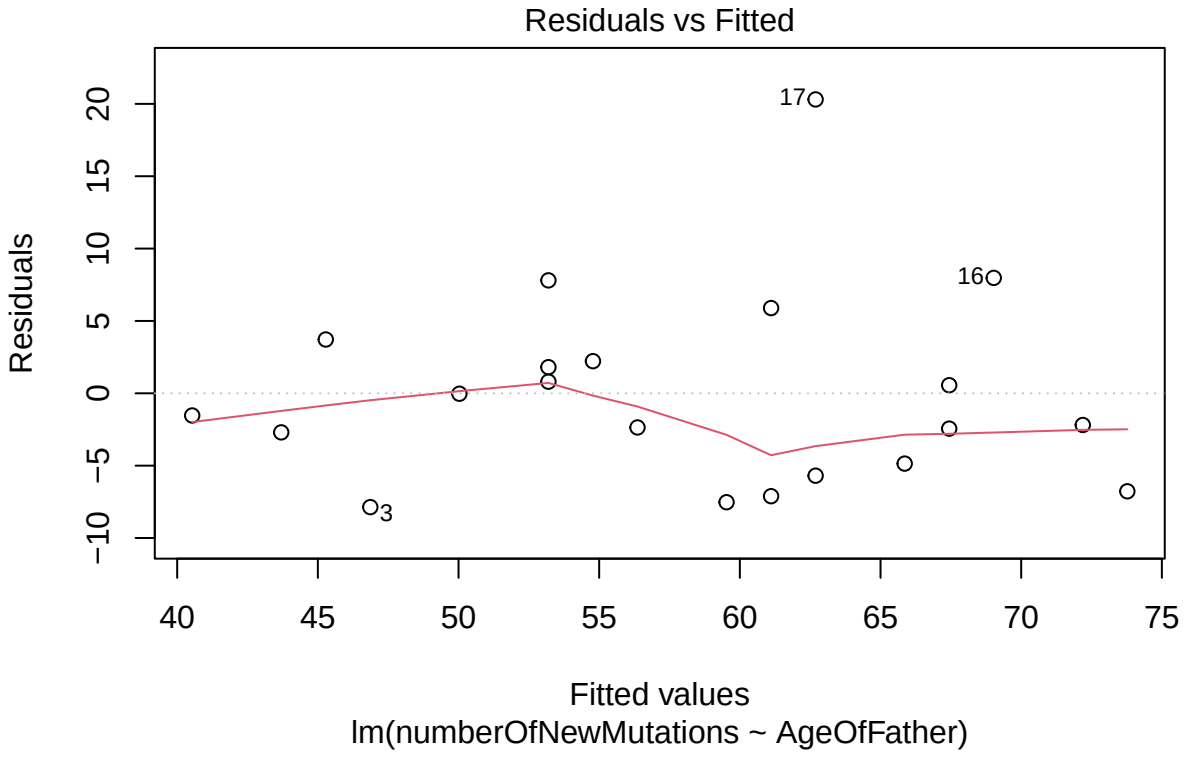


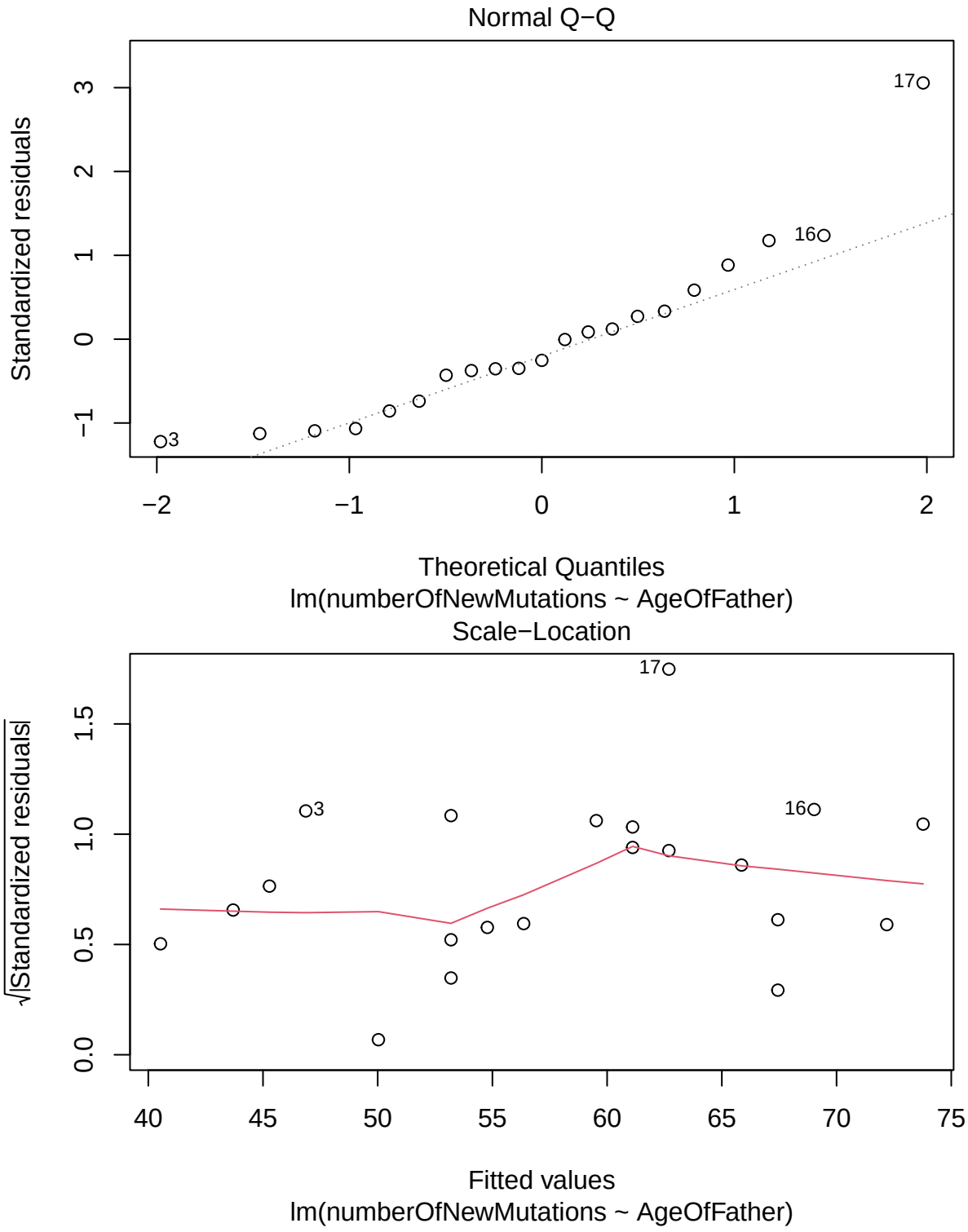
We see that the number of mutations seem to increase with age! Moreover, the relationship appears to be a straight line, suggesting that we can eaily model this relationship as:

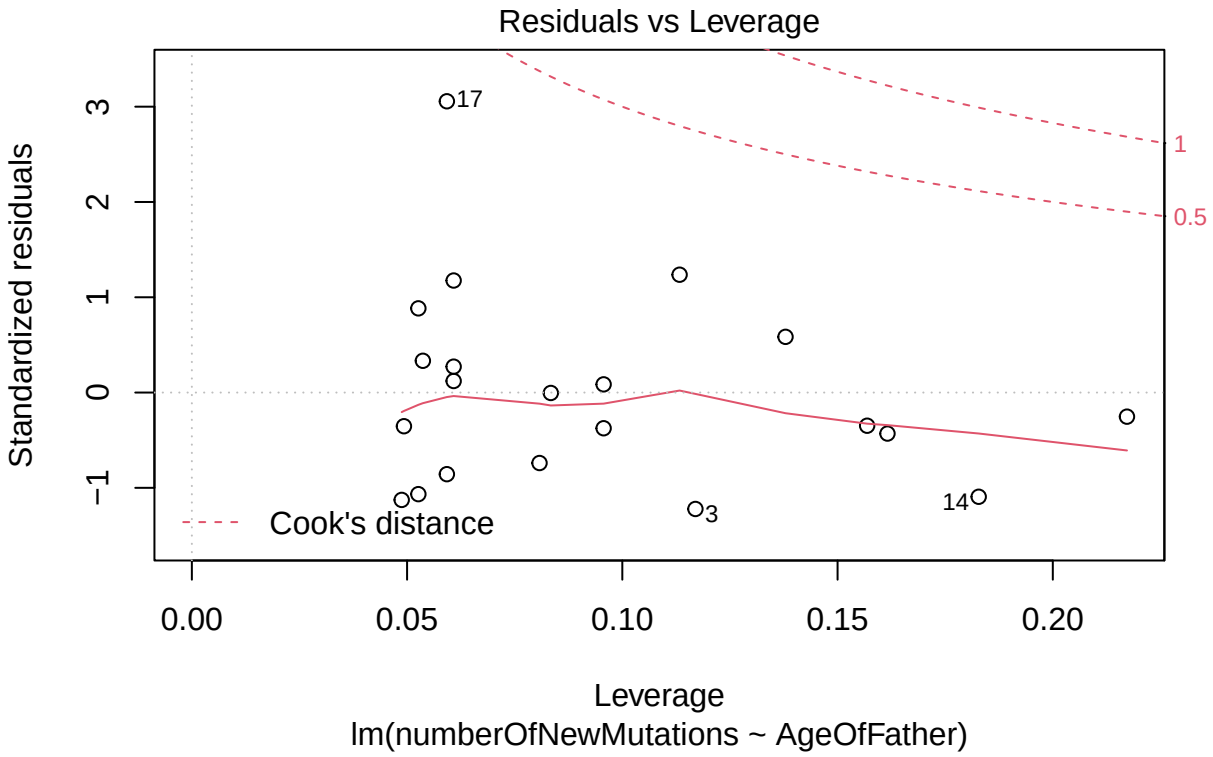
```
mut.lm <- lm(numberOfNewMutations~AgeOfFather,data=mut)
```

We'll assume (from the original research) that subjects were chosen randomly and are independent. Therefore, let's check the assumptions of equal variance and normality:

```
plot(mut.lm)
```



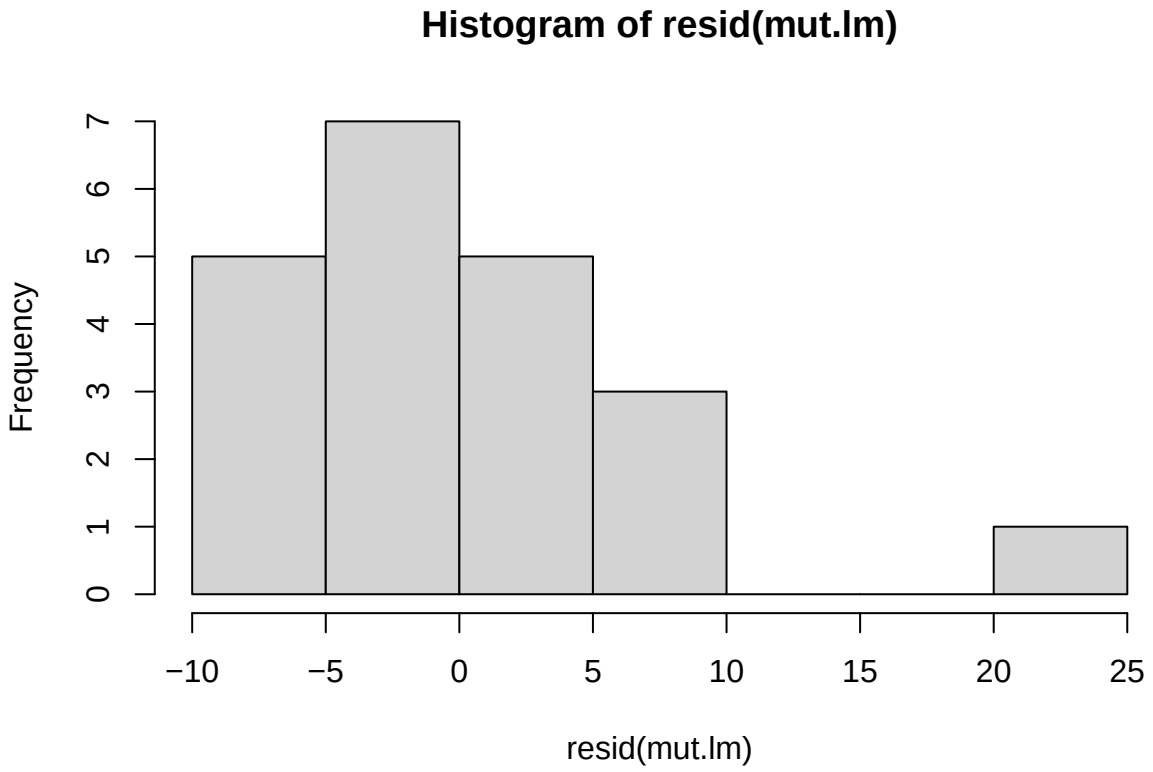

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(numberOfNewMutations ~ AgeOfFather)
Scale−Location
√|Standardized residuals|
Fitted values
lm(numberOfNewMutations ~ AgeOfFather)



These plots suggets that the assumptions seem OK:

- The first plot suggests the variance of the residuals (the vertical spread of residual) is relatively similar along the x-axis, suggesting equal variance. Notice one unusual residual, with the number, 17 beside it; the '17' indicates that this unusual measurement lies on line 17 of the dataframe, `mut`. If we were worried that this measurement affected our conclusions, we could run the analysis a second time without this observation and compare the results: if the conclusions are the same, we will know that observation 17 has little impact on our conclusions. If, instead, the conclusions differ markedly, then we'd conclude that observation 17 strongly affects our conclusions as we should report the results from both analyses (i.e., with and without observation 17) to let the reader decide how to interpret the results. -The second plot suggests that normality is OK... this plot is a bit ugly, but overall is OK. This alternative plot (`hist(resid(mut.lm))`), below, suggests the residuals are a bit skewed, nut, normality is likely sufficient.
- The third plot generally looks OK (i.e., indicated equal variance), but the wobble in the red line draws our attention to observation 17 again. To save space (and your reading), we will not analyze the data twice (with and without observation 17), but you should give this a try!

```
hist(resid(mut.lm))
```



Let's check the results from this model:

```
anova(mut.lm)
```

```
## Analysis of Variance Table
##
## Response: numberOfNewMutations
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## AgeOfFather  1 1818.33 1818.33  38.754 5.584e-06 ***
## Residuals   19  891.48   46.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the small p-value, we have strong evidence that `AgeOfFather` is related to the number of mutations passed to offspring.

NOTE that `anova()` does not give estimates of the slope or intercept. To obtain these estimates, use summary():

```
summary(mut.lm)
```

```
##
## Call:
## lm(formula = numberOfNewMutations ~ AgeOfFather, data = mut)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.865 -4.859 -1.534  2.221 20.307
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.2090     7.0494   2.157    0.044 *
```

```
## AgeOfFather   1.5828     0.2543   6.225 5.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.85 on 19 degrees of freedom
## Multiple R-squared:  0.671,  Adjusted R-squared:  0.6537
## F-statistic: 38.75 on 1 and 19 DF,  p-value: 5.584e-06
```

This output indicates that the slope of the line fitted through the data has a y-intercept of 15.2090 (SE = 7.0494) and a slope of 1.5828 (SE = 0.2543).

Think about what that slope means. It indicates that for every year that a father ages, we can expect (on average) that he'll pass on 1.58 more mutations to his offspring. That seems like a big effect if age (especially given that the maximum number of new mutations in the dataset is only 83)! Moreover, this dataset does not have particularly old fathers: the age of fathers ranges from 16 to 37, which seems relatively young to me (I was 40 when we had our second child)! If this trend continues for older fathers, the overall difference in mutation numbers passed to offspring from young vs. much older fathers will be even greater.

Question 4 - MOLE-RATS

This question involves mole rats, which are mammals with distinct social castes (like bees, with a Queen, workers and male bees). The only individuals in a colony that reproduce include a single queen with several males that she mates with. Everyone else is a "worker", who do everything from gather food to care for the young to defend the colony. However, it is possible that "workers" come in at least two varieties: "frequent" and "infrequent" workers. Scantlebury et al. (2006) wished to test whether aspects of physiology differed between these two putative types of workers. They measured daily energy expenditure for individuals believed to belong to the two types of workers. They also measured body mass because they expected that body mass will affect energy expenditure, and they wished to account for this possible effect when testing for differences in physiology between the two (putative) types of workers.

Here are the data:

```
mr <- read.table("moleRats.csv",sep=',',header=TRUE)
str(mr)
```

```
## 'data.frame':    35 obs. of  3 variables:
##  $ caste   : chr  "worker" "worker" "worker" "worker" ...
##  $ lnMass  : num  3.85 3.99 4.11 4.17 4.25 ...
##  $ lnEnergy: num  3.69 3.69 3.69 3.66 3.87 ...
```

```
summary(mr)
```

```
##     caste               lnMass         lnEnergy
##  Length:35          Min.   :3.850   Min.   :3.555
##  Class :character   1st Qu.:4.248   1st Qu.:3.902
##  Mode  :character   Median :4.511   Median :4.190
##                     Mean   :4.540   Mean   :4.193
##                     3rd Qu.:4.844   3rd Qu.:4.489
##                     Max.   :5.263   Max.   :5.043
```

```
head(mr)
```

```
##    caste   lnMass lnEnergy
## 1 worker 3.850148 3.688879
## 2 worker 3.988984 3.688879
## 3 worker 4.110874 3.688879
## 4 worker 4.174387 3.663562
## 5 worker 4.248495 3.871201
```

```
## 6 worker 4.262680 3.850148
```

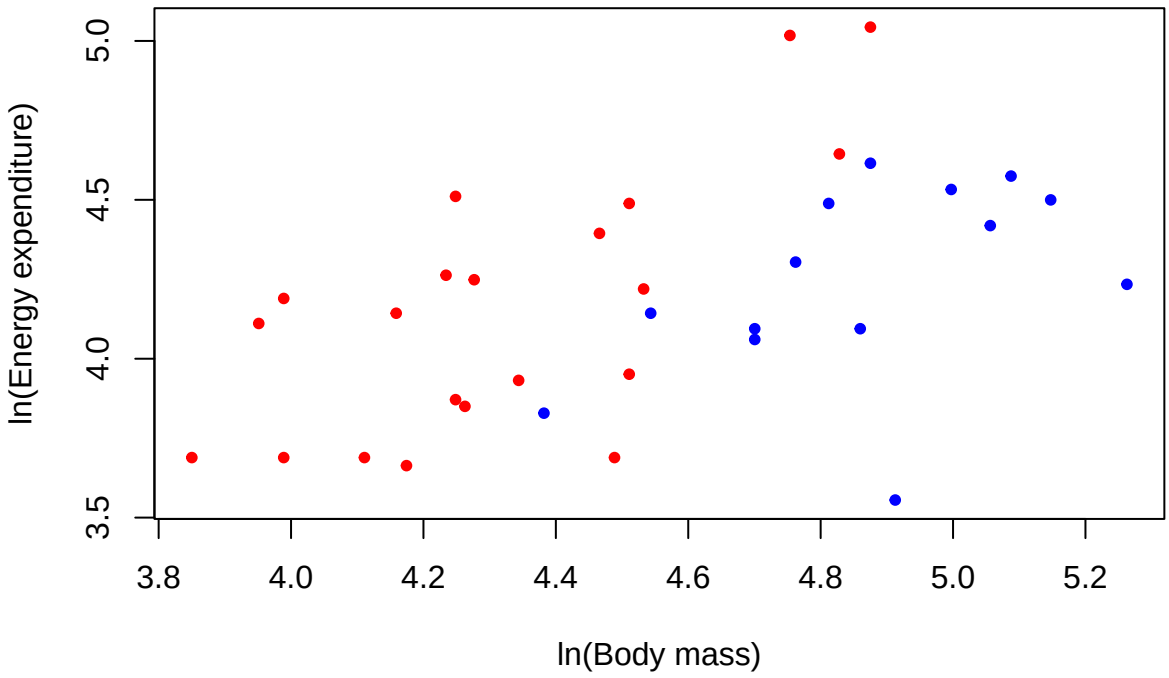
We can see that the column `caste` denotes the two worker types: `lazy` and `worker`. The data in columns `lnMass` and `lnEnergy` denote each mole rat's mass and energy expenditure, respectively; both variables have already been log-transformed. We notice that the variable, `caste` is currently type, `chr`; let's change it to a Factor:

```
mr$caste <- factor(mr$caste)
str(mr)
```

```
## 'data.frame':    35 obs. of  3 variables:
##  $ caste   : Factor w/ 2 levels "lazy","worker": 2 2 2 2 2 2 2 2 2 2 ...
##  $ lnMass  : num  3.85 3.99 4.11 4.17 4.25 ...
##  $ lnEnergy: num  3.69 3.69 3.69 3.66 3.87 ...
```

Now that that data are in the state we wish, let's plot the data. `lnEnergy` will be the dependent variable:

```
plot(lnEnergy~lnMass, xlab="ln(Body mass)", ylab="ln(Energy expenditure)",
     pch=20, col=ifelse(caste=="worker", "red", "blue"),data=mr)
```

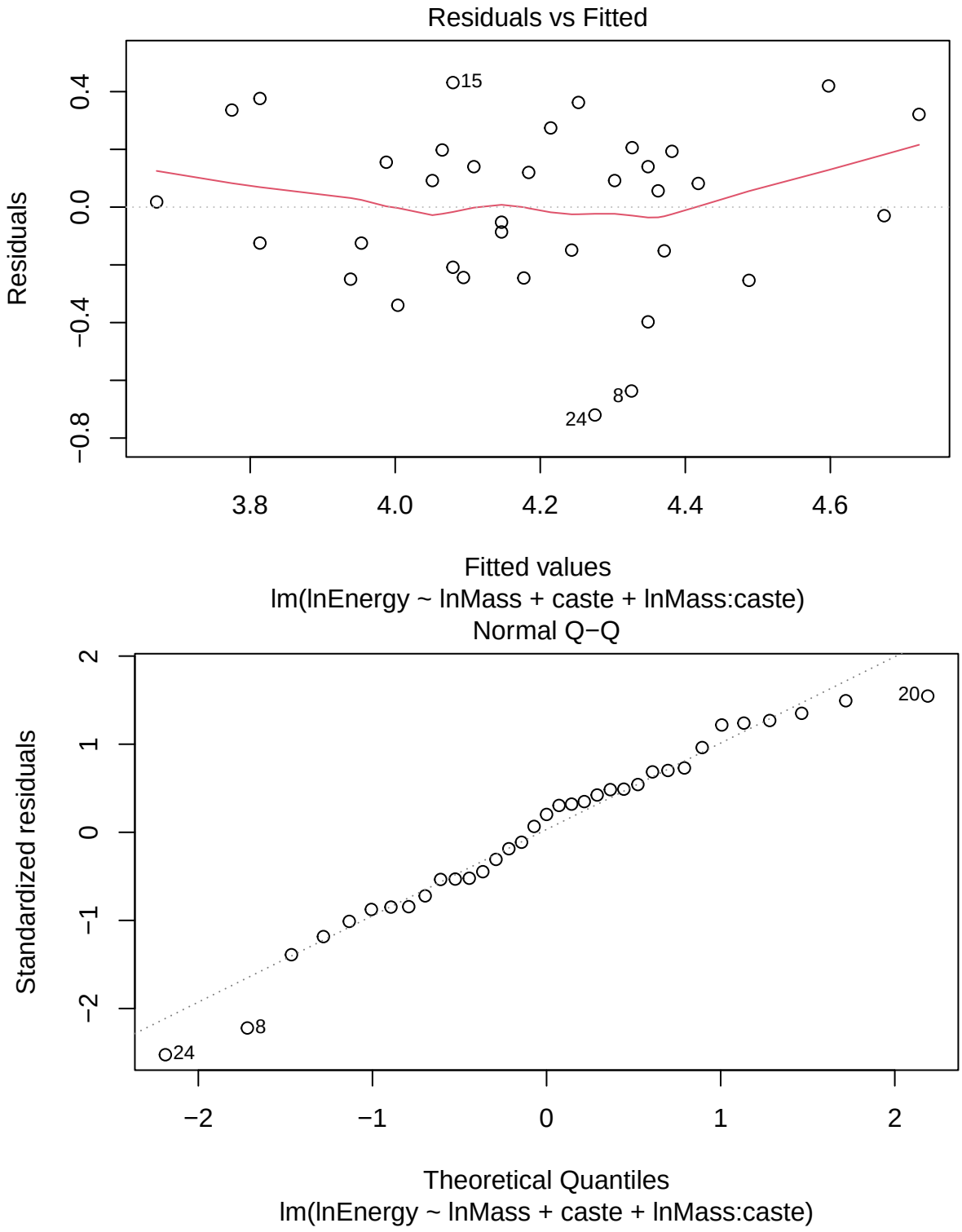


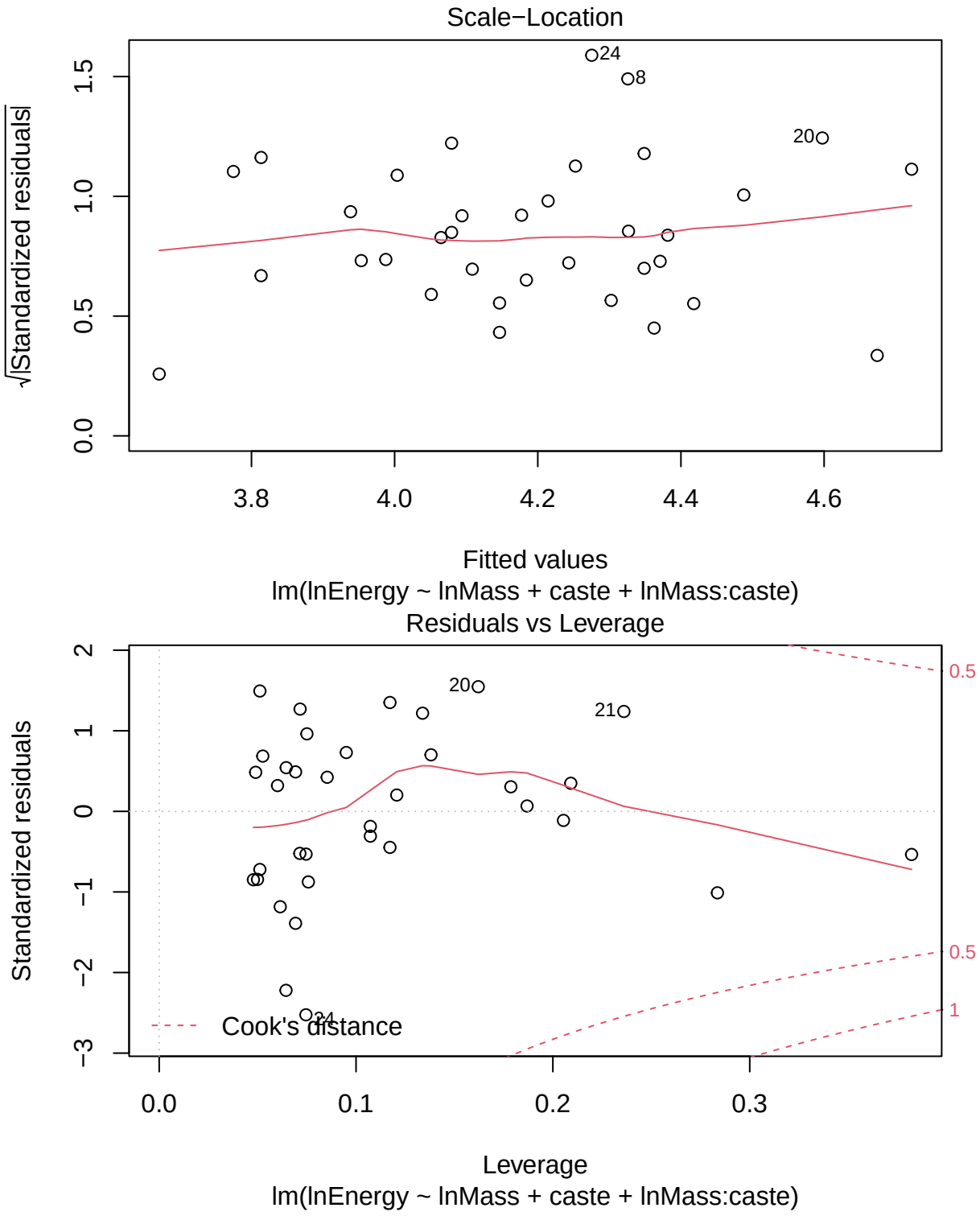
These data indicate that:

- the energy expenditure of `worker` mole rats (red points in the plot) tend to be higher than their `lazy` counterparts (blue points).
- energy expenditure seems to increase with body mass for both caste types
- it is difficult to predict whether the slopes will be different for the two `caste` types, as there are too few data to guess effectively. We'll need to run a model to learn more.

Let's model the data; for now, we'll not use type 3 sum of squares (we'll do that later when we calculate p-values):

```
mr.lm <- lm(lnEnergy~lnMass + caste + lnMass:caste,data=mr)
plot(mr.lm)
```


Residuals vs Fitted
Residuals
0.4
0.0
−0.4
−0.8
15
8
24
3.8
4.0
4.2
4.4
4.6
Fitted values
lm(lnEnergy ~ lnMass + caste + lnMass:caste)
Normal Q−Q
Standardized residuals
2
1
0
−1
−2
20
8
24
−2
−1
0
1
2
Theoretical Quantiles
lm(lnEnergy ~ lnMass + caste + lnMass:caste)



We'll follow the author's lead and say that the data are randomly selected and independent. What about equal variance and normality of residuals?

- The first plot shows that the variance (i.e., vertical spread of the residuals) is relatively equal along the x-axis, implying that the data meet the assumption of equal variance;
- The second plot suggests that the residuals are nicely normally distributed.
- The third plot also indicates that the data meet the assumption of equal variance.

Now that we're satisfied that the data meet the assumptions, let's check our results:

```
summary(mr.lm)
```

```
##
## Call:
## lm(formula = lnEnergy ~ lnMass + caste + lnMass:caste, data = mr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72004 -0.17990  0.05631  0.19551  0.43128
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.2939     1.6691   0.775   0.4441
## lnMass               0.6069     0.3428   1.771   0.0865 .
## casteworker         -1.5713     1.9518  -0.805   0.4269
## lnMass:casteworker   0.4186     0.4147   1.009   0.3206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2965 on 31 degrees of freedom
## Multiple R-squared:  0.4278, Adjusted R-squared:  0.3725
## F-statistic: 7.727 on 3 and 31 DF,  p-value: 0.0005391
```

The intercept refers to the y-intercept of `lazy` caste members; likewise, the coefficient `lnMass` refers to the slope for `lazy` mole rats, which equals 0.6069. The slope for `worker` mole rats equals `0.6069 + 0.4186 = 1.0255`. These slopes do seem to be different (one slope is two-thirds larger than the other). But, these slopes were estimated with relatively small sample sizes (only 14 and 21 individuals in the two `caste` groups), so this apparent difference in slopes may simply arise due to sampling error. Indeed, the p-value for `lnMass:casteworker` equals, 0.3206, indicating little evidence that the slopes differ. Let's calculate p-values, remembering to use type 3 sum of squares (remember to add the `contrasts` option to our linear model):

```
mr.lm.t3 <- lm(lnEnergy~lnMass+caste+lnMass:caste,data=mr,
               contrasts = list(caste = contr.sum))
```

```
library(car)
Anova(mr.lm.t3, type  = 3)
```

```
## Anova Table (Type III tests)
##
## Response: lnEnergy
##               Sum Sq Df F value    Pr(>F)
## (Intercept)  0.02385  1  0.2713 0.6061609
## lnMass       1.36179  1 15.4922 0.0004363 ***
## caste        0.05696  1  0.6481 0.4269408
## lnMass:caste 0.08956  1  1.0188 0.3206094
## Residuals    2.72494 31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the interaction ($p = 0.3206094$) indicates that we have very little evidence to suggest that the slopes for the relationship between `lnMass` and `lnEnergy` differs between the `caste` types (it also matches the output from `summary(mr.lm)`, above).

We'll proceed by assuming that the slopes do not differ between the levels of `caste`. We'll follow two different approaches to analyze these data.

APPROACH 1 - estimate the main effects using emmeans

If we're satisfied in the belief that the slopes are similar for the two `caste` levels, we can proceed by characterizing the main effects. (Remember that we could not interpret these main effects if we had concluded that `lmMass` and `caste` interacted to determine `lnEnergy`.)

The question that motivated the study was to determine whether the physiology differed between the two putative `caste` types. Therefore, we'll focus on addressing this question, using `emmeans`:

```
mr.emmeans.caste <- emmeans(mr.lm.t3, "caste")
```

```
## NOTE: Results may be misleading due to involvement in interactions
```

```
mr.emmeans.caste
```

```
##  caste  emmean    SE df lower.CL upper.CL
##  lazy     4.05 0.136 31     3.77     4.33
##  worker   4.38 0.082 31     4.21     4.55
##
## Confidence level used: 0.95
```

These results indicate that the `emmean` (and SE) `lnEnergy` (having adjusted for effects of `lnMass`) tends to be higher for `worker` mole rats (4.38 (0.082)) than for `lazy` mole rats (4.05 (0.136)). (Remember that these values are for log-transformed data.)

Let's compare these estimated, marginal mean (`emmean`) values using the `pairs()` function:

```
mr.pairs.caste <- pairs(mr.emmeans.caste)
mr.pairs.caste
```

```
##  contrast      estimate    SE df t.ratio p.value
##  lazy - worker   -0.329 0.159 31 -2.068  0.0470
```

The effect size (and SE) of `caste` equals -0.329 (0.159). We also see that, unlike the output from `Anova()`, above, the p-value (0.0470) indicates moderate, or 'suggestive' evidence for a difference in energy expenditure between levels of `caste`.

We can calculate 95% CI's for this effect size like this:

```
confint(mr.pairs.caste)
```

```
##  contrast      estimate    SE df lower.CL upper.CL
##  lazy - worker   -0.329 0.159 31   -0.654 -0.00457
##
## Confidence level used: 0.95
```

APPROACH 2 - drop the interaction from the model

I'll provide less commentary for this approach, but outline the code:

```
mr.lm.t3.main <- lm(lnEnergy~lnMass+caste,data=mr,
               contrasts = list(caste = contr.sum))
```

```
library(car)
Anova(mr.lm.t3.main, type  = 3)
```

```
## Anova Table (Type III tests)
##
## Response: lnEnergy
##              Sum Sq Df F value    Pr(>F)
## (Intercept) 0.00111  1  0.0126    0.9112
## lnMass      1.88152  1 21.3923 5.887e-05 ***
```

```
## caste       0.63747  1  7.2478    0.0112 *
## Residuals   2.81450 32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mr.emmeans.caste.main <- emmeans(mr.lm.t3.main, "caste")
mr.emmeans.caste.main
```

```
##  caste  emmean    SE df lower.CL upper.CL
##  lazy     3.96 0.101 32     3.75     4.16
##  worker   4.35 0.077 32     4.19     4.51
##
## Confidence level used: 0.95
```

```
mr.pairs.caste.main <- pairs(mr.emmeans.caste.main)
mr.pairs.caste.main
```

```
##  contrast      estimate    SE df t.ratio p.value
##  lazy - worker   -0.393 0.146 32 -2.692  0.0112
```

Once again, we find moderate or 'suggestive' evidence that energy levels differ between the `worker` and `lazy` caste levels.

```
confint(mr.pairs.caste.main)
```

```
##  contrast      estimate    SE df lower.CL upper.CL
##  lazy - worker   -0.393 0.146 32   -0.691  -0.0957
##
## Confidence level used: 0.95
```

On the log scale, the `worker` mole rats expend 0.393 more energy than `lazy` mole rats (for the population and conditions exampled), once accounting for differences in body size; this difference plausibly ranges from 0.0957 to 0.691.

Question 5

As you might have guessed, Marland et al. (2020)'s approach to interpret their data is not appropriate. To understand why, let's consider their goal or hypothesis. They appear to be interested in making a statement about the effect of BSA at time 24 relative to the effect of BSA at time 0. In other words, they wish to know whether the effect of BSA depends on the context: time 0 vs. time 24. This perspective implicitly requires that the authors test whether the size or nature of the difference between BSA and Control differs between the two time periods; i.e., their hypothesis invokes an *interaction* between Treatment (BSA, Control) and Time period (0, 24). If Marland et al. (2020) had analyzed their data with a 2-Factor GLM (or something similar) they could have tested for such an interaction: if they found evidence for an interaction then they would be justified in stating that the effect of BSA (vs. Control) differs between time periods. In contrast, the analysis they used, involving 2 t-tests, does *not* involve an explicit test of the hypothesis that the effect of BSA differs between the Time periods (the hypothesis of interest); therefore, their conclusions are not appropriate. By using 2 t-tests, they only evaluated differences between BSA vs. Control at each Time, separately, and **did not explicitly tests whether the effect of BSA differed between times**.

Note that Marland et al. (2020)'s mistake occurs commonly: it is *not* correct to assume that a 'significant' result in one context but not in another means that effects differ (in size, and generally) between the contexts. If we wish to test for 'context dependence' of results, we require an analysis of an interaction.